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 to the VII/A series, and the lifetime is 14 to 17 years.

It is standard practice to have a spare satellite in orbit for 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.

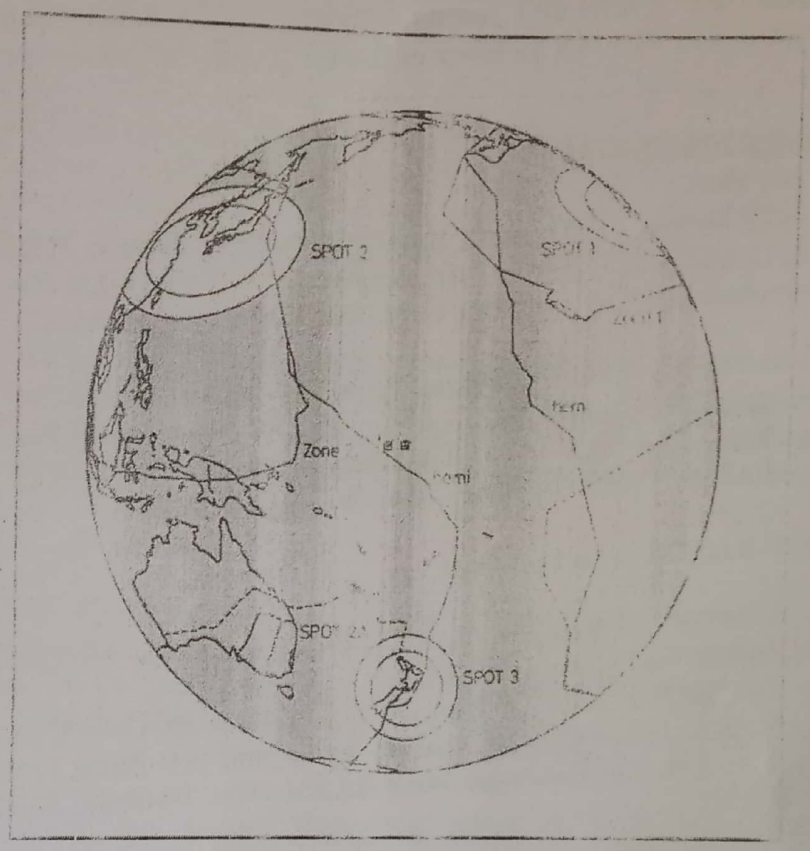


Figure 5.2 Region of glob

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,

...ll India Radio and Doordarshan. The overall coordination and management of INSAT system matters with the Secretary level INSAT Coordination Committee.

INSAT satellites provide transponders in various bands (C, S, Extended C and Ku) to meet the television and communication needs of India. Some of the satellites also have the Very High Resolution Radiometer (VHRR), CCD cameras for meteorological 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(Previous). 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 36 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 L-Band Transponders With 42 Dbw EIRP Over India And An MSS Payload Similar To Those On INSAT-3B And INSAT-3C.
INSAT-4 Series:

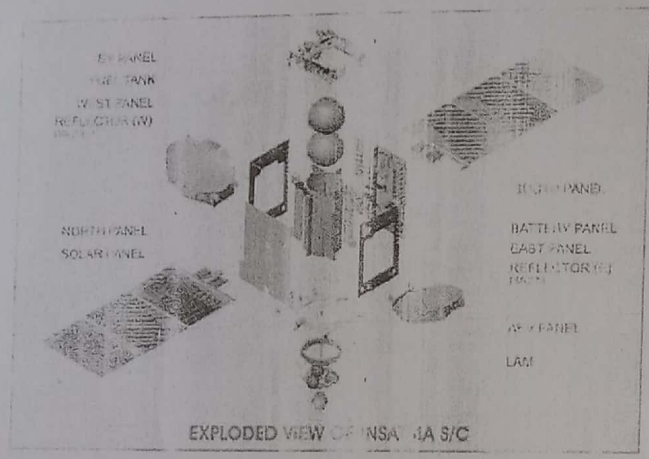


Figure 5.3 INSAT 4A

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 boundary, 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.

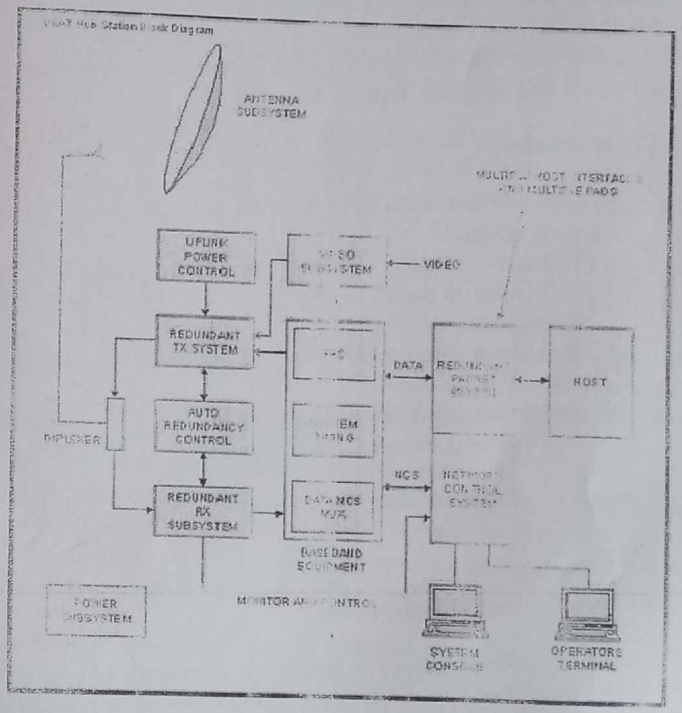


Figure 5.4 VSAT Block Diagrams

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 organization has exclusive access to its own VSAT network.

Time division multiplex 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).

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.

3.4 Mobile satellite services:

3.4.1 GSM:

3.4.1.1 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 MHz 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 Services 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.

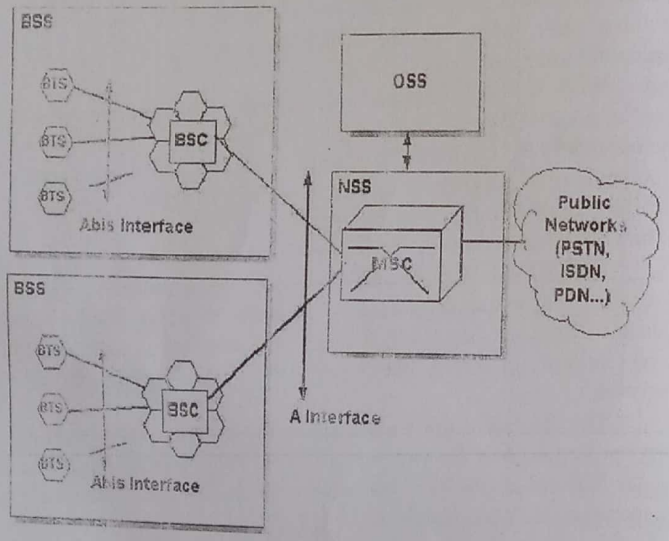


Figure 5.5 GSM Block Diagrams

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

9.5 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.

5.4.2 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.

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.

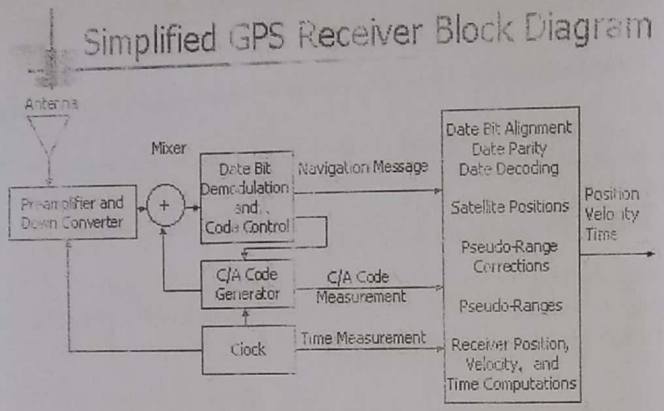


Figure 5.6 GPS Block Diagrams

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.

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.

5.5. 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.

- Development of GNSS Ground Based Augmentation System (GBAS) Continues
  - Local Area Augmentation System (LAAS)
- GNSS is Cornerstone for National Airspace System

5.10 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.

5.10.1 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 guardband 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.

Uplink MHz	1	3	5	RHCP		31
Downlink MHz	17324.00 12224.00	17353.16 12253.16	17382.32 12282.32	. . .		17761.40 12661.40

Uplink MHz	2	4	6	LHCP		32
Downlink MHz	17338.58 12238.58	17367.74 12267.74	17411.46 12296.50	. . .		17775.98 12675.98

Figure 5.12 DBS Service

5.10.2 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.

5.10.3 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.

5.11 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

	1	3	5	RHCP		31
Uplink MHz	17324.00	17353.16	17382.32	. . .		17761.40
Downlink MHz	12224.00	12253.16	12282.32			12661.40

	2	4	6	LHCP		32
Uplink MHz	17338.58	17367.74	17411.46	. . .		17775.98
Downlink MHz	12238.58	12267.74	12296.60			12675.98

Figure 5.12 DBS Service

5.10.2 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.

5.10.3 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.

5.11 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.

5.11.1 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.

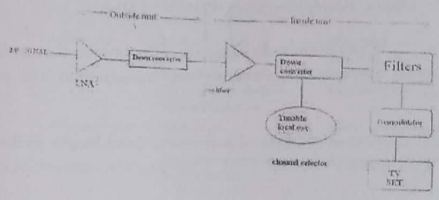


Figure 5.13 DTH Service

5.11.2 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.

5.12 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
  - o DAB-S (Satellite) – using QPSK – 40 Mb/s
  - o DAB-T (Terrestrial) – using QAM – 50 Mb/s
  - o 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.

5.13 Worldspace services:

WorldSpace (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. WorldSpace delivers the latest tunes, trends and information from around the world and around the corner.

WorldSpace subscribers benefit from a unique combination of local programming, original WorldSpace content and content from leading brands

G. Prasath. MSc

Table 22.1. *Current status of organ replacement technology. (Adapted from The Biomedical Engineering Handbook ed J D Bronzino (Boca Raton, FL: Chemical Rubber Company).)*

Clinical standing	Artificial organ	Transplantation
Generally accepted	Heart lung machine	Blood transfusion
	Large joint prostheses	Corneal transplants
	Bone fixation systems	Banked bone
	Cardiac pacemakers	Bone marrow
	Large diameter vascular grafts	Kidney, cadaveric donor
	Prosthetic heart valves	Heart
	Intra-aortic balloon pumps	Liver
	Implantable lenses	Heart/lung
	Hydrocephalus shunts	
	Dental implants	
	Skin or tissue expanders	
Accepted with reservations	Maintenance haemodialysis	Kidney, living related donor
	Chronic ambulatory	Whole pancreas
	Peritoneal dialysis	
	Breast implants	
	Sexual prostheses	
	Small joint prostheses	
	Extracorporeal membrane	
	Oxygenation in children	
	Cochlea prostheses	
Limited application	Implantable defibrillator	Pancreatic islets
	ECMO in adults	Liver lobe or segment
	Ventricular-assist devices	Cardiomyoplasty
	Artificial tendons	
	Artificial skin	
	Artificial limbs	
Experimental	Artificial pancreas	Gene transfer
	Artificial blood	Embryonic neural tissue
	Intravenous oxygenation	Bioartificial pancreas
	Nerve guidance channels	Bioartificial liver
	Total artificial heart	
Conceptual stage	Artificial eye	Striated and cardiac muscle
	Neurostimulator	Functional brain implants
	Blood pressure regulator	Bioartificial kidney
	Implantable lung	
	Artificial trachea	
	Artificial oesophagus	
	Artificial gut	
	Artificial fallopian tube	

22.3.1. *Artificial heart valves*

How do natural heart valves work?

In the normal human heart, the valves maintain a unidirectional flow of blood with minimal frictional resistance, whilst almost completely preventing reverse flow. They act passively: the moving parts of the valve, the tissue

leaflets, have negligible inertia and open and close in response to pressure changes in the blood generated by the contraction of the surrounding myocardium (see section 2.6). All four valves sit within a flat fibrous supporting structure which is part of the fibrous skeleton of the heart. This separates the ventricles from the atria and the ventricular outflow conduits (the aorta and pulmonary artery), and is perforated by four openings two larger ones for the atrio-ventricular valves (mitral and tricuspid valves) and two smaller ones for the aortic and pulmonary valves (figure 22.9).

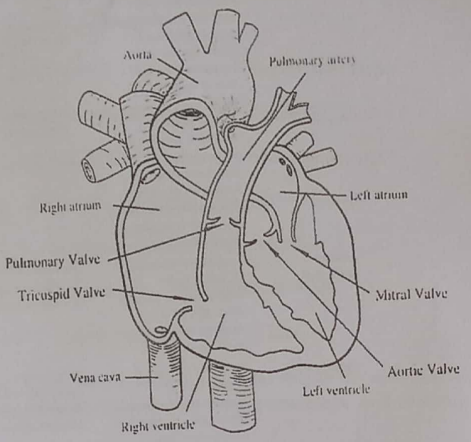


Figure 22.9. *Anatomy and structure of heart valves.*

Why do natural valves sometimes need to be replaced?

Whereas a normal heart valve opens passively with little measurable transvalvular pressure difference, and closes effectively with minimal leakage, diseased valves cause major haemodynamic abnormalities and an increased workload for the heart. Heart valve disease affects primarily the valves on the left side of the heart and is associated with increased leakage, increased resistance to forward flow or a combination of the two. The terms 'insufficiency', 'incompetence' or 'regurgitation' are used to indicate backflow through the closed valve. Excessive regurgitation leads to a progressive enlargement of the cardiac chamber upstream of the valve which now fills from both directions. The term 'stenosis' describes a blocking or narrowing of the valvular orifice and requires an increased amount of energy to drive the blood through. The restriction results in an abnormally high pressure difference across the valve and, with time, leads to hypertrophy of the cardiac chamber upstream of the stenotic valve.

The consequences of valve pathology will depend on the ability of the heart to adapt to the increasing demands. The severity of valve abnormalities can be assessed using ultrasonic imaging which can be used to observe retrograde and turbulent flow around heart valves *in vivo* (see section 19.7).

In cases where cardiac function is severely compromised the damaged natural valve can be replaced with an artificial substitute. This involves the replacement of the moving parts of the diseased valve. Over the years, two distinctly different types of prosthetic valves have been developed: *mechanical valves* which are made of a variety of synthetic materials and *bioprosthetic valves* made of chemically modified biological tissues. Most valve mechanisms have a rigid frame to house the moving flow occluder, whilst anchorage of the device also necessitates suitable accessory structures. The specialized terms used to describe the components of an artificial valve are defined in table 22.2.

Table 22.2. *Nomenclature for prosthetic heart valve components.*

Term	Meaning
Mechanical valve/prosthesis	Heart valve substitute manufactured entirely from man-made materials
Tissue valve/bioprosthesis	Valve substitute manufactured, in part, from chemically treated biological material
Poppet or occluder	Mobile component, typically a ball or disc, which moves to open and close valve
Housing	Assembly which retains the occluder
Frame or stent	Rigid or semi-rigid support for flexible leaflets of a tissue valve
Cusps or leaflets	Flexible components of a tissue valve which open and close in response to flow
Sewing ring	Fabric cuff surrounding stent or housing, used by surgeon to anchor the valve in place by suturing

What does valve replacement involve?

Whilst a detailed description of the operative technique is outside the scope of this book, a typical procedure is described briefly in the section which follows. In order to replace the aortic valve, the surgeon will open the chest, cutting through the sternum to expose the heart. The patient is then placed on cardiopulmonary bypass (see section 22.3.2) to maintain the oxygenation and circulation of the blood. To reduce tissue damage, the body temperature is cooled to 34 °C by passing the blood through a refrigerated heat exchanger. The ascending aorta is clamped, and the heart is arrested by potassium cold cardioplegia. This ensures that the heart is stopped in a relaxed state. The damaged aortic valve tissue is cut away and replaced by a prosthesis which is sutured into the aortic annulus below the coronary arteries. The aorta is closed and the left ventricle is vented to remove any trapped air. The aortic clamp is removed and the heart defibrillated. The patient is rewarmed and removed from cardiopulmonary bypass.

Treatment with anti-coagulants is required during cardiopulmonary bypass to prevent thrombosis and thromboembolism. In addition, patients receiving mechanical valves will require controlled anti-coagulation for the rest of their lives. Bioprosthetic valves are less inherently thrombogenic and may not require long-term anti-coagulants.

A brief history of valve development

The first clinical use of a mechanical heart valve was performed by Dr Charles Hufnagel in 1952, who partially corrected aortic incompetence by implanting an acrylic ball valve into the descending aorta. The introduction of cardiopulmonary bypass in 1953 enabled open heart procedures to be performed and, in 1960, Harken and Starr implanted ball valves enclosed in metal cages, the former in the aortic position, the latter as a mitral valve replacement (see figure 22.10).

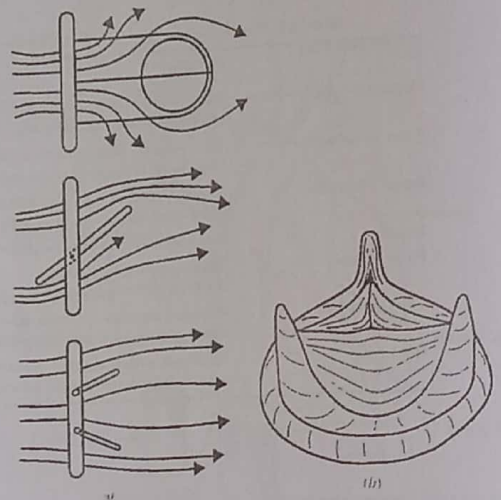


Figure 22.10. *(a) Three types of mechanical prosthetic valve. A caged ball valve is shown at the top, a tilting disc valve in the middle and a bileaflet valve at the bottom. (b) A porcine bioprosthetic valve.*

It was soon realized that the central occluder of the caged-ball valve presented a degree of obstruction to the blood flow. This was particularly significant in patients with a narrow aortic root. In addition, in the mitral position the high profile of the cage could impinge on the wall of a small ventricle. For these reasons, attempts were made to find less obstructive designs. In consequence, low-profile tilting-disc valves were introduced. Tilting-disc valves presented significantly less resistance to blood flow than the caged-ball valves, and appeared to have lower incidences of thrombosis and thromboembolism. The first tilting disc valves had rigid plastic (Teflon or Delrin) discs. These were replaced by pyrolytic carbon, a material which combines durability with low thrombogenicity, in 1971.

A significant development came in 1978 with the introduction of the all pyrolytic carbon valve by St Jude Medical Inc. This valve comprises two semi-circular discs or leaflets which, in the open position, present minimal flow disturbance. The success of this valve and its relatively low incidence of thrombotic complications have led to the development of a number of other bileaflet designs which currently account for 75% of valve implants.

Although there have been significant improvements in the design of mechanical valves over the years, all mechanical valves cause flow disturbances which may ultimately lead to thrombosis or thromboembolism. For this reason, patients with mechanical valves are required to undergo long-term anti-coagulant therapy.

Many valve designers have sought biological solutions to this problem using natural tissues to fabricate valves. These valves have been developed in parallel with mechanical designs. Pioneers used free-hand constructed valves made of the patient's own tissue. This technique was first reported in 1966 using fascia lata, a sheet of fibrous tissue which covers the muscles of the thigh. The valves gave poor results with premature tearing or calcification and many failed within 5 years of implantation.

The next advance came in 1969 with the introduction of bioprosthetic valves in the form of frame-mounted pig aortic valves. The tissue was chemically modified by glutaraldehyde fixation (tanning). Pig valves had two main disadvantages: they were difficult to obtain in small sizes and, as they were mounted on the inside of a support frame, they presented an obstruction to flow. In an attempt to overcome these problems a valve was introduced in which the leaflets were fashioned from glutaraldehyde-fixed bovine pericardium. In this case, the leaflets were attached to the outside of the frame thus maximizing the area of orifice available for flow. These valves could also be manufactured in a full range of sizes.

An alternative has been the use of human allografts (or homografts). These are human aortic valves which are obtained from cadavers. Prior to use they are cryopreserved or treated with antibiotics. The first allograft procedure was carried out by Ross in 1962. As allografts are normally implanted without a stent, the level of technical skill and operative time required for implantation is much greater than that required for a frame-mounted bioprosthesis. However, as the long-term results have proved encouraging, in recent years, there has been a renewed interest in their use. The use of homografts is restricted by the available supply rather than the demand.

The most recent trend for aortic valve replacement is the use of a stentless bioprosthesis. In acknowledgement of the success of the unmounted allografts, isolated or joined cusps, sometimes surrounded by a conduit of biological tissue, are sutured directly to the patient's aorta. The tissues employed are usually of bovine or porcine origin and are chemically treated. Unstented valves have improved haemodyamics when compared with traditional stented bioprostheses, as they avoid the narrowing of the orifice which is associated with the use of a stent and a sewing ring, thus giving minimal resistance to flow. However, like allografts, they are more demanding in terms of surgical skill and may prove difficult to replace should the need arise.

Research is also being carried out in an attempt to produce a leaflet valve made entirely from man-made materials. To date, attempts to manufacture flexible leaflet valves from synthetic polymers have not been

Table 22.3. *Summary of the material composition of nine key designs of mechanical valves developed over the course of 30 years.*

Year	Name	Type	Poppet	Material
1959	Hufnagel	Ball	Polypropylene	Methacrylate
1964	Starr–Edwards 1000	Ball	Silastic	Stellite
1968	Wada–Cutter	Tilting disc	Teflon	Titanium
1969	Bjork–Shiley	Tilting disc	Delrin	Stellite
1970	Lillehei–Kaster	Tilting disc	Pyrolitic carbon	Titanium
1971	Bjork–Shiley	Tilting disc	Pyrolitic carbon	Stellite
1977	Medtronic–Hall	Tilting disc	Pyrolitic carbon	Titanium
1977	St Jude Medical	Bi-leaflet	Pyrolitic carbon	Pyrolitic carbon
1991	Jyros	Bi-leaflet	Vitreous carbon	Vitreous carbon

successful. No synthetic material yet produced is proven to exhibit the flexural durability of natural valve cusps (see table 22.3).

What are the design criteria?

The aim of the valve designer is to produce a device which:

- Is durable (capable of functioning for 35–40 million cycles per year for up to 30 years)
- Has a large orifice area and presents the lowest possible resistance to blood flow
- Creates a minimal degree of flow separation and stasis
- Does not induce regions of high shear stress
- Has minimal regurgitation.
- Causes minimal damage to blood and plasma constituents
- Is manufactured from materials which are non-thrombogenic
- Is easy to implant, is quickly incorporated into, and tolerated by, the patient's tissues
- Creates no noise.
- Can be simply and consistently manufactured at an acceptable cost

No current design, other than the native valve, meets all of the above criteria. All prosthetic valves are inherently stenotic when compared with the natural valve. The implantation of any artificial valve unavoidably restricts the tube-like flow paths found in the normal heart and introduces a degree of narrowing. In addition, the shapes of the natural valves change during the different phases of the cardiac cycle and the introduction of a stent or housing imposes a fixed geometry to the valve orifice. Artificial valve design, to date, has been a compromise between an unattainable ideal and the technically feasible.

To simplify matters we should concentrate on a small number of key requirements which we can go some way towards attaining.

Valves should:

- Function efficiently and present the minimal load to the heart.
- Be durable and able to function for the life-time of the patient
- Not cause thrombus formation or promote the release of emboli

Current designs

Currently, seven basic configurations of heart valve substitutes are produced commercially: caged-ball valves, tilting-disc valves and bi-leaflet valves, frame-mounted porcine aortic and bovine pericardial bioprosthetic valves and stentless valves. Cryopreserved human valves are also available. Data presented in 1993 showed that the total world market was of the order of 130 000 valves. The USA accounted for 60 000 implants per year and 5000 valve implants were performed every year in the UK.

Table 22.4 summarizes the advantages and disadvantages of mechanical and bioprosthetic valves.

Which valve is best?

There is no simple answer to this question and a number of factors must be considered, including the particular circumstances of the individual patient. The major advantage of mechanical valves is long-term durability. The major advantage of bioprosthetic valves is low thrombogenicity. Enhanced durability coupled with the need for life-long anti-coagulation associated with mechanical valves must be weighed against the better quality of life, freedom from anti-coagulation-related complications such as haemorrhage, but limited durability and risks of reoperation which are associated with a bioprosthetic valve. A recent overview of valve usage suggests

Table 22.4. A comparison of mechanical and tissue heart valves.

Valve type	Advantages	Disadvantages
Mechanical	Long-term durability	Unnatural form
	Consistency of manufacture	Patient usually requires long-term anti-coagulant therapy
Tissue	More natural form and function	Uncertain long-term durability
	Less need for long-term anti-coagulant therapy	Consistency of manufacture is more difficult *In vivo* calcification

that bioprosthetic valves are used in 40% of the patients in the USA, in 25% in the UK and in as many as 80% in Brazil.

Many surgeons recommend mechanical valves for patients below the age of 65 and for valve replacements in young patients where durability is of utmost importance. In the case of children and adolescents, calcification of bioprostheses is particularly rapid and severe. Bioprosthetic valves are considered for elderly patients, for patients in developing countries where access to anti-coagulant control may be limited and for patients who are unlikely to comply with the stringent requirements of anti-coagulation therapy. They are essential for patients for whom anti-coagulation therapy is contra-indicated, for example, in women in the early pregnancy.

It must be remembered that, even though currently available valves may not be ideal, the alternative for patients with native valve failure is progressive cardiac failure and death.

Evaluation of valve performance

A true assessment of valve performance can only be obtained from long-term clinical studies. However the initial quantitative information must be obtained by laboratory evaluation using flow simulators. These tests routinely include measurements of pressure difference for a range of simulated cardiac outputs under both steady and pulsatile flow conditions and regurgitation in pulsatile flow. A knowledge of the mean pressure difference ($\overline{\Delta P}$) across the open valve during forward flow enables the effective orifice area (EOA) to be obtained. The EOA gives a measure of the degree of obstruction introduced by the valve and may be calculated in cm^2 from the following formula given by Yoganathan (1984):

$$\mathrm{EOA} = \frac{Q_{\mathrm{rms}}}{51.6\sqrt{\overline{\Delta P}}}$$

where Q_{rms} is the rms flow rate in $cm^3\,s^{-1}$ over the cardiac cycle. Note the constant 51.6 that is included. It should be appreciated that this equation is only a guide to the effectiveness of a valve.

Calculations of energy loss enable estimates to be made of the total load the valve presents to the heart. Laser Doppler and flow visualization techniques provide information about flow velocities, shear and shear stress fields (see section 2.6 in Chapter 2). These data allow predications of the likelihood of damage to blood cells to be made.

Pulsatile flow testing is carried out using hydrodynamic test rigs, or pulse duplicators. These model the left side of the heart with varying levels of sophistication and attention to anatomical variation. There is no single universally accepted design and many involve a compromise between accurate simulation and

ease of use. It is likely that, in the future, the use of *in vitro* techniques will be superseded by computational fluid-dynamic analyses (CFD).

Wear and durability can also be investigated in the laboratory. Valves are tested at accelerated rates of up to 20 Hz. In this way, the effects of 10 years' mechanical wear can be simulated in a period of 8 months. For mechanical valves, subsequent surface analysis enables the degree of wear to be quantified for individual valve components. Tissue valves are examined for evidence of tissue damage and tears.

22.12 Cardiopulmonary bypass

Concept of bypass

'Bypass' is a term employed by surgeons to indicate that fluid normally circulating through an organ is diverted around it, either to reduce the functional work-load and allow the organ to heal, or to isolate the organ for the duration of a surgical procedure.

During cardiopulmonary bypass (heart/lung bypass) the blood is diverted away from the heart and lungs. As this is incompatible with life beyond a few minutes, surgical procedures involving the heart and main blood vessels must be coupled with artificial maintenance of cardiorespiratory function by a heart–lung machine. This is a mechanical system capable of pumping blood around the body and oxygenating it by means of an appropriate gas exchange unit. Such a system is obviously a safety-critical system.

A heart–lung machine was first used for the treatment of pulmonary embolism in 1937 and cardiopulmonary bypass was first used for open-heart surgery in 1953.

Uses of cardiopulmonary bypass

- As a temporary substitute for heart and lung function during surgery.
- As an extracorporeal membrane oxygenation system to assist respiratory exchange.
- To maintain life after severe damage to heart or lung (myocardial infarction (MI), trauma, pulmonary embolism).
- For short-term assist during invasive therapy (lung lavage).
- For the treatment of respiratory imbalance (hypercapnia).

A typical circuit is shown schematically in figure 22.11. The blood is drained from the patient by cannulation of the inferior and superior vena cavae. The heart and lungs are isolated by cross-clamping the aorta downstream of the aortic valve and the venae cavae at the entrance to the right atrium. The venous blood enters the extracorporeal circuit and is transported to the oxygenator where it is oxygenated and carbon dioxide is removed. The example shown incorporates a specific type of oxygenator, a 'bubble' oxygenator. When using this type of device a defoamer and a bubble trap are required to remove gaseous emboli which, if allowed to enter the patient's circulation, may cause brain, lung and kidney damage. A heat exchanger enables the blood to be cooled in a controlled manner inducing systemic hypothermia. Before being returned to the patient the oxygenated blood is filtered. This arterial filter removes microaggregates and any residual gas micro-bubbles. Blood is then returned to the body by means of a cannula in the aorta.

Blood released into the chest during surgery is sucked through a second blood circuit from the surgical field and returned to the system.

Oxygenators

Two types of oxygenator are currently in clinical use. These are direct contact and membrane types.

Direct contact types are usually 'bubble' type oxygenators which allow direct contact between the blood and gas. Oxygen is bubbled through a series of compartments containing venous blood. This process

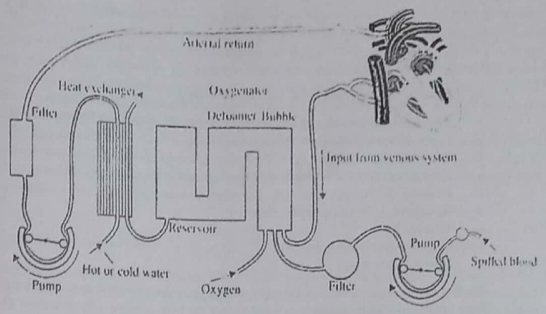


Figure 22.11. *A typical circuit for cardiopulmonary bypass.*

causes foaming of the blood. Defoaming is then carried out by passing oxygenated blood over silicone coated screens. Contact between the blood and bubbles will result in damage to blood elements and protein denaturation due to the high interfacial energies involved. Foaming precipitates fibrin and increases platelet activation.

In membrane-type oxygenators blood is separated from the gas phase by a permeable polymeric sheet or tube.

Three types of membrane are currently used. Each has advantages and disadvantages.

- *Homogeneous*: the membrane takes the form of a continuous sheet of solution/diffusion membrane.
- *Microporous*: has a high void volume.
- *Composite*: thin film of solution/diffusion polymer on a microporous substrate.

Homogeneous. Gas diffuses into the membrane polymer at the interface and from the polymer into the blood. This process is slow, requiring long perfusion times and a large area for adequate exchange. (Materials used include PTFE, polyurethane and polysiloxane.)

Microporous. Pores are introduced during the manufacturing process (e.g. porous polypropylene).

Composite. A film of homogeneous polymer on a microporous substrate (e.g. 25 μm layer of polysulphone cast onto porous polypropylene.)

The choice of membrane will depend on the balance between gas permeability, strength and blood compatibility. There are two common geometries of membrane oxygenator design. These are multiple flat channels and multiple hollow fibre types. The area of membrane required to obtain the correct level of blood oxygenation will depend on the design and can be calculated if it is assumed that complete saturation of the haemoglobin is required.

Pumps

We can list the design requirements for a suitable pump

- Be capable of flow rates up to 10 l min^{-1} and be able to achieve this against a pressure of 180 mmHg
- Cause minimal clotting and thrombus formation
- Not promote gas emboli
- Have no hot spots which might damage the blood
- Be easily sterilized
- Be capable of being calibrated accurately
- Be reliable
- Cause low shear and turbulence
- Give pulsatile flow

The last design requirement listed is controversial. As pulsatile flow is more complex to achieve than steady flow, the use of a more sophisticated pump must be fully justified. The benefit of pulsatile flow remains a subject for debate. There is some suggestion that pulsatile flow is associated with an increase in O_2 consumption, reduced lactate accumulation and increased capillary blood flow to the brain.

Roller pumps are commonly employed. These have the advantages that the blood is only in contact with the tubing and little priming is required. However, there are disadvantages to the roller pump which does cause shear forces in the blood, will continue to appear to work against a blockage and causes stresses in the tube which may eventually crack.

The purpose of the heat exchanger is to control the blood temperature thus preventing progressive uncontrolled cooling. This is essential as abrupt temperature gradients result in cell damage and the release of gas from solution in the plasma. Filters are placed in the arterial return line and between the cannula used to clear the operative site and the oxygenator in order to remove particulate debris from the blood, thus preventing damage to the lungs, brain or kidney.

There is no ideal design of filter. If the pore size is too small the resistance of the circuit may rise as the filter blocks. In addition, the filter itself may cause blood damage. A typical design is made up of pleated polyester and has a pore size of about 40 μm.

22.3.4 *Haemodialysis, blood purification systems*

Our final example of a safety-critical system is that of haemodialysis. Dialysis is the removal of substances by means of diffusion through a membrane. Dialysis is used to replace the normal function of the kidneys in a patient with kidney failure. The loss of kidney function can be either acute or chronic. In acute renal failure, which can be caused by accident or disease, the kidneys will eventually recover their normal function. In the absence of dialysis the patient would die before the kidneys recovered. In chronic renal failure, the kidneys are permanently damaged and, in the absence of either a kidney transplant or regular dialysis, the patient will die.

Two types of dialysis are used. In peritoneal dialysis, the dialysing fluid is run into, and then out of, the patient's abdomen. This is a relatively simple technique that does not need either expensive equipment or access to the circulation, and it is used for certain patients with acute renal failure. Continuous ambulatory peritoneal dialysis (CAPD) has made peritoneal dialysis suitable for long-term use in chronic renal failure. In haemodialysis, blood is continuously removed from the patient, passed through an artificial kidney machine, and then returned to the patient.

Chronic renal failure patients who have not had a kidney transplant and who are selected as suitable for dialysis will be treated either by haemodialysis or peritoneal dialysis. Alternatively, a kidney can be removed from a live donor (usually a close relative) or from a person who has just died, and can be used to replace the kidneys in the chronic renal failure patient.

Chronic renal failure patients may be trained to use their own haemodialysis machine in their own home. This has many advantages. The risks of cross-infection are much reduced, because all the patients are effectively isolated from one another, the quality of the patient's life is improved, and the cost is reduced. In addition, the patient does not tie up expensive hospital facilities and staff, so that many more patients can be treated. It is worth emphasizing that this is a revolution in patient care as the patient is responsible for his own life-support system, and for doing many things that are usually the province of the doctor or nurse. Obviously the safety-critical aspects of the equipment design are extremely important.

The patient will need two or three dialysis sessions every week, each of several hours duration. The dialysis machine must always be working, and should it fail, it must tell the patient what is wrong. It must, in this situation, be repaired quickly—the patient's life depends on it. Most patients can manage three days without dialysis, but the machine must be repaired by the fourth day.

The function of the normal kidney

The two kidneys are bean-shaped organs, about 12 cm long and 150 g in weight. They are situated on the back of the abdominal wall. The top of the kidneys lies beneath the bottom two or three ribs and each contains about a million nephrons (figure 22.12). The nephron has two parts, the glomerulus and the tubule. The function of the glomeruli is to filter the plasma which is circulating in the capillary loops within Bowman's capsule. This is a passive process—it does not require any energy. The blood pressure in the capillary loops is about 60 mmHg (8 kPa), and about 25% of the cardiac output goes to the kidneys.

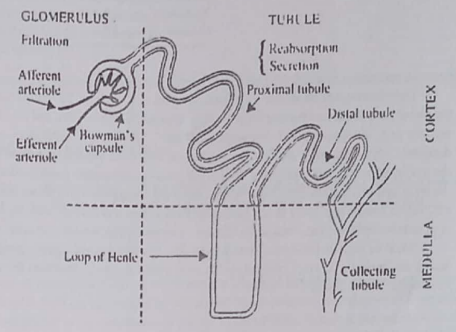


Figure 22.12. *Diagram of a single nephron. (Redrawn from A J Wing and M Magowan 1975 The Renal Unit (London: Macmillan).)*

The output of filtrate is about 1 ml/s/kidney, i.e. about 180 l day^{-1}. The total plasma volume is about 3 l, so that the plasma is filtered through the kidneys about 60 times a day. The filtrate then passes into the tubules. The total length of the tubules in each kidney is about 50 km. The tubules re-absorb electrolytes, glucose and most of the water, giving a total urine output about 1–2 l day^{-1}. This is an active process, which uses energy, and is continuously adjusted to maintain the correct fluid and electrolyte balance in the body.

The composition of the blood is very complex. The most important electrolytes are sodium, potassium, chloride, bicarbonate and calcium. The concentration of these electrolytes in normal plasma is given in table 22.5. The molecular weight of the substance, in grams, dissolved in 1 l of water, gives a concentration of 1 mol l^{-1}. The processes of filtration and re-absorption, together with secretion from the distal tubules, maintain the correct level of the electrolytes. Any departure of the electrolyte levels from normal will have an immediate effect on the health of the patient. If the serum sodium level is elevated, the patient's blood pressure will increase. Potassium changes the excitability of the cells, and an increase of serum potassium above 6 mmol l^{-1} can cause cardiac arrest without warning. An increased calcium level will cause the acid output of the stomach to increase, which can result in bleeding from peptic ulcers. A decrease in the calcium level will cause bone diseases. Most metabolic processes are very sensitive to the pH of the blood, which depends on the bicarbonate concentration.

Table 22.5. *The concentration of the more important electrolytes in normal plasma.*

	mmol l^{-1}
Sodium	132–142
Chloride	100
Bicarbonate	25
Potassium	3.9–5.0
Calcium	2.5

The electrolytes have low molecular weights (e.g. sodium 23, potassium 40). Organic chemicals in the blood have higher molecular weights (e.g. urea 60, bilirubin 600), and proteins have very high molecular weights (e.g. albumin 60 000, fibrinogen 400 000). Diffusion of substances across the dialysis membrane decreases with increasing molecular weight. Failure of the kidneys results in uraemia (retention of urea in the blood) and the concentration of urea in the plasma rises steeply from the normal level of 5 mmol l^{-1}. All the body systems are affected, both by the increasing concentration of waste products in the blood, and by the electrolyte imbalance. Urea is not particularly toxic, and it is thought that one important function of dialysis is to remove unidentified 'middle molecules' with molecular weights between 300 and 1500.

It can be seen that dialysis is much more than just the removal of waste products from the blood. Dialysis has to maintain the correct electrolyte balance within the body, maintain the correct pH of the blood, and control the fluid balance.

History of dialysis

Although the first haemodialysis in an animal was performed more than 90 years ago, the use of the technique in humans had to await the development of suitable membranes and anti-coagulants. Cellophane (cellulose acetate) membranes were developed during the 1920s and 1930s, and the commercial production of heparin started in the middle of the 1930s. The first haemodialysis on a human was performed by Kolff in 1943. He used a cellulose tube wrapped round a cylindrical drum which rotated in a bath of dialysis fluid.

The principle of haemodialysis remains essentially the same. Blood is withdrawn from the body and the patient is heparinized to prevent clotting of the blood in the extracorporeal circuit. The blood passes through an artificial kidney, in which it is separated from the dialysis fluid by a semi-permeable membrane. Electrolytes and small molecules pass freely through the pores in the membrane. The large molecules and the blood cells are unable to pass through the pores. Similarly, any bacteria present in the dialysis fluid are unable to pass into the blood.

- The method of operation of a cardiac defibrillator
- Safety-critical features of a haemodialysis system

The definition of a safety-critical device should relate to the consequences of failure of the device. Failure might be either a failure to function in a therapeutic device, or a failure to give the right answer in a diagnostic device where the consequences can be life threatening.

Here are some examples of safety-critical systems:

Aircraft	Surgical robots
Lifts	Telesurgery
Cars	Pacemakers
Heart valves	Defibrillators
Implants	Hydrocephalus shunts
Haemodialysis equipment	Infusion pumps
Computer software	Lasers
Linear accelerators	

We can list some of the questions we need to ask when considering the safety of a system:

What if it fails to function?
Does it fail safe?
What is the quality of the materials and components?
What are the constructional standards?
What are the maintenance requirements?
What margins of safety are allowed for?
What is the reliability?

22.2 CARDIAC ELECTRICAL SYSTEMS

We will continue our consideration of safety-critical systems by dealing with the design of two very common electromedical devices. The first is the cardiac pacemaker and the second the defibrillator.

22.2.1. *Cardiac pacemakers*

Electrical stimulators are widely used: physiotherapists use them in order to exercise muscles; anaesthetists test for muscular relaxation during surgery by observing the response to stimulation of a peripheral nerve. A growing use of peripheral nerve stimulators is for the relief of pain, although the mechanism for this effect is poorly understood. Cardiac pacemakers and defibrillators are used to stimulate cardiac muscle directly. The pacemaker corrects for abnormalities in heart rate, whereas cardiac defibrillators are used to restore a fibrillating heart to normal sinus rhythm. The way in which electricity interacts with tissue was covered in Chapter 8 and in section 10.2 of Chapter 10 neural stimulation was described.

Rhythmic contraction of the heart is maintained by action potentials which originate at a natural cardiac pacemaker. There are actually two pacemaker areas in the heart, the sinoatrial (SA) node and the atroventricular (AV) node, but the SA node normally dominates because it has the higher natural frequency. The normal course of events is that an impulse from the SA node is propagated through the myocardium, spreading over the atria to the AV node where there is a small delay before the impulse is conducted over the ventricles causing depolarization of the musculature. The normal cardiac rhythm is called sinus rhythm.

Any defects in conduction of the cardiac impulse can cause a change in the normal sinus rhythm and this is called an arrhythmia. Heart block occurs when the conduction system between atria and ventricles fails. This will not usually stop the heart because other pacemaking areas of the ventricles will take over or, if the blockage is not complete, some impulses may get through from the atria, but the heart rate will fall. This is called bradycardia (slow heart rate) and it may mean that the heart cannot supply the body's demands and so dizziness or loss of consciousness may occur.

There are three types of heart block:

- In first-degree block the delay at the AV junction is increased from the normal 0.1 to 0.2 s.
- In second-degree block some impulses fail to pass at all but a few get through to the ventricles.
- In complete block no impulses get through and so the ventricles pace themselves, but at a very much reduced heart rate of typically 40 beats per minute.

In all these cases an artificial pacemaker can be used to increase the heart rate to a level where the cardiac output is adequate to meet the body's needs.

The 'His bundle' which connects the atria and ventricles may be damaged by any one of several causes, such as poor blood supply to the bundle, ageing or accidental damage during surgery. Complete heart block can give rise to occasional fainting attacks and this is then described as Stokes–Adams syndrome. If left untreated the prognosis is poor, with 50% mortality after one year.

Brief history of pacemakers

The heart can be stimulated by applying an electrical current for a short time and then repeating this process about once every second. If the current is applied directly to the heart by a pair of wires the current required is only a few milliamperes and can be supplied by a battery. However, the first cardiac pacemakers used in the early 1950s were external devices which applied the current to cutaneous chest electrodes. They were effective but required quite high currents and were painful to the patient and could cause chest burns.

In 1958 the first implanted pacemakers were used. They used stainless steel electrodes sewn onto the myocardium and were implanted at the time of cardiac surgery. The process of implantation required major surgery with associated risks. A major improvement, first tried in about 1965, was to introduce electrodes into the cardiac chambers by passing them along a large vein. By passing the electrodes through the vena cava and right atrium into the right ventricle the need for major surgery was removed.

Very many technical improvements have been made to pacemakers since 1965 and this has improved the reliability and efficacy of the devices, but the basic principles have not changed. There are about 250 000 devices implanted worldwide each year so the financial implications are great.

Output requirements and electrodes

Internal pacemakers typically apply a rectangular pulse of 1 ms duration and amplitude 10 mA to the heart. As the resistance presented by the electrodes and tissue is about 500 Ω we can calculate the power requirement:

$$\text{power} = I^2R = 5 \times 10^{-3} = 50\ \text{mW}.$$

This is the power required during the pulse. As this pulse of 1 ms duration is repeated at about 1 s^{-1} (60 bpm) we can calculate the average power as 50 μW.

If we supply the pacemaker from a battery then we must know what output voltage we might require. In the above example the pulse amplitude is 5 V (10 mA × 500 Ω) so we could use a 5 V battery. The capacity of a battery is often quoted in ampere hours (A h) and a typical small cell will have a capacity of 1 A h. Now we can calculate the average output current in the above example as 10 μA as the 10 mA flows for one thousandth of each second. Our battery could supply this current for $1/10^{-5}$ h, i.e. about 11 years.

This 'back of the envelope calculation' we have just made shows that we can hope to power a pacemaker for a very long time from a small battery. Our calculation did not take into account the power requirements of the circuitry of the pacemaker, but even if we allow another 10 μA for this the battery life should still be more than 5 years. The output pulses have to be applied to the tissue through an electrode: actually two electrodes as a circuit has to be produced. The cathode is the electrode placed in the ventricle (see section 10.2) as this is where the stimulation is required. Very many types of electrode have been used for pacemakers using metals such as platinum, silver, stainless steel, titanium as well as various alloys. Carbon electrodes have also been used. The subject of electrodes is complex and too difficult to treat here. The electrode can be thought of as a transducer because it has to allow electrons flowing in the pacemaker wires to give rise to ionic flow in the tissue (see section 9.2). This reaction must be stable and non-toxic. Electrodes vary in area between about 10 and 100 mm^2. In all cases a capacitor is inserted in series with the output of the pulse generator in order to prevent any accidental application of DC to the electrodes and also to ensure that in normal operation there is no net current flow into the tissue and so no electrolysis.

Types of pacemaker

A pacemaker must consist of a pulse generator and electrodes. Most pacemakers are implanted or internal types where the entire device is inside the body. However, external pacemakers also exist where the pulse generator is external to the body and the electrodes are located either on or within the myocardium.

External pacemakers can be used on patients with a temporary heart arrhythmia that might occur in critical post-operative periods or during cardiac surgery.

Internal pacemakers are implanted, with the pulse generator put in a surgical pouch often below the left or right clavicle (figure 22.1). The internal leads may then pass into the heart through the cephalic vein.

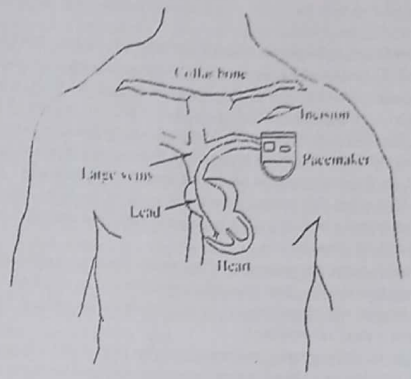


Figure 22.1. *The transvenous lead from the pacemaker enters the subclavian vein and is guided under x-ray control into the heart. The pacemaker is installed in a subcutaneous pouch.*

The simplest type of pacemaker produces a continuous stream of output pulses at about 70 bpm. There are many disadvantages to this simple approach, one being that the heart rate will not vary in response to what the patient is doing. Another disadvantage is that power may be wasted if heart block is not complete because

some beats could occur naturally without the pacemaker. In addition this competition between naturally occurring beats and the pacemaker output may not lead to the most effective cardiac performance. These disadvantages to fixed-rate pacing (often referred to as competitive or asynchronous pacing) have led to the development of a range of more complex devices. These are illustrated in figure 22.2.

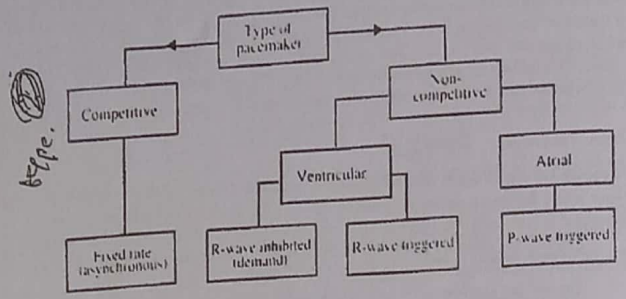


Figure 22.2. *Categories of pacemaker*

The alternative to fixed-rate or competitive pacing is non-competitive pacing. In this case the pacemaker records the ECG produced by the heart and produces an output in response to this signal. Non-competitive types can be subdivided into ventricular- and atrial-triggered devices (see figure 22.2). Atrial-triggered devices produce an output triggered by the *P*-wave of the ECG (see section 16.2) which is generated by the atria and will not be affected by the heart block. Ventricular-triggered devices use the *R*-wave of the ECG in one of two ways. In a demand-type pacemaker an output pulse is only produced in the absence of a naturally occurring R-wave, i.e. the R-wave is used to inhibit the output of the pacemaker. If the pulse rate falls below the pre-set rate of the pacemaker then output pulses will again be produced. However, in a standby R-wave triggered device an output pulse is produced in response to every R-wave and if one does not occur when expected then the pacemaker will generate one.

For testing purposes at the time of implantation and to enable checks, such as the state of the battery, to be carried out periodically most demand pacemakers can be set into a fixed-rate mode. This is often done by means of a magnet which actuates a reed relay inside the pacemaker. The magnet is placed on the skin above where the pacemaker is implanted.

Ventricular demand-type pacemakers are the most commonly used. However, atrial-triggered devices are used in complete heart block where the normal vagal and hormonal control of the atria allows heart rate to vary in response to demand. In order to record the P-wave electrodes have to be placed in the atria in addition to the pacing electrodes in the ventricles. A block diagram of a typical pacemaker is shown in figure 22.3. The timing circuit, which consists of an *RC* network, reference voltage and comparator, determines the basic pacing rate of the pulse generator. The pulse width circuit determines the output pulse duration and a third *RC* network gives a delay to limit the maximum rate of pulses. The sensing circuit is inhibited during the output pulse in order to stop overload of the circuitry. The sensing circuit detects a spontaneous R-wave and resets the timing capacitor so no output is produced. The voltage monitor has two functions. One is to detect a low battery voltage and use this to reduce the fixed-rate output as a means of signalling a low battery. The other is to increase the output pulse width when the battery voltage falls so that the output pulse energy remains constant.

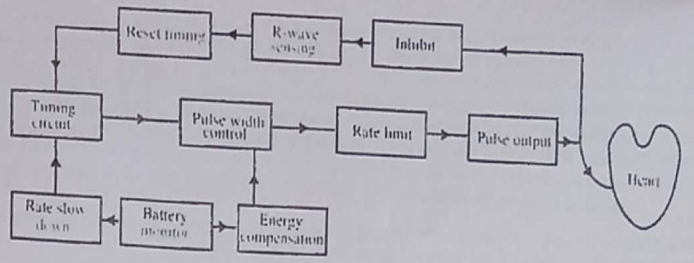


Figure 22.3. *Block diagram of a typical pacemaker*

There are agreed international codes for the description of pacemakers. These classify devices in terms of the heart chamber which is paced, the heart chamber which is sensed and what type of response is made, i.e. inhibition or triggering. However, the advent of pacemakers with microprocessors and the possibility of very complex algorithms to determine the output of the pacemakers have made it very difficult to describe devices by means of a simple code.

Power sources

It is essential that the pacemaker has a reliable power source which will last as long as possible. It is possible to replace an exhausted pacemaker but surgery is necessary and battery failure if it is not detected at an early stage can give rise to a life-threatening situation. The ideal power source should be very small, have a long life, i.e. a high capacity, be unaffected by body temperature, be easy to test so that exhaustion can be predicted, be cheap, be unaffected by autoclaving and give an output of at least 5 V. No gases must be produced. Types which have been used include mercury cells, nuclear-powered thermoelectric generators and lithium cells.

Many types of power source have been proposed and several used in pacemakers. The most exotic is the nuclear-powered pacemaker which contains a capsule of plutonium (^{238}Pu) which has a half-life of 87 years. This radioactive isotope emits α-particles which generate heat when they are absorbed within the pacemaker; the heat can then be used to generate electricity from a large number of thermocouples which form a thermopile. This type of power source is actually very reliable and has a life of more than 20 years, but there are many disadvantages, so that whilst some remain implanted new ones are not used. The disadvantages are firstly, that plutonium is very toxic so that the encapsulation has to ensure that there is zero possibility of leakage and secondly, that the cost is high. Plutonium is so toxic that cremation of a device would be a significant hazard so that strict control has to be exercised over the disposal of nuclear-powered devices. The radiation dose to the patient is actually very small.

For many years mercury cells were widely used for powering pacemakers but unfortunately their life was only about 2 years. A life of at least 5 years had been anticipated but the deterioration within the body proved to be more rapid than anticipated so that by about 1970 the mercury batteries were the most severe constraint on pacemaker life. In the early days lead failure and problems of encapsulation of the electronics had been the major problems but these had been largely overcome by 1970. People then tried all sorts of alternatives such as piezoelectric devices, self-winding watch-type mechanisms, fuel cells, rechargeable cells and electrochemical generation from body fluids. None of these were very successful.

The battery most commonly used today is the lithium iodine cell. This offers a much longer life than the mercury cell and has the advantage that no gases are given off so that it can be hermetically sealed. It

provides a higher voltage output than the 1.35 V of the mercury cell, as shown in figure 22.4. The stable discharge curve is an advantage in that the output of the pacemaker will not change during the life of the battery. However, it is also a disadvantage in that it is quite difficult to assess when a battery is coming to the end of its safe life. It is for this reason that quite complex procedures have to be used to check the batteries, including switching the pacemaker into fixed-rate mode so that the drain on the battery is constant.

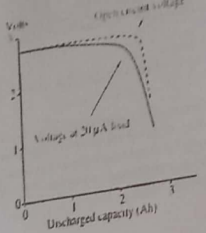


Figure 22.4. Typical discharge curve of a lithium iodine cell

Safety requirements

Many of the requirements for safe operation of a pacemaker have already been mentioned. In the design of the pulse generator attempts are made to ensure that under any single-fault condition the generator will revert to a fixed-rate mode. Attention has also to be given to the effect of electromagnetic interference (see sections 10.4 and 22.2.2) on the performance of demand-type pacemakers. Patients with pacemakers are warned not to expose themselves to the outputs of powerful radio transmitters and devices such as airport security checking devices. They should also avoid magnetic resonance scanners which produce high magnetic and electromagnetic fields. However, inadvertent exposure is possible so that the design of the pacemaker should be such that it will revert to a fixed-rate mode when interference is present.

Hazards do not always concern the electrical output of the pacemaker. The leads from the pacemaker to the heart can also malfunction. There can also be problems of compatibility between the pacemaker encapsulation and body tissues (see Chapter 4) and of course sterilization is necessary before implantation of the pacemaker.

Five methods have been used for sterilization of pacemakers:

- Cold chemicals
- Radiation
- Ethylene oxide gas
- Steam, 120 °C for 15 min is commonly used
- Dry heat at about 150 °C.

Radiation can affect the electronics of a pacemaker, particularly MOSFET devices. Autoclaving at 120–130 °C and dry heat will damage most cells and some electronic components. This leaves gas sterilization and cold chemicals. Of these gas sterilization is best but it is not always available.

The technology of pacemakers has advanced rapidly over the past 40 years with advances in power sources, biomaterials, sensors and computing. Many pacemakers now incorporate rate-responsive and can be regarded as intelligent implants. These can sense a change from normal cardiac performance and select a suitable therapy which may be a type of pacing, but also perhaps a defibrillation pulse in reverse or a therapeutic drug. Physiological variables such as temperature, pH, Po_2, Pco_2, respiration, glucose or blood pressure can be sensed and used by the implanted microprocessor to select the appropriate therapeutic action.

22.2.2. *Electromagnetic compatibility*

Many items of medical equipment now use quite sophisticated electronic control systems. In principle these can both be affected by external electromagnetic fields and also produce such interfering fields. Pacemakers can obviously be a hazard to the patient if they are affected by interference. These problems are covered by national and international agreements on electromagnetic compatibility (EMC). In Chapter 10 we discussed how various types of interference can affect bioelectric measurements. However, a whole range of electromedical devices may be affected by electromagnetic interference.

Examples of equipment which might be affected by external fields:

Infusion pumps	Computers
Patient monitoring equipment	Linear accelerators
Demand pacemakers	EMG/EEG/ECG
Surgical robots	Hearing aids
Defibrillators	Software
Lasers	

Equipment which might produce interference includes:

Computers	Radio communications
Mobile phones	Thermostats
Surgical diathermy/electrosurgery	Electrostatic materials
Linear accelerators	Transformers
Physiotherapy diathermy	

EMC is a growing problem which might be tackled in various ways:

Design of the equipment
Earthing/shielding/filtering/separation
Design of the hospital

In the case of pacemakers there is an intrinsic difference in the susceptibility of unipolar and bipolar devices to interference. The bipolar device is less likely to be affected by interference because the electrodes are close together and hence require a high field gradient to produce an interfering voltage.

22.2.3. *Defibrillators*

Our second detailed example of a safety-critical system is the defibrillator. Defibrillators are devices that are used to apply a large electric shock to the heart. They are used to restore a normal sinus rhythm to a heart which is still active but not contracting in a co-ordinated fashion. The cause of fibrillation is commonly ischaemia of heart tissue but less common causes are electric shock, drugs, electrolyte disorders, drowning and hypothermia. The use of a defibrillator on a patient following a heart attack is an emergency procedure, as the pumping action of the heart has to be restarted within a few minutes if the patient is to survive. The defibrillator is therefore a 'safety-critical' device; if it fails to work when required then the patient will die.

Defibrillators have a long history in that some animal work was done in 1899, but emergency human defibrillation was not carried out until the 1950s. They have been in widespread use since the 1960s.

Fibrillation

Cardiac muscle is intrinsically active in that it will contract periodically in the absence of any neural connections. If a piece of cardiac muscle is removed from the heart and placed in an organ bath then oscillating electrical potentials can be recorded from the piece of tissue. Now in the intact heart all the pieces of cardiac muscle interact with each other such that they all oscillate at the same frequency and so contract regularly at that frequency. However, if part of the heart muscle is damaged or disrupted then the interaction can be disrupted and fibrillation can occur. All the pieces of cardiac muscle are still oscillating but at different frequencies so that there is no co-ordinated contraction.

Either the ventricles or the atria can fibrillate but the consequences to the patient are very different. Under atrial fibrillation the ventricles still function but with an irregular rhythm. Because atrial filling with blood does not depend upon atrial contraction there is still blood for the ventricles to pump, so that whilst the patient may be aware of the very irregular heart beat blood is still circulating. Ventricular fibrillation is much more dangerous as the ventricles are unable to pump blood so that death will occur within a few minutes. Ventricular fibrillation is not self-correcting so that patients at risk of ventricular fibrillation have to be monitored continuously and defibrillation equipment must be immediately to hand.

Fibrillation is obvious in the resulting ECG as shown. Figure 22.5(*a*) shows a normal ECG and figure 22.5(*b*) the result of cardiac fibrillation. There are still potential changes during fibrillation but they are apparently random and the amplitude is less than occurs during normal sinus rhythm. Care has to be taken in recognizing fibrillation from the ECG as other conditions may change the appearance of the ECG waveform. Figures 22.5(*c*)–(*e*) show the appearance of atrial fibrillation, atrial flutter and ventricular tachycardia.

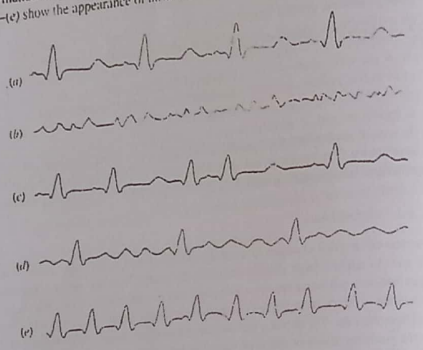


Figure 22.5. *Diagrams of: (a) a normal ECG; (b) fibrillation; (c) atrial fibrillation; (d) atrial flutter; (e) ventricular tachycardia.*

A defibrillation shock aims to totally stop the heart, or at least to stop enough cells to inhibit fibrillation, in the hope that the heart will restart in an orderly fashion after the shock. If sufficient current is used to stimulate all the musculature of the heart then when the shock stops all the heart muscle fibres enter their refractory period at the same time, after which normal heart rhythm may resume.

Of the two major categories of muscle, striated (skeletal) and smooth, cardiac muscle is most like smooth muscle. Skeletal muscle is composed of long muscle fibres and does not produce activity unless it receives impulses along an associated motor nerve. Smooth muscle is composed of much shorter and smaller cells and it is intrinsically active. Neural impulses and circulating hormones will influence the activity but smooth muscle can contract in isolation. Cardiac muscle is intrinsically active.

Skeletal muscle can be electrically stimulated and will produce an action potential in response to a stimulus of duration 100 μs or less (see figure 10.4). Cardiac muscle is more difficult to stimulate, in part because it is intrinsically active so that our ability to stimulate the muscle depends upon what the muscle is doing at the time we apply our stimulus. Longer duration pulses are required to stimulate cardiac muscle than striated muscle, although high-amplitude stimuli of short duration will have the same effect as lower amplitude impulses of longer duration. This is illustrated in figure 22.6 in the form of a curve often referred to as a strength–duration curve. The 'current' curve shows that the amplitude of current required for defibrillation decreases as duration increases, but that there is a minimum current required whatever the stimulus duration. This is called the rheobase. Take note that the current required is several amps, which is a very high current to apply to the human body.

Figure 22.6 also shows the charge (Q) and energy (E) associated with the current pulse. The charge is the product of the current I and the pulse duration D. The energy is I^2RD where R is the resistance into which the current is delivered. It can be seen that there is a minimum energy required at a pulse width of about 4 ms.

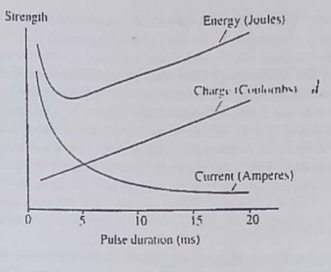


Figure 22.6. *A strength–duration curve for defibrillation applied across the thorax. Curves showing the associated charge and energy are also shown.*

If $E(\Delta) = \dfrac{V_{rms}}{51.0 \times \Delta P}$

where a and b are constants, then $I = \dfrac{a}{D} + b$

and $\text{energy}(E) = \left(\dfrac{a}{D} + b\right)^2 RD = \left(\dfrac{a^2}{D^2} + b^2 + \dfrac{2ab}{D}\right) RD$

$$\frac{dE}{dD} = -\frac{a^2 R}{D^2} + b^2 R$$

This will be zero at the turning point in E and hence

$$\frac{a^2}{D^2} = b^2 \quad \text{therefore} \quad D = \pm\frac{a}{b} \quad \text{and} \quad I = 2b$$

The minimum energy requirement will be for a duration such that the current is double the rheobase.

One reason for the large current which is required for defibrillation is that only a small fraction of the current applied to the chest will actually flow through the heart. If defibrillation is applied directly to the heart during open heart surgery then smaller currents are required.

Another important variable related to the current and pulse duration required for defibrillation is the size of the subject. It has been shown that larger animals require higher currents to defibrillate than smaller ones. It is also true that larger animals are more likely to suffer from fibrillation than smaller ones. Large people require higher defibrillation currents than small people and, in particular, children require low currents.

Pulse shapes The stimulus waveform which is used in most striated muscle stimulators, such as those used in physiotherapy, is rectangular. However, defibrillators produce more complex waveforms. The reason for this is partly physiological, but it is also related to the way in which the impulses are generated. It is actually quite difficult to generate a very high current rectangular pulse, whereas it is relatively easy to charge up a capacitor to a high voltage and then discharge it through the patient. However, the current waveform then produced has a high peak and a long exponential tail. There is evidence that the long exponential tail can refibrillate the heart and so reverse the defibrillation. For this reason a damped exponential current waveform is used as illustrated in figure 22.7. The pulse width is about 4 ms which corresponds approximately to the minimum energy requirement which was shown in figure 22.6. In the next section we will consider how waveforms of the shape shown in figure 22.7 can be generated.

Design of defibrillators

About 50 A of current is required for defibrillation across the chest. Now if we know the resistance into which this current has to pass we can calculate the power requirement. The resistance will depend upon the size of electrodes used and the size of the patient but a typical figure is 50 Ω.

$$\text{power required for defibrillation} = I^2 R = 125\ \text{kW}$$

This is a large amount of power. The first defibrillators in clinical use were called AC types and they simply used the mains supply fed through a transformer and a switch for defibrillation. Unfortunately the maximum power which can be drawn from most mains supply outlet sockets is about 15 kW which is just about sufficient to defibrillate a child but not an adult.

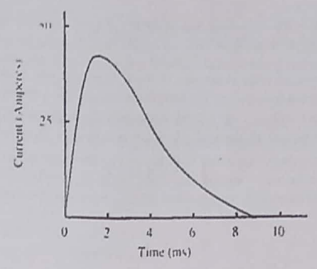


Figure 22.7. *Output waveform from a DC defibrillator.*

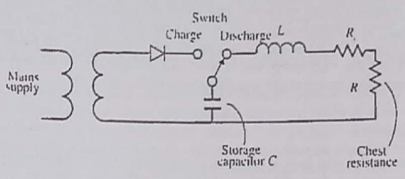


Figure 22.8. *Basic circuit diagram of a DC defibrillator.*

The solution to the problem of limited power availability from a mains supply socket is the DC defibrillator. This uses a large capacitor, which is charged to a high voltage from the mains supply and then discharged rapidly through the patient. A circuit diagram is given in figure 22.8, where the storage capacitor is marked as C. We can determine the necessary size of this capacitor relatively easily.

The voltage V we require is at least 2500 V to produce a current of 50 A into 50 Ω.

The duration of the discharge must be about 4 ms, so the time constant formed by C and R must be about 4 ms. For R equal to 50 Ω this gives C as 80 μF.

The *energy* stored in this capacitor when charged to 2500 V is given by $\frac{1}{2}CV^2$ which is 250 J. In fact, most defibrillators have a maximum stored energy of 400 J, although the maximum output is only required on large patients.

In the previous section on pulse shapes it was argued that a pulse output with a long tail is undesirable because this can cause refibrillation. In order to remove the long tail of the capacitor discharge an inductor L is usually added in series with the defibrillator output as shown in figure 22.8.

We can calculate the shape of the output current as follows. First, by applying Kirchhoff's law we obtain

$$L\frac{d^2 i}{dt^2} + (R_1 + R)\frac{di}{dt} + \frac{i}{C} = 0 \qquad (22.1)$$

where i is the current, L the inductance, C the storage capacitor, R the resistance presented by the patient and R_i is the internal resistance of the defibrillator circuit. The solution of this differential equation has three forms, the first is oscillatory, the second an aperiodic discharge and the third, the fastest aperiodic discharge when the circuit is critically damped. The solution in this case is given by

$$i = \frac{CV(R+R_i)}{4L^2} te^{-\frac{(R+R_i)}{2L}t} \quad (22.2)$$

and

$$L = (R+R_i)^2 \frac{C}{4} \quad (22.3)$$

This gives a typical value for L of 40 mH.

Stored and delivered energy. Not all the energy stored in the capacitor will be delivered into the chest of the patient because of losses within the output circuit of the defibrillator. The main source of loss is the resistance of the inductor and this is represented as R_i in figure 22.8. Now the *stored energy* in the capacitor charged to voltage V is $\frac{1}{2}CV^2$, but is easy to show that the *delivered energy* to the patient will be $\frac{1}{2}CV^2 \times R/(R+R_i)$. Typically the delivered energy will be about 10% less than the stored energy so that a defibrillator of 400 J maximum stored energy might actually deliver a maximum of 360 J.

Electrodes

Obviously, electrodes have to be placed on the thorax in the best position to ensure that current will flow through the myocardium. If the electrodes have too high an impedance (see section 9.2.3) or are not optimally placed then an unnecessarily high pulse output will be required for defibrillation. The energy output from a defibrillator is sufficient to cause tissue burns so these might also arise if electrode contact impedance is too high.

Chest electrodes and direct heart electrodes usually consist of bare metal electrodes made of a non-corrosive material. Very often output switches are incorporated into the back of the electrodes for easy use by the operator. Typical chest electrode size is 10 cm diameter. Direct heart electrodes are usually smaller and look a bit like spoons on the end of handles so they can be placed in contact with the myocardium. Chest electrodes are usually used with a conductive jelly to reduce contact impedance to about 50–100 Ω. If jelly is not used then burns can occur and defibrillation will be less effective.

When electrodes are placed on the chest, they are usually applied with both electrodes on the anterior chest wall. One electrode is placed over the apex of the heart, which is about the fifth intercostal space in the midclavicular line on the left side of the chest. The other is placed in the second intercostal space adjacent to the sternum.

Use of defibrillators

There are basically three uses for defibrillators:

- The first is direct defibrillation of the heart during surgery. During cardiac surgery the heart may spontaneously fibrillate or the surgeon may intentionally produce fibrillation. The maximum output used to defibrillate directly is about 50 J.
- The second major use of defibrillators is for cardioversion. This is a synchronized shock which is applied across the chest to correct for atrial fibrillation, atrial flutter or ventricular tachycardia. Energies from about 20–200 J are used for cardioversion.
- The third use for defibrillators is emergency defibrillation in cases of ventricular fibrillation. Cardiopulmonary resuscitation is often used to keep the patient alive until the defibrillator is ready for use. Often a first pulse of 200 J will be followed by higher shocks if the first is not successful in restoring a normal sinus rhythm.

Some defibrillators are battery powered and sufficiently small to be portable. Many have ECG/EKG monitoring included and other features, such as impedance detection circuitry to detect poor electrode contact, recording of the actual delivered energy and various alarms to detect dangerous conditions, are also included. Most units have the ECG/EKG monitoring facility and a synchronization circuit, to enable the operator to be certain of the diagnosis. Obviously the outcome to be avoided is the precipitation of fibrillation in a normal heart.

Implanted defibrillators

Totally automatic implantable defibrillators have been developed in the past few years. These are like pacemakers, but treat tachycardias rather than bradycardias and can defibrillate with a pulse of up to about 30 J. Both ventricular fibrillation and ventricular tachycardia are treated, because of the rapid fatality of ventricular fibrillation and the high frequency with which ventricular tachycardia can develop into ventricular fibrillation. These devices have been found to greatly reduce mortality in patients known to be at high risk from sudden cardiac death.

The implanted defibrillator contains sensors to detect cardiac electrical activity in the same way as demand pacemakers and signal processing in order to make the correct decision as to when treatment is required. This is obviously a safety-critical system. Implanted defibrillators contain ECG/EKG sensing leads, a pulse generator, signal processing electronics and electrodes which are placed on the epicardium or in the right ventricle. Obviously such a device makes great demands upon the battery power supply which has to supply large output pulses without impairing its ability to supply the decision making electronics. Current devices are able to supply about 100 shocks of 25 J before the battery is exhausted.

22.3 MECHANICAL AND ELECTROMECHANICAL SYSTEMS

We continue our examples of safety-critical systems by considering some mechanical and electromechanical medical devices.

Disease can be addressed in one of two ways:

- by returning a malfunctioning organ to health using chemical or physical agents, or
- by substituting a functional counterpart, either after removing the sick organ, or ignoring it entirely.

Beyond a certain stage of failure, it is often more effective to replace a malfunctioning organ than seeking in vain to cure it. This has given rise to 'spare parts medicine' where a whole range of implanted artificial organs have been developed. These implants must often be considered as *safety-critical systems*.

The use of implants gives rise to very many complex issues. These concern medical, social, managerial, economic, legal, cultural and political aspects. We will only consider the engineering design aspects. Table 22.1 lists some of the systems which have been developed over the past few decades. We will just consider the design aspects of three of these developments. The first is heart valve substitutes, the second cardiopulmonary bypass and the third haemodialysis systems.

Table 22.1. *Current status of organ replacement technology. (Adapted from The Biomedical Engineering Handbook ed J D Bronzino (Boca Raton FL: Chemical Rubber Company).)*

Clinical standing	Artificial organ	Transplantation
Generally accepted	Heart-lung machine Large joint prostheses Bone fixation systems Cardiac pacemakers Large diameter vascular grafts Prosthetic heart valves Intra-aortic balloon pumps Implantable lenses Hydrocephalus shunts Dental implants Skin or tissue expanders	Blood transfusion Corneal transplants Banked bone Bone marrow Kidney, cadaveric donor Heart Liver Heart/lung
Accepted with reservations	Maintenance haemodialysis Chronic ambulatory Peritoneal dialysis Breast implants Sexual prostheses Small joint prostheses Extracorporeal membrane Oxygenation in children Cochlea prostheses	Kidney, living related donor Whole pancreas
Limited application	Implantable defibrillator ECMO in adults Ventricular-assist devices Artificial tendons Artificial skin Artificial limbs	Pancreatic islets Liver lobe or segment Cardiomyoplasty
Experimental	Artificial pancreas Artificial blood Intravenous oxygenation Nerve guidance channels Total artificial heart	Gene transfer Embryonic neural tissue Bioartificial pancreas Bioartificial liver
Conceptual stage	Artificial eye Neurostimulator Blood pressure regulator Implantable lung Artificial trachea Artificial oesophagus Artificial gut Artificial fallopian tube	Striated and cardiac muscle Functional brain implants Bioartificial kidney

22.3.1. *Artificial heart valves*

How do natural heart valves work?

In the normal human heart, the valves maintain a unidirectional flow of blood with minimal frictional resistance, whilst almost completely preventing reverse flow. They act passively; the moving parts of the valve, the tissue

Nandhini Na

Some sound level meters include a slow/fast response switch. In the 'slow' position the meter response is slowed by using a low-pass filter which will make the meter less subject to sudden changes in sound level. The slow position allows a more accurate measurement of the average noise level to be obtained than the fast position.

15.2.6. *Normal sound levels*

Table 15.1 gives the sound pressure levels both in pascals and in decibels, corresponding to nine circumstances.

Damage to the ear occurs immediately for sound levels of about 160 dB. Normal atmospheric pressure is about 10^5 Pa and 160 dB is 2×10^3 Pa so that damage occurs at about 0.02 atm. The threshold of hearing is the other extreme. This pressure represents 2×10^{-10} atm; if we were to measure this pressure with a mercury manometer then the mercury level would only change by 1.5×10^{-10} m.

The range of sound levels which are encountered in normal living is very wide, although there has been increasing pressure in recent years to limit the maximum sound levels to which people are exposed. There is no international agreement on standards for occupational exposure but most of the developed countries have adopted a limit of 90 dB for continuous exposure over a normal 8 h working day, with higher levels allowed for short periods of time. In some countries the level is set below 90 dB.

In a room where hearing tests are carried out, the background noise level should not be greater than 40 dBA and a level below 30 dBA is preferred. Lower noise levels are needed if 'free field' testing is to be carried out (see ISO 8253-2 for more detailed guidance). The use of sound-reducing material in the walls, floor and ceiling of the audiology test room is often necessary. Noise-reducing headsets are a cheap way of reducing the background noise level for a patient.

Table 15.1. *Nine typical sound pressure levels, expressed on the dBA scale.*

Sound pressure ($N m^{-2}$ = Pa)	Sound pressure level (dBA)	Circumstances
2×10^3	160	Mechanical damage to the ear perhaps caused by an explosion
2×10^2	140	Pain threshold, aircraft at take-off
2×10^1	120	Very loud music, discomfort, hearing loss after prolonged exposure
2×10^0	100	Factory noise, near pneumatic drill
2×10^{-1}	80	School classroom, loud radio, inside a car
2×10^{-2}	60	Level of normal speech
2×10^{-3}	40	Average living room
2×10^{-4}	20	Very quiet room
2×10^{-5}	0	Threshold of hearing

15.3. BASIC MEASUREMENTS OF EAR FUNCTION

A measure of speech comprehension is the most desirable feature of a hearing test. Tests are used in which speech is presented to the subject at a range of intensities and their ability to understand is recorded. Speech audiometry is a valuable test of hearing, although the results depend not only on the hearing ability of the subject but also upon their ability to comprehend the language which is used. Sounds other than speech are

also used: a tuning fork can be used by a trained person to assess hearing quite accurately. Sources of sound such as rattles are often used to test a child's hearing: the sound level required to distract the child can be used as evidence of their having heard a sound.

In this section an account is given of some commonly used hearing tests. In pure-tone audiometry a range of pure tones is produced and the subject is asked whether they can hear the sounds. Middle-ear impedance audiometry is another type of hearing test which enables an objective measurement to be made of the mechanical function of the middle ear. Otoacoustic emissions are a further objective measure of hearing but the origin of these emissions is not well understood.

15.3.1 *Pure-tone audiometry: air conduction*

The pure-tone audiometer is an instrument which produces sounds, in the form of pure tones, which can be varied both in frequency and intensity. They are presented to the patient either through headphones for air conduction measurements, or through a bone conductor for bone conduction measurements. In the test situation the patient is instructed to listen carefully and respond to every sound. This response may be to raise a finger or to press a button; if the patient is a child then they may be asked to respond by moving bricks or some other toy when they hear the sounds. The threshold level is said to be the minimum intensity at which the tone can be heard on at least 50% of its presentations.

The audiometer contains an oscillator which produces a sinusoidal waveform. The frequency of this sine wave can be changed, and the minimum available frequencies are 250, 500, 1000, 2000, 4000 and 8000 Hz. The output from the oscillator is taken to an audio amplifier and then into an attenuator, which may be either stepped or continuously variable. A standard range would be from −10 to 120 dB. The output from the attenuator is taken to the headphones. The input connection to the amplifier can be interrupted by a switch which allows the sound to be presented as bursts of a pure tone. This is the most common way of presenting sounds for manual audiometry.

A loud sound is presented to the patient and the intensity reduced slowly until they can no longer hear the sound. The threshold found by this method will not be the same as that which is found if the sound intensity is increased slowly from zero to the point where it can first be heard. For this reason it is important that a consistent test procedure be adopted. There is no universal agreement on the procedure to be adopted in pure-tone audiometry, but the following is a widely used system. The sounds are initially presented in decreasing intensity and then both upward and downward changes are made close to the threshold to determine the level at which 50% of the sounds are heard.

Procedure for routine air conduction pure-tone audiometry

Place the headphones comfortably on the patient, making sure that the red phone is over the right ear. Spectacles can be most uncomfortable when headphones are worn and are therefore best removed.

Start at a level of 50 dB and 1000 Hz.

Present tones of about 2 s in duration with varying intervals (1–3 s).

If the tone is heard, then reduce the level in 10 dB steps until it is no longer heard. If the starting tone is not heard, then raise the level in 20 dB steps until it is heard, and then descend in 10 dB steps.

From the first level at which the tone is not heard, first raise the level in 5 dB steps until it is heard, then down in 10 dB steps until it is not heard, then up again in 5 dB steps. This enables two ascending threshold measurements to be made.

After testing at 1000 Hz proceed to 2000, 4000 and 8000 Hz. Repeat the reading at 1000 Hz and then make measurements at 500, 250 and 125 Hz.

Great care must be taken to vary the interval between the tones in order to detect where incorrect responses are given.

15.3.2 *Pure-tone audiometry: bone conduction*

Instead of presenting the sound vibrations through headphones a vibrator can be attached over the mastoid bone behind the ear. The vibrator is usually attached by a sprung band passing over the head. Sounds presented by this means bypass the eardrum and middle ear and are able to stimulate the inner ear directly. A patient with disease of the middle ear, such that sounds are attenuated in passing through the middle ear, may have a raised threshold to sound presented through headphones but a normal threshold to sound presented through the bone conductor.

The procedure for making a threshold determination through a bone conductor is the same as that which was described for air conduction. The results of both air and bone conduction threshold measurements are presented graphically as shown in figure 15.8. Different symbols are used for the right and left ears and also for air and bone conduction thresholds.

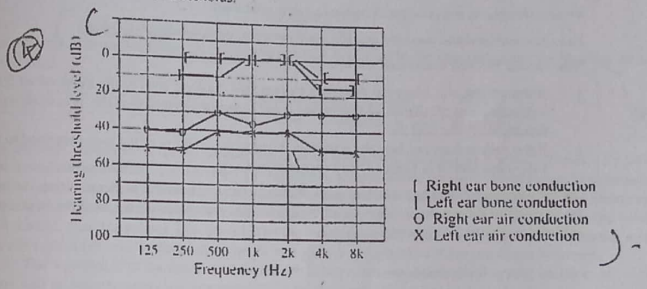


Figure 15.8. *This pure-tone audiogram shows the variations in hearing level for both air- and bone-conducted sounds. The patient has normal bone conduction thresholds but a 40 dB loss for air conduction.*

15.3.3 *Masking*

A hearing loss which only affects one ear is called unilateral and a loss to both ears, but of different degrees, is called asymmetrical. If a patient has a much greater hearing loss in one ear than the other then it is possible that sounds presented to the poor ear may be heard in the good ear. 40 dB is the minimum reduction in intensity of a sound presented to one ear but heard in the other ear. If the difference in pure-tone thresholds between the two ears is greater than 40 dB then special techniques have to be used in testing.

In order to obtain a 'true' threshold when testing the poor ear, a masking noise is presented to the good ear to prevent cross-over. The masking noise of choice is narrow-band noise; this is a random noise such as the hiss which is produced by a high-gain audio amplifier, but filtered to present a one-third octave band of noise centred on the test frequency.

The criteria to assess when masking is needed, are

- where the difference between left and right unmasked air conduction thresholds is 40 dB or more, and
- where the unmasked bone conduction threshold is at least 10 dB better than the worst air conduction threshold

This is necessary because sounds are conducted through the skull with very little loss of intensity, a sound presented through the mastoid bone on one side of the head can be heard at the same intensity on the other side

Procedure for masking

Measure the threshold of the masking signal in the non-test ear

Present the tone to the poor ear at unmasked threshold level.

Introduce narrow-band masking into the good ear at the masking signal threshold.

Now present the tone to the poor ear again:

1 If the patient still hears the tone then increase the masking level to the good ear in 5 dB steps up to a maximum of 30 dB above threshold. If the tone is still heard then this is considered to be the true threshold for the poor ear.

2 If the patient does not hear the tone then increase the intensity of the tone presented to the poor ear in 5 dB steps until it is heard. Then proceed as in 1.

The test is not considered satisfactory until the tone in the poor ear can be heard for an increase of 30 dB in the masking to the good ear. (See the reference to 'Recommendations for masking in pure-tone threshold audiometry', in the bibliography.)

15.3.4. *Accuracy of measurement*

Pure-tone audiometry gives a measure of hearing threshold over a range of sound frequencies. However, the measurement is a subjective one because it depends upon the co-operation of the patient and their ability to decide when a sound can be heard. Hearing threshold will vary amongst a group of normal people; it can also change from day to day and is affected by exposure to loud sounds. For these reasons a range of −10 to +15 dB is normally allowed before a threshold measurement is considered to be abnormal.

Very many factors can contribute to inaccuracies in measurement but only a few can be mentioned here. These factors can arise either from the equipment or from the operator.

Pure-tone audiometry equipment should be calibrated at least once a year using an artificial ear; this is a model of an ear with a microphone included so that the actual sound level produced by headphones can be measured. In addition, a routine weekly test of an audiometer should be made by the operator by testing his or her own hearing. If the threshold readings change by more than 5 dB and there is no reason for their hearing to have been affected, then the audiometer is probably at fault and should be recalibrated.

There are many ways in which the operator can obtain inaccurate results. Switch positions or displays can be misread or the threshold plotted incorrectly on the audiogram. Correct placement of the earphones or the bone conductor is very important; if the earphone is not placed directly over the ear canal significant errors can arise.

In addition to disease many other factors can change hearing thresholds. Aspirin, some antibiotics, and menstruation are just three factors which, it has been claimed, can cause changes. The common cold can cause the Eustachian tubes to become partially blocked and this will change the threshold. An audiologist must be alert to these factors which might explain an abnormal hearing threshold.

Some explanation of how hearing defects can be diagnosed from the audiogram is given in section 15.4.

15.3.5. *Middle-ear impedance audiometry: tympanometry*

This is a technique for measuring the integrity of the conduction between the eardrum and the oval window to the inner ear by measuring the acoustic impedance of the eardrum (see figure 15.1). The primary function of the middle ear is that of an impedance matching system, designed to ensure that the energy of the sound wave is transmitted smoothly (with minimum reflection) from the air in the outer ear to the fluid in the inner ear. Middle-ear impedance audiometry (tympanometry) is a technique for measuring the integrity of this transmission system. If the middle ear is defective (whether due to a mechanical defect or physical inflammation) then the impedance matching might be lost and most of the energy of an applied sound will be absorbed or reflected. The acoustic impedance (see section 3.4.2) of the eardrum and middle ear is analogous to an electrical impedance. If the ear has a low impedance then an applied sound will be transmitted with very little absorption or reflection. If the middle ear is inflamed then the impedance may be high and most of an applied sound will be absorbed or reflected.

Electrical impedance is measured by applying a potential, V, across the impedance and recording the current, I, which flows. Then,

$$\text{impedance} = V/I.$$

The analogy in acoustics is that an alternating pressure is applied to the impedance and the resulting airflow is recorded.

$$\text{acoustic impedance} = \frac{\text{pressure}}{\text{flow}} = \frac{\text{pressure}}{\text{velocity} \times \text{area}} \left(\frac{\text{N m}^{-2}}{\text{m s}^{-1}\,\text{m}^2}\right).$$

The acoustic impedance is measured in acoustic ohms which have the units of N s m^{-5}. Figure 15.9 shows how sound can be applied as an alternating pressure to a volume whose acoustic impedance is to be measured for a given constant flow. The sound pressure at the entrance to the volume will be proportional to the acoustic impedance. If the acoustic impedance doubles, then the sound pressure level at the entrance to the volume will double. Now the volume flow provided by the loudspeaker can elicit two responses: the pressure within the ear canal might rise or the eardrum might move. In practice both of these responses will occur.

The wavelength of the sound used for the test is large compared with the length of the ear canal, so that there will be approximately uniform pressure in the ear canal. The frequency normally used is in the range 200–220 Hz, which corresponds to a wavelength in air of 1.7 m.

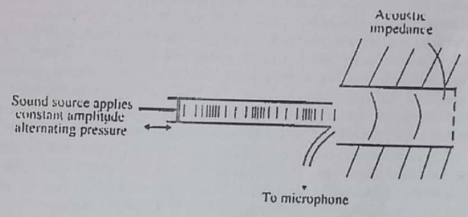


Figure 15.9. *Sound presented to a cavity which will have a certain acoustic impedance. The relative magnitude of the absorption and reflection of the sound determines the intensity of sound which is measured by the microphone.*

There are several designs of equipment which can be used to measure the acoustic impedance from the external auditory meatus. It is difficult to separate the impedance of the ear canal from that of the tympanic membrane and the middle ear. A complete analysis is outside the scope of this book. The measurement of acoustic impedance is widely used but, in most cases, only relative values of impedance are measured. The rest of this short section will be devoted to a qualitative description of the technique.

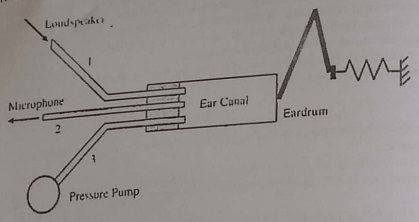


Figure 15.10. *A system for measuring the acoustic impedance of the eardrum and middle ear.*

A probe containing three tubes is introduced into the external ear canal; the tip of the probe is in the form of a plug which makes an airtight seal with the walls of the ear canal. The first of the tubes is connected to a sound source and the second to an air pump which enables the static pressure between the probe and the tympanic membrane to be controlled. The third tube is connected to a microphone which feeds an amplifier and recorder (figure 15.10). Under normal circumstances the pressure in the middle ear is equal to atmospheric pressure, the Eustachian tube having the function of equating middle-ear pressure to that close to the pharynx. If for any reason there is a pressure difference across the tympanic membrane then this stress will increase the stiffness of the membrane and hence its impedance. The system shown in figure 15.10 can be used to apply a positive pressure to the tympanic membrane and thus increase its impedance. The sound applied down the coupling tube will be reflected from the tympanic membrane back into the microphone tube. If the positive pressure is now reduced, then less sound will be reflected until a minimum impedance is reached when the pressure on both sides of the eardrum is the same. If the pressure is further reduced to a negative value then the impedance will rise again. In figure 15.11, the output from the impedance meter has been plotted as a graph of impedance versus the pressure applied to the eardrum. This is the result for a normal ear which has a well-defined minimum impedance when the applied pressure is zero.

Most impedance meters used clinically are calibrated by measuring the impedance of a known volume within the range 0.2–4.0 ml. The larger the volume, the smaller will be the impedance. However, if the reciprocal of impedance is used instead of impedance then there is a linear relationship with volume. The reciprocal of impedance is compliance, which is analogous to conductance in electrical terms. (Strictly, the reciprocal of impedance is admittance, which depends on the frequency of the applied pressure, but at the frequencies used in clinical impedance meters compliance and admittance are equal.) In the case of the ear a floppy eardrum will have a high compliance and a taut eardrum a low compliance. If compliance is used instead of impedance then the impedance curve of figure 15.11 can be replotted with the vertical axis calibrated as an equivalent volume (figure 15.11). This type of display is called a tympanogram and is widely used. Some examples of how otitis media (fluid in the middle ear) or a perforated eardrum affect the shape of the tympanogram curve are given in section 15.4.2.

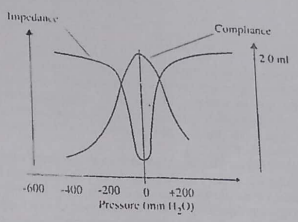


Figure 15.11. *Acoustic impedance plotted as a function of the pressure applied to an eardrum. The equivalent compliance is plotted as a tympanogram.*

Acoustic impedance measurements can be made at frequencies ranging from 100 Hz to several kilohertz and both the amplitude and phase of the impedance can be recorded. In the routine clinic a frequency of about 220 Hz is often used. The impedance is usually recorded as the applied steady pressure is changed from +200 mm of water pressure to −400 mm of water (+2 kPa to −6 kPa). In a normal ear, the minimum impedance is usually found between +100 and −100 mm of H_2O.

Stapedius reflex

There are two muscles within the middle ear: the tensor tympani and the stapedius. These muscles respond to acoustic stimulation. A loud sound introduced into one ear normally provokes bilateral contraction of the stapedius muscle. The muscle acts on the ossicles to stiffen the tympanic membrane. The intensity of sound normally required to cause this reflex is about 80 dB above threshold. The increase in stiffness of the eardrum changes the impedance of the ear. Observation of the impedance change resulting from the stapedius reflex contraction can be of some value in assessing hearing threshold.

15.3.6. Measurement of oto-acoustic emissions

The ear may be thought of as a transducer which converts sounds into a stream of nerve impulses. The way in which this is carried out is still not clear but it appears to be an active process. Part of the evidence for this is the observation that the ear actually produces sounds. Spontaneous emissions can be recorded in about 50% of normally hearing subjects although they are at very low levels and need very quiet conditions and very careful recording in order to be observed.

It was found by Kemp in 1978 that evoked oto-acoustic emissions could be recorded. Oto-acoustic emissions can be defined as *acoustic energy produced by the cochlea and recorded in the outer ear canal*. By applying a short acoustic stimulus to the ear a microphone placed within the ear canal can then be used to record an emitted sound. It can be shown that the emissions are not simply passive acoustic reflections of the sound; they occur over a period of about 20 ms following the stimulus, which is much too long a period for simple acoustic reflection; the response is related to the stimulus in a nonlinear manner, whereas a simple reflection would be linear, and the emissions cannot be recorded from ears with damaged outer hair cells.

There has been considerable experimental work carried out since 1978 to determine the origin of otoacoustic emissions and the conclusion is that the origin is related to the function of structures related to the outer hair cells within the cochlea. Because these emissions can be recorded relatively easily and they are sensitive to small amounts of hearing loss the technique has been developed as a means of testing the function of the cochlea objectively. Most studies show that oto-acoustic emissions cannot be recorded if the hearing loss is greater than 30 dB.

Figure 15.12 shows the acoustic emissions recorded from a normal subject following a stimulus with an intensity which was varied over the range 35–80 dB SPL. It can be seen that the response appears over a period of about 10 ms and has a maximum amplitude of about 500 μPa, i.e. 28 dB SPL. Now in a test subject the background noise level in the ear canal, even in a very quiet environment, may be about 30 dBA SPL because of the noise produced by blood flow, breathing, muscle contractions and joint movements in the head. In order to record oto-acoustic emission signals filtering and then averaging must be used. In a typical recording the stimulus will be repeated at 50 pps and 500 or more signals will be averaged.

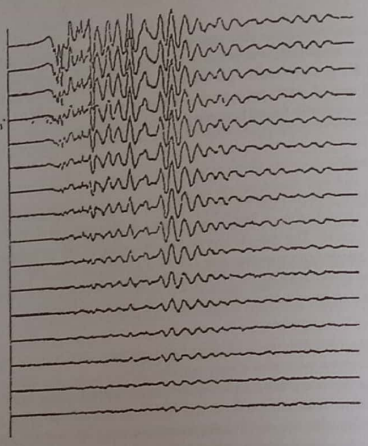


Figure 15.12. *Oto-acoustic emissions recorded in a normal adult for stimuli of intensity varied over the range 35–80 dB SPL (bottom to top of diagram). Time runs from left to right on a linear scale from 0 to 20 ms. The vertical scale is the same for all traces and the largest response is about 500 μPa (SPL 28 dB). It can be seen that the responses change with stimulus intensity but not in a linear fashion. In each case two average signals are presented so that the consistency of the response can be assessed. (From Grandori et al 1994 Advances in Otoacoustic Emissions, Commission of the European Communities, ed F Grandori.)*

Oto-acoustic emissions are usually recorded as the response to a transient stimulus. However, they can also be recorded either during the frequency sweep of a low-level tone stimulus or by recording the distortion products produced when two continuous sine waves of different frequencies are applied.

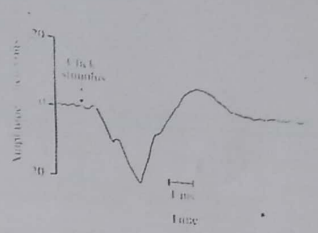


Figure 15.20. *Typical cochlear response obtained during an electrocochleography test. The stimulus was presented 2000 times at 80 dB above the threshold.*

15.6 HEARING AIDS

Hearing aids were first introduced in the 1930s, but were cumbersome devices and produced very poor sound quality. These aids used carbon granule microphones and amplifiers. Carbon granules change their electrical resistance when they are subject to pressure changes and so they can be used to modify an electrical current as sound changes the pressure applied to a diaphragm. Miniature valves superseded the carbon granule amplifiers and piezoelectric microphones replaced the carbon granule microphones. In the 1950s, transistors were introduced and currently most aids use integrated circuit amplifiers. Future aids are likely to be all digital devices. Piezoelectric microphones are still used, although ceramic materials are used as the piezoelectric element. A piezoelectric material has a crystal structure such that, when pressure is applied, shared electric charges are redistributed and so a potential is produced across the material. The diaphragm of a ceramic microphone is directly connected to the piezoelectric ceramic so that movement of the diaphragm gives a proportional potential difference across the material.

The need for a hearing aid is usually assessed on the basis of hearing tests, and table 15.2 gives some basic information on the classification of auditory handicap.

15.6.1 *Microphones and receivers*

Microphones and receivers are both transducers, the first converting from sound to electrical energy and the second vice versa. The 'receiver' is rather inappropriately named because it is the earpiece which actually produces the amplified sound, although it does allow the patient to receive the sound. In current hearing aids the microphone and the receiver are the largest components, with the exception of the battery. Ceramic microphones are the most commonly used but magnetic types are also in use; the magnetic type consists of a diaphragm connected to a ferromagnetic armature which is within the magnetic field produced by a coil. Movement of the diaphragm causes the armature to move and thus induces a potential in the coil. Most receivers are also magnetic types that use the current through the coil to move a metal cone attached to a diaphragm.

The coupling between the receiver and the ear canal is very important as it modifies the frequency response of the aid. Old aids of the body-worn type had a separate receiver placed directly over the ear canal but aids which are worn behind the ear contain both the microphone and receiver so that the sound has to be conducted to the ear canal through a short plastic tube. Many aids are now placed within the ear and the receiver makes acoustic connection very close to the eardrum.

Table 15.2. *Some of the problems which are associated with different levels of hearing loss.*

Average hearing loss	Speech understanding	Psychological implications	Need for a hearing aid
25 dB HL	Slight handicap: difficulty only with faint speech	Child may show a slight verbal deficit	Occasional use
35 dB HL	Mild handicap: frequent difficulty with normal speech	Child may be educationally retarded. Social problems begin in adults	Common need for hearing aid
50 dB HL	Marked handicap: difficulty even with loud speech	Emotional, social and educational problems more pronounced	The area of most satisfaction from a hearing aid
65 dB HL	Severe handicap: may understand shouted speech but other clues needed	Pronounced educational retardation in children. Considerable social problems	Hearing aids are of benefit—the extent depends upon many factors
85 dB	Extreme handicap: usually no understanding of speech	Pronounced educational retardation in children. Considerable social problems	Lip reading and voice quality may be helped by hearing aid

15.6.2. *Electronics and signal processing*

The three most important factors in the specification of the performance of a hearing aid are gain, frequency response and maximum output.

Gain can be varied by using the volume control and, in many aids, a range of 0–60 dB (1–1000) is provided. The maximum possible gain is usually limited by acoustic feedback from the receiver to the microphone which will cause a howl or oscillations if the gain is increased too far.

Frequency response should, ideally, cover the whole audio bandwidth but in practice, the performance of the receiver and the microphone limit the bandwidth. A typical frequency response for an aid is shown in figure 15.21.

Maximum output may be the most important part of the specification. A normal person might hear sounds with intensities ranging from 0 to about 90 dB. If these sounds are to be amplified by 60 dB then the range of intensities to be produced by the aid should be 60–150 dB. It is very difficult to produce sound at a level of 150 dB without distortion; many aids will only produce about 110 dB and the very best aids 140 dB. However, in most cases the maximum output required is much less than 150 dB.

The maximum output is limited both by the receiver and the electrical power available. The power supply for a hearing aid has to be provided from a battery. The current consumption in the quiescent state, i.e. no sound, may be about 100 μA and in a noisy environment 5 mA. If the battery is to last for more than a week of normal use then the battery capacity must be greater than about 60 mA h, e.g. for 100 h usage made up of 10 h at 5 mA and 90 h at 100 μA would require 59 mA h.

The trend in hearing aid design is to make them more versatile and adaptable to the hearing loss of a particular patient. Several computer programmes are available which will select a hearing aid appropriate to patient audiometric characteristics. The enormous recent advances in electronics have enabled many improvements to be made in hearing aid design. Not only can aids be made smaller but also sophisticated

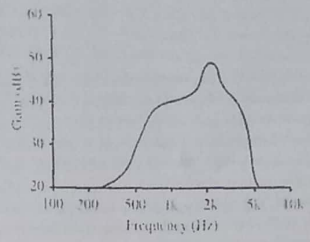


Figure 15.21. *Representative frequency response of a hearing aid.*

signal processing can be included. Digitized hearing aids have been developed to maximize high-frequency gain, filter out non-speech signals and exactly mirror a particular hearing loss. One way to enable a wide range of input sounds to be accommodated with a more limited output range is to use sound compression. Sound compression increases the range of sounds which can be amplified. Complex frequency responses can be provided and matched to the hearing loss of a particular patient. This field is changing so rapidly that the reader is advised to study manufacturers' literature as well as more detailed texts.

15.6.3. *Types of aids*

The range of hearing aid types is very wide but they can be classified according to where they are worn. Five major categories are:

1. Body worn.
2. Behind the ear (BTE).
3. In the ear (ITE).
4. In the canal (ITC).
5. Completely in the canal (CIC).

Body-worn aids can be relatively large which enables high-quality components and large batteries to be used. For these reasons the body-worn aid usually gives the largest 'maximum output' and the best sound quality. However, behind-the-ear and within-the-ear aids are usually more acceptable than the body-worn aids because they are more convenient and more socially acceptable. Body-worn aids are now rarely used, except where very high output is required.

The BTE aid hangs behind the ear with a tube connecting to the canal via an ear mould. The relatively large size allows complex circuitry to be included, but they are more visible than an ITE aid. The fact that the microphone is outside the pinna means that no natural ear resonance effects are obtained.

The ITE aids are popular and are often custom-made to fit the ear. Quite high gain is possible without feedback if the aid fits well into the ear. The aids are quite visible, however, and the microphone can be subject to wind noise. The maximum output is lower than that of a BTE aid.

The ITC aids sit in the concha portion of the ear and extend down into the ear canal. They have the advantage of being less visible than other aids, a reduced effect of wind noise and the ability to use natural resonances within the ear. Disadvantages include short battery life and an increased chance of feedback at higher gains.

CIC aids sit entirely within the canal and are almost completely invisible to casual view. However the constructional demands are high and performance may be limited. Battery life is short and control of the device difficult. Some ear canals are not sufficiently large to accommodate them.

The total performance of a hearing aid is determined by the microphone characteristics, amplifier characteristics, receiver/ear mould characteristics and the way in which these elements might interact. The ear mould is the plastic plug which is made to fit a particular ear and the sound-conducting tube which connects the aid to the mould. The acoustic properties of the plastic mould are relatively unimportant but it must make a tight seal to the walls of the ear canal. An analysis of the physics involved in the performance of the coupling is difficult and not appropriate to this introductory text.

15.6.1 *Cochlear implants*

Cochlear implants convert sound signals into electrical currents which are used to stimulate auditory nerve cells via electrodes placed within the cochlea. They are fundamentally different from normal acoustic hearing aids in that they replace the function of the hair cells within the inner ear. Implants are only used in patients who are profoundly deaf, whose hearing loss exceeds 105 dB. This is a small patient group representing less than 1% of the population.

These devices sample the sound and then use a processed version to electrically stimulate the auditory nerve within the cochlea. The microphone and signal processing are carried out within a body-worn unit. The output is then inductively coupled to the implant via a transmitter coil which is often held in place over the mastoid by magnets within both the transmitter coil and the implant. Implants allow for the perception of sound when the middle and inner ear are damaged but the auditory nerve is intact. Cochlea implantation is expensive and hence there is rigorous selection of patients in order to maximize the success of an implantation. The selection takes into account motivation, expectations, emotional state, age, intelligence and general fitness, as well as the expected benefit to hearing. Alessandro Volta in about 1800 was the first person to apply an electrical stimulus to the ear. Systematic research into electrical stimulation of the auditory nerve was carried out in the 1960s and 1970s, but self-contained implants were first developed in the 1980s both in Melbourne, Australia and in the USA.

Principles of operation

There are about 30 000 fibres in the auditory nerve and each of them is normally sensitive to a narrow range of frequencies over a dynamic range of intensity of about 30–40 dB. The intensity of sound is coded by the rate at which cells fire and the number of cells which are excited. Sounds of different frequencies are attenuated by different amounts as they are transmitted into the cochlea. It appears that the brain uses the spatial distribution of cells stimulated within the cochlea to determine the frequency of a sound and the temporal distribution of cell impulses to determine the intensity of a sound. This last sentence is oversimplified but it is the basis upon which cochlea implants have been designed.

In a cochlear implant either a single electrode or an array of electrodes are usually implanted, via a hole drilled close to the round window, inside the scala tympani, which places the electrodes close to the auditory nerve cells. Figure 15.22 shows an array of electrodes that can be placed within the cochlea. The electrodes are typically made from platinum–iridium and up to 22 electrodes are used. In some cases the electrodes are in pairs (bipolar) and an electrical stimulus is applied between the two, but in other cases the stimulus is applied between one electrode (the cathode) placed close to the nerve cells and an indifferent electrode placed further away. The principles of neural electrical stimulation are covered in sections 10.2 and 16.5. It is very important that tissue damage cannot occur as a result of the electrical stimuli applied via the electrodes. It is generally accepted that the charge density per pulse should be below 0.2 $\mu C\ mm^{-2}$ to minimize the risk of long-term damage.

If a single pulse has an amplitude of 0.5 mA and a duration of 100 μs then the charge per pulse is 0.05 μC. If the electrode is a square of sides 0.5 mm then the charge density per pulse will be 0.2 $\mu C\ mm^{-2}$.

In order to limit the current which can flow a constant current as distinct from a constant voltage stimulus is usually used. The constant current stimulus also has the advantage that the threshold for stimulation does not depend upon electrode impedance and so will not change with time as the electrode interface changes. In addition to limiting the charge which can be delivered it is important to avoid electrode polarization and the release of toxic products from electrochemical reactions by preventing any DC current flow. This is usually done by using a charge-balanced bipolar stimulus with both positive and negative phases.

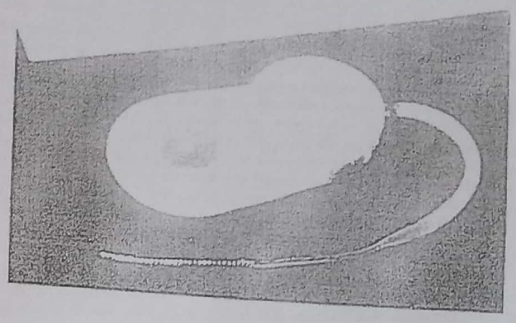


Figure 15.22. *A cochlear implant. The array of electrodes which are passed through the round window and placed within the scala tympani can be seen. The scale shown is in inches. (Courtesy of Cochlear Ltd.)*

Sound coding

Whilst most implants make some attempt to mimic normal cochlear function there is nonetheless a very wide range of strategies. Systems are generally multi-channel, and processing techniques range from mapping the frequency of the stimulus to electrode location in the cochlea to more complex coding strategies. These might code the stimuli as pulses, square waves or amplitude-modulated high-frequency carriers. All the implants seek to include the frequency information contained in human speech which is mainly between 300 and 3000 Hz. The square wave and pulse types are usually designed to track voice pitch or first formant. The carrier types amplitude modulate a carrier at about 15 kHz with a filtered audio signal.

Cochlear implants are still developing rapidly and a standard method of sound coding and impulse delivery has not yet emerged.

Performance and future

For adults, cochlear implants provide their users with significant benefits even though they do not replace the function of a normal ear. They help the wearer to recognize environmental sounds such as bells, calls and knocks at the door. They can also help improve lip-reading and the user's speech is likely to be more intelligible as they can monitor their own voice. For children, the pre- as well as postlingually deaf are now considered for implantation. Many children have learnt language through the use of an implant.

Patients are carefully selected and an implant programme includes a long period of training, learning and assessment after the implant procedure.

15.6.5. *Sensory substitution aids*

Cochlear implants do not enhance what remains of an impaired sense of hearing. They substitute one sensation for another in that sound is used to electrically stimulate the auditory nerve. Nonetheless, the sensation is perceived as coming from the ears. Many attempts have been made to use senses other than hearing to carry sound information. Research groups have tried converting a sound into a visual stimulus, an electrical stimulus or a vibratory stimulus. Some of these sensory substitution aids have reached the point of clinical application as part of the range of aids which might help the profoundly deaf.

The best developed of these sensory substitution aids is the vibrotactile aid. These devices convert sound into a modulated vibration which can then be used to vibrate a finger or elsewhere on the body. Such aids can certainly be used to give the wearer some awareness of environmental sounds. However, the information carrying capacity of such an aid is very limited and is almost certainly insufficient to convey normal speech. The range of frequencies which can be usefully used in a tactile aid is between 20 and 500 Hz as below 20 Hz individual cycles are perceived and above 500 Hz sensitivity falls rapidly. The percentage change in frequency which can be identified is about 25% which compares poorly with the ear where a change of about 0.3% can be heard.

The range of intensities which can be perceived using a vibrotactile stimulus is about 55 dB and the smallest detectable percentage change is about 2 dB. The 55 dB range is very much less than the range of the ear but the detectable change in intensity is similar to that of the ear. Temporal resolution is often measured as the ability to detect gaps or periods of silence and values of about 10 ms are quoted for vibrotactile input.

From this evidence it is clear that the tactile sense can be used for communication but it will have a much smaller information transfer capacity than the auditory system.

Information transfer rate can be defined as follows:

> Information transfer rate = the number of binary bits of information which can be communicated each second.

The information transfer rate required for speech transmission is usually considered to be about 20 bits s^{-1}. The number of bits which can be transmitted using a single-channel vibrotactile aid is at best about 4 bits s^{-1}. It is possible that by using several channels applied to the fingers for example that 20 bits s^{-1} might be communicated but this has not yet been proven.

Vibrotactile sensory substitution aids have been compared with cochlear implants (Stevens 1996) and it has been concluded that a tactile aid may offer advantages over a cochlear implant for certain categories of patient. In particular, those who are prelingually deaf may obtain limited benefit from a cochlear implant.

an important safeguard as the muscle will always relax before contracting again, thus ensuring that the heart will continue to act as an effective pump even if stimuli are arriving at many times the normal rate. The absolute refractory period is followed by a relative refractory period during repolarization, during which a larger than normal stimulus is needed to initiate depolarization. A premature beat during this period (or an external electrical stimulus) can cause ventricular fibrillation.

The electrocardiogram

The electrocardiogram recorded from the right arm and the left leg has a characteristic shape shown in figure 16.15. The start of the P wave is the beginning of depolarization at the SA node. The wave of depolarization takes about 30 ms to arrive at the AV node. There is now a delay in conduction of about 90 ms to allow the ventricles to fill. The repolarization of the atria, which causes them to relax, results in a signal of opposite sign to the P wave. This may be visible as a depression of the QRS complex or may be masked by the QRS complex. After the conduction delay at the AV node, the His–Purkinje cells are depolarized, giving rise to a small signal which is usually too small to be visible on the surface. The conduction through the His–Purkinje system takes about 40 ms, and the depolarization and contraction of the ventricles then begins, giving rise to the QRS complex. Finally, repolarization of the ventricles takes place. This is both slower than the depolarization and takes a different path, so that the resulting T wave is of lower amplitude and longer duration than the QRS wave, but has the same polarity.

The mechanical events in the heart give rise to characteristic heart sounds, which can be heard through a stethoscope or can be recorded using a microphone on the chest wall. The first sound is low pitched and is associated with the closure of the atrio-ventricular valves as the ventricles start to contract. The second, a high-pitched sound, is associated with the closure of the aortic and pulmonary valves as the ventricles relax. Other sounds are usually the result of heart disease.

16.2.2. *The electrocardiographic planes*

The heart can be thought of as a generator of electrical signals that is enclosed in a volume conductor—the body. Under normal circumstances we do not have access to the surface of the heart and must measure the electrical signals at the surface of the body. The body and the heart are three dimensional, and the electrical signals recorded from the skin will vary depending on the position of the electrodes. Diagnosis relies on comparing the ECG/EKG from different people, so some standardization of electrode position is needed. This is done by imagining three planes through the body (figure 16.16).

The electrodes are placed at standard positions on the planes. The frontal plane is vertical and runs from left to right. The sagittal plane is also vertical but is at right angles to the frontal plane, so it runs from front to back. The transverse plane is horizontal and at right angles both to the frontal and the sagittal plane.

ECG/EKG monitoring, which simply checks whether the heart is beating or not, uses electrodes placed in the frontal plane. To diagnose malfunctions of the heart, the ECG/EKG is recorded from both the frontal and the transverse plane. The sagittal plane is little used because it requires an electrode to be placed behind the heart—often in the oesophagus.

The frontal plane ECG/EKG: the classical limb leads

The ECG is described in terms of a vector, the cardiac vector. The electrical activity of the heart can be described by the movement of an electrical dipole which consists of a positive charge and a negative charge separated by a variable distance. The cardiac vector is the line joining the two charges. To fully describe the cardiac vector, its magnitude and direction must be known. The electrical activity of the heart does not consist of two moving charges, but the electric field which is the result of the depolarization and repolarization of the cardiac muscle can be represented by the simple model of a charged dipole.

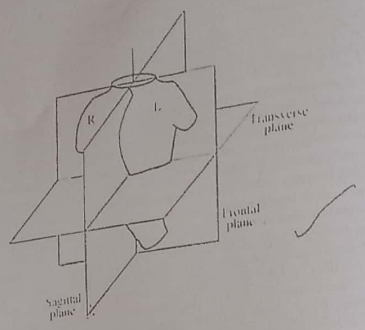


Figure 16.16. *The electrocardiographic planes*

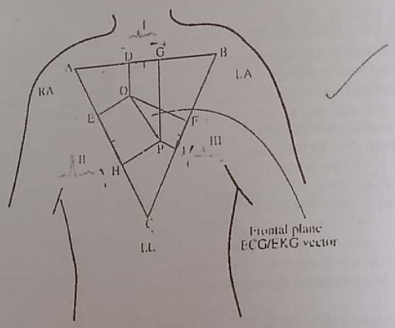


Figure 16.17. *Einthoven's triangle*

In the physical sciences, a vector is usually described by its length in two directions at right angles (e.g. the *x*- and *y*-axes on a graph). With the frontal plane ECG/EKG, it is usual to describe the cardiac vector by its length in three directions at 60° to each other. The resulting triangle (figure 16.17) is known as Einthoven's triangle, and the three points of the triangle represent the right arm (RA), the left arm (LA) and the left leg (LL). Because the body is an electrical volume conductor, any point on the arm, from the shoulder down to the fingers, is electrically equivalent, and recording from the left leg is electrically equivalent to recording from anywhere on the lower torso.

The three possible combinations of the three electrode sites are called leads I, II and III, and convention stipulates which is the positive electrode in each case.

Lead I	RA (−) to LA (+)
Lead II	RA (−) to LL (+)
Lead III	LA (−) to LL (+).

If the amplitude of the signals in the three leads is measured at any time during the cardiac cycle and plotted on the Einthoven triangle, the direction and amplitude of the cardiac vector can be found. In practice, 'cardiac vector' refers to the direction and amplitude of the cardiac vector at the peak of the R wave.

The use of Einthoven's triangle assumes that the human torso is homogeneous and triangular. This, of course, is not true but it is ignored in practice as the interpretation of the ECG/EKG is empirical and based on the correlations between the shape of the ECG/EKG and known disorders of the heart.

In order to work out the direction of the cardiac vector, recordings from leads I, II and III must be made. Draw an equilateral triangle ABC (figure 16.17) and mark the centre point of each side. Measure the height of the R wave on the same ECG/EKG complex for each of leads I, II and III. This is taken as the algebraic sum of the R and S waves, i.e. measure from the lowest point on the S wave to the highest point on the R wave. Note whether this is positive or negative. Using a suitable scale (e.g. 5 cm = 1 mV), draw each of the R wave amplitudes in the correct direction along the appropriate side of the triangle (DG, EH, FI). Place the centre of the R wave vector at the centre of the side of the triangle. Draw in the perpendiculars from each end of the vectors (DO, EO and FO; HP, GP and IP). The point of intersection, O, is the beginning of the cardiac vector, and the point of intersection, P, is the end. Draw in the cardiac vector OP. In practice, the measurements will not be perfect. The three lines will not meet at a point P, but will form a small triangle, within which is the end of the cardiac vector.

The normal cardiac vector direction depends on age and body build. The direction of lead I, from right to left, is taken as 0°. (Remember that we are looking from the front of the body, so that this runs from left to right on the diagram.) In young children, the axis is vertically downwards at +90°. During adolescence, the axis shifts to the left. A tall thin adult will have a relatively upright axis, whereas a short stocky adult might have an axis between 0° and −30°. An axis between −30° and −180° is referred to as left-axis deviation, and an axis between +90° and +180° is referred to as right-axis deviation.

The frontal plane ECG/EKG: augmented limb leads

Leads I, II and III are referred to as bipolar leads, because the measured signal is the difference in potential between two electrodes. Unipolar measurements are made by recording the potential at one electrode with respect to the average of the other two potentials. These are referred to as aVR, aVL and aVF (augmented vector right, left, and foot). The combinations are

aVR	(LA + LL)/2(−)	wrt RA(+)
aVL	(RA + LL)/2(−)	wrt LA(+)
aVF	(RA + LA)/2(−)	wrt LL(+)

The three unipolar leads have a direct vector relationship to the bipolar leads

$$aVR = -(I + II)/2$$
$$aVL = (I - III)/2$$
$$aVF = (II + III)/2$$

They are, in fact, the projection of the frontal plane cardiac vector onto three axes which are rotated 30° to the left from the Einthoven triangle. The direction and size of the cardiac vector can obviously be determined from the unipolar lead recordings in the same way as from the bipolar lead recordings.

The transverse plane ECG/EKG

The transverse plane ECG/EKG is recorded unipolarly with respect to an indifferent electrode formed by summing the signals from the left and right arms and the left leg (LA + RA + LL). Six electrodes are usually used, labelled V1 to V6. The electrodes are placed close to the heart and their position is more critical than the position of the frontal plane electrodes. They are placed on a line running round the chest from right of the midline to beneath the left axilla (figure 16.18). V1 and V2 are placed in the fourth intercostal space immediately to the right and left of the sternum. V3 is placed halfway between V2 and V4, and V4 is placed in the fifth intercostal space directly below the middle of the left clavicle. V4, V5 and V6 all lie on the same horizontal line, with V5 directly below the anterior axillary line (the front edge of the armpit), and V6 directly below the mid-axillary line (the mid-point of the armpit). The electrical signals recorded from the transverse plane electrodes are also shown in figure 16.18.

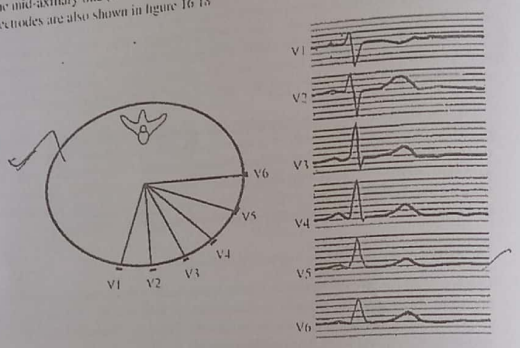


Figure 16.18. *The position of the chest electrodes for leads V1 to V6. Typical waveforms are also shown.*

The sagittal plane ECG/EKG

The sagittal plane ECG is rarely recorded. The indifferent electrode is again formed by the summation of the signals from the right and left arms and the left leg, and the active electrode is placed behind the heart. This is done using an oesophageal electrode, consisting of an electrode at the end of a catheter. The catheter is placed through the nose and down the oesophagus, until the electrode lies in the same horizontal plane as the heart.

For diagnostic purposes, the ECG/EKG is recorded on paper with either the six frontal plane leads only or the six frontal planes and the six transverse plane electrodes. For monitoring purposes, the ECG/EKG is displayed on a VDU screen, though provision may be made for recording abnormal stretches of the ECG either to paper or to memory. If a long-term record of the ECG/EKG of the patient's normal work is required, a magnetic recorder will be used. This is described in section 16.2.4.

It is commonplace nowadays for a patient to have an in-dwelling cardiac catheter, either for recording the ECG/EKG and pressure from within the heart, or for pacing using an external pacemaker. The presence of a direct electrical connection to the heart greatly increases the danger of small electrical leakage currents causing fibrillation. If the patient has an in-dwelling cardiac catheter, then ECG/EKG equipment which is isolated from earth must be used—this applies to equipment connected either to the cardiac catheter or to surface electrodes. All modern ECG equipment is isolated from earth.

Monitoring or recording equipment must be capable of measuring the ECG/EKG, which has an amplitude of about 1 mV, whilst rejecting the interfering common-mode signal due to the presence of the 50/60 Hz mains supply. The subject of amplifier design and interference rejection was dealt with in Chapters 9 (section 9.3.3) and 10 (section 10.3).

Electrodes and filters

Electrodes are dealt with in detail in section 9.2. Two types of electrode are commonly used for ECG recording. A six- or 12-lead ECG/EKG recording will only take a few minutes at the most, but may be done on a very large number of patients each day. Plate electrodes are used with either saline-soaked pads or gel pads which are held on to the arms and legs with rubber straps. Suction electrodes are used for the transverse plane electrodes. For long-term monitoring, where the ECG/EKG may be recorded continuously for several days, disposable silver–silver chloride electrodes are used. These have a flexible backing material and use an electrode jelly formulated to give a minimum of skin reaction. Plate electrodes should not be used for long-term recording, as any corrosion of the plate can give rise to unpleasant skin reactions. With good skin preparation, the impedance of the electrodes will be less than 10 kΩ, so that an amplifier input impedance of 1 MΩ is adequate. In practice, the electrodes will not have exactly the same impedance. The electrode impedance will act as a potential divider with the common-mode input impedance of the amplifier. To achieve 80 dB common-mode rejection with a 10 kΩ difference in impedance between the electrodes, a common-mode input impedance of 100 MΩ is required (see section 10.3.3).

Interference on the ECG/EKG is often caused by the electrical signals from any muscles which happen to lie between the electrodes. The majority of the EMG spectrum lies above the frequency range required for recording the ECG/EKG, so that most of the EMG interference can be removed by suitable filtering of the signal. For long-term monitoring, the electrodes are placed on the chest so that the signals from the arm and leg muscles are eliminated. The bandwidth needed for diagnosis (which requires accurate reproduction of the waveshape) is 100 Hz, whilst 40 Hz is adequate for monitoring. The lowest frequency of interest in the ECG/EKG is at the repetition rate, which is not normally lower than about 1 Hz (60 b.p.m.). However, because the waveshape is important, a high-pass filter at 1 Hz cannot be used, because the distortion due to the phase shift of the filter will be unacceptable. The usual solution is to reduce the centre frequency of the high-pass filter until there is no significant phase shift at the lowest frequency of interest. A low-frequency 3 dB point of 0.05 or 0.1 Hz is usually used. The introduction of digital filters is enabling the 0.1 Hz cut-off to be increased so that improved baseline stability is obtained without risk of waveform distortion.

The electrocardiograph

The electrocardiograph usually records the ECG/EKG on paper with 1 mm divisions in both directions and every fifth line emphasized. The standard paper speed is 25 mm s^{-1} (1000 ms cm^{-1}), with a sensitivity of 10 mm mV^{-1} (1 mV cm^{-1}). An historical oddity is that the amplitude of the various parts of the ECG/EKG is quoted in millimetres (assuming the standard calibration). As there is nothing fundamental about the calibration, it would be more logical to quote the amplitude in millivolts.

The standard ECG/EKG for inclusion in the patient's notes is recorded using a portable electrocardiograph, which may record one, or three lead positions simultaneously. The three-lead machines may switch between the leads automatically to give a fixed length of recording from each lead, ready for mounting in the notes. If the lead switching is to be done manually, the recording for each lead would be continued until 5 or 10 s of record has been recorded free from artefacts.

First of all, the patient should be encouraged to relax. Taking an ECG/EKG may be routine for the technician, but it is not routine for the patient, who may think that something in the test is going to hurt or they may be apprehensive about the results. The skin should be cleaned gently and the electrodes applied. For an automatic recorder, all 12 electrodes will have to be applied. For a one-channel recording, the three electrodes will be moved to the appropriate sites between each recording. Check that there is no mains interference—it may be necessary to earth yourself by touching the machine or the electrocardiograph may have to be moved to reduce the interference. If the patient is relaxed, there should be no EMG interference.

16.2.4 *Ambulatory ECG/EKG monitoring*

The traditional method of studying the ECG/EKG of a patient with suspected abnormalities that are not visible on a standard ECG/EKG recording is to confine the patient to bed for a few days with an ECG/EKG monitor connected, and tell a nurse to look for abnormalities on the ECG/EKG. Automatic arrhythmia detectors are also available which will do the nurse's job without fatigue, but this is still a very expensive method of diagnosis, and it may not be successful because many ECG/EKG abnormalities occur as a result of stress during the normal working day. Monitoring the ECG/EKG of patients during their normal working day is both cheaper and more effective. The monitoring is usually done using a small digital recorder.

The heart contracts about 100 000 times in 24 h. If a 24 h long recording were replayed onto an ECG/EKG recorder with a paper speed of 25 mm s^{-1}, the record would be 1.26 km long. Some form of automatic analysis is obviously needed, and this is usually performed by a special purpose computer. First of all, the R wave must be detected reliably. Most of the energy in the R wave lies between 10 and 30 Hz. The ECG is therefore passed through a bandpass filter and full wave rectified (because the R wave may have either polarity) to give a trigger signal. The R–R interval can be measured, and alarm limits set for low and fast heart rates (bradycardia and tachycardia). More sophisticated analyses can be performed. The results of the analysis can be made available as trend plots or as histograms, and the analyser will write abnormal sections of ECG on a chart recorder for visual analysis by a cardiologist. One person can analyse about 50 24 h recordings per week using an automatic analyser.

16.3 ELECTROENCEPHALOGRAPHIC (EEG) SIGNALS

The EEG technician's major role is to provide the medical specialist with a faithful recording of cerebral electrical activity, but in order to do this the technician must have an understanding of both the recording equipment and the characteristics of the EEG and its source. Electroencephalograph simply means a graph of the electrical changes from the *enkephalos* (Greek for brain).

The EEG arises from the neuronal potentials of the brain but, of course, the signals are reduced and diffused by the bone, muscle and skin which lie between the recording electrodes and the brain. There is a technique called electrocorticography (ECoG) where electrodes are placed directly on the cortex during surgery, but this is not a routine technique. The advantage of ECoG is that the electrodes only record from an area of the cortex of about 2 mm diameter, whereas scalp electrodes record from an area about 20 mm in diameter.

16.3.1 *Signal sizes and electrodes*

The EEG is one of the most difficult bioelectric signals to record because it is very small, this probably explains why the ECG/EKG was first recorded in about 1895, but the EEG was not recorded until 1929. There were simply no methods of recording signals as small as the EEG in the first decades of the 20th Century.

The normal EEG has an amplitude between 10 and 300 μV and a frequency content between 0.5 and 40 Hz. If electrodes are applied perfectly and the very best amplifier is used, there will still be a background noise of about 2 μV p–p, which is significant if the EEG is only 10 μV in size. Every care must be taken to reduce interference and to eliminate artefacts such as those which patient movement can produce, if a good EEG is to be recorded.

The best electrodes are Ag–AgCl discs which can be attached to the scalp with collodion. The scalp must be degreased with alcohol or ether and abraded before the electrode is held in place; collodion is run round the edge of the electrode and allowed to dry. Electrolyte jelly is then injected through a hole in the back of the disc electrode to form a stable scalp contact.

In a routine EEG clinic it is normally much too time consuming to apply many disc electrodes with collodion, which has to be removed with acetone after the test, and so electrode skullcaps are often used. The skullcap is an elastic frame which can be used to hold saline pad electrodes in place. The electrodes are a chlorided silver core with a saline loaded (10 g l^{-1}) cotton material around the core. These electrodes can be attached quickly and give good results.

Electrodes can be placed all over the scalp and different combinations used for recording. The most commonly used electrode placement system is the 10–20 system, so named because electrode spacing is based on intervals of 10% and 20% of the distance between specified points on the head. These points are the nasion and inion (the root of the nose and the external occipital protuberance at the back of the head), and the right and left pre-auricular points (the depressions felt in front of the upper part of the ear opening). The 10–20 system is shown in figure 16.19. In this diagram the letters correspond to anatomical areas of the brain as follows. O, occipital; P, parietal; C, central; F, frontal; FP, frontal pole; T, temporal; and A, auricular. Nineteen electrodes are used in the 10–20 system.

16.3.2 *Equipment and normal settings*

An EEG machine is basically a set of differential amplifiers and recorders. The distinguishing features are that there is usually a minimum of eight channels—and in many cases 16 or more channels—on the recorder, and there may be provision for 44 input electrode connections. The eight-channel machines are normally portable types. The amplifier outputs are usually digitized so that a computer can be used for further analysis and display.

Sixteen differential amplifiers will have a total of 32 input connections plus one earth connection. The input selector switches allow the correct combination of electrodes to be connected to the differential amplifiers; on some machines every electrode may be selected separately, but in others a complete combination (called a montage) can be selected by one switch. If the electrodes are selected individually then it must be remembered that each differential amplifier has both a non-inverting and an inverting input. These + and − inputs usually correspond to white and black wires, respectively, and must be connected correctly.

There are internationally agreed 'normal' or 'standard' settings for an EEG recording, these are listed below:

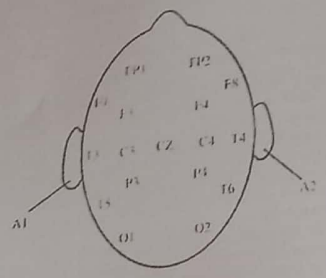


Figure 16.19. The 10–20 system of electrode placement

Chart speed. Speeds of 15, 30 and 60 mm s^{-1} are usually provided but 30 mm s^{-1} is the standard setting.

Gain setting. Switched settings are usually given but 100 μV cm^{-1} is the standard for routine recording.

Time constant. The low-frequency response of an EEG recorder is usually quoted as a time constant (TC) and not as a −3 dB point. 0.3 s is the standard time constant: it corresponds to a −3 dB point of 0.53 Hz.

Filters. The high-frequency response of an EEG recorder is quoted as a −3 dB point. 75 Hz is the standard setting but other values such as 15, 30 and 45 Hz are available to reduce interference which cannot be eliminated by other means.

A calibration facility is included so that the gain settings can be checked. The calibration allows a signal of say 100 μV to be introduced at the inputs of the differential amplifiers. This type of calibration does not check that the electrodes have been carefully applied and are performing correctly. Many machines include an electrode impedance test circuit which allows every electrode to be tested: an impedance below 10 kΩ is necessary for the best recording. Some machines also include a facility called a biological test whereby one electrode on the body is driven with a standard test signal; this test signal should appear equally on all channels if all the electrodes and amplifiers are functioning correctly.

16.4.7 *Normal EEG signals*

It is not possible in this short section to describe the 'normal EEG'. What we can do is to outline a normal recording procedure and to give one example of an EEG tracing.

A complete EEG test will take about 30 min and it be essential that the test is conducted in a quiet environment. The room must be both acoustically quiet and electrically quiet if interference is not to be troublesome. A location which is remote from sources of interference such as operating theatres and physiotherapy departments is best. Only one person should normally be in the room with the patient: a bell call system can always be used to bring rapid help if required. Resuscitation equipment, including oxygen, should always be on hand. In some cases, for example, young children, it may be necessary to have two people present in the room. Application and testing of electrodes may take about 10 min, after which the patient is asked to relax for the duration of the test. In order to record the EEG during a range of states the patient is first asked to relax with their eyes closed for 5–10 min; a further shorter recording is then made with the eyes open. Following this the patient is asked to hyperventilate (breathe as fast as possible) for about 3 min. Hyperventilation is a form of stimulation to the brain as oxygen levels are increased and carbon dioxide levels decreased; another form of stimulation which can make EEG abnormalities more obvious is a flashing light. Flashes at a repetition frequency of 15 s^{-1} are often used as this can precipitate abnormal rhythms in patients suffering from epilepsy.

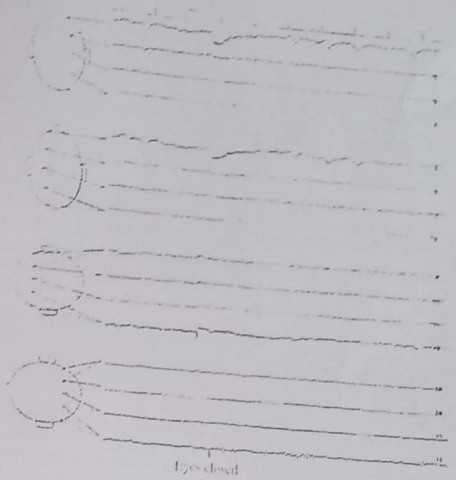


Figure 16.20. A 16 channel EEG recorded from a normal subject. Time runs from left to right and at the top, one second marker blips are shown, the amplitude of these blips corresponds to 75 μV. When the eyes are closed an artefact can be seen on channels 1 and 5 and a regular rhythm within the alpha range (8–13 Hz) appears on several channels. (Courtesy of Dr J A Jarratt, Department of Neurology, Royal Hallamshire Hospital, Sheffield.)

Figure 16.20 shows a normal 16-channel EEG recording. The waveform varies greatly with the location of the electrodes on the scalp. There are, however, certain characteristics which can be related to epilepsy, seizures and a number of other clinical conditions.

Epilepsy can give rise to large-amplitude spike and wave activity and localized brain lesions may give distinctive, large-amplitude, slow waves. An alert and wide-awake normal person usually displays an unsynchronized high-frequency EEG, whereas if the eyes are closed a large amount of rhythmic activity in the frequency range 8–13 Hz is produced. As the person begins to fall asleep, the amplitude and frequency of the waveforms decrease. Many more 'normal' patterns can be described.

Problems and artefacts

There are very many practical problems which arise when recording an EEG. Some of the problems are associated with the equipment, but other problems originate from the patient whose co-operation must be obtained if spurious signals from movement of the eyes or muscles are to be avoided. The list which follows includes some of the more common causes of artefacts on an EEG trace: electrode artefacts are usually the most troublesome.

Eye potentials. There is a potential of several millivolts between the back and front of the eyes. This potential gives rise to current flow through the tissues surrounding the eyes and this current will change as the eyes move. The effect can be seen as large deflections of the EEG trace when the eyes are moving.

ECG. The ECG is not usually a major problem in EEG recording, but if the recording electrodes are spaced a long way apart an ECG will be recorded and seen as sharp regular deflections on the recording. The artefact which results if the patient has an implanted cardiac pacemaker is very large and cannot be removed.

Electrode artefacts. If the patient moves or the wires leading to the electrodes are disturbed, then the electrochemical equilibrium underneath the electrodes will be changed and so potential changes will occur. Another effect can occur if the patient is perspiring as this will also disturb the electrochemical equilibrium under the electrodes and give rise to quite large potential changes. These changes are usually slow baseline changes on the EEG.

There are very many more sources of spurious signals on the EEG trace. Ways in which electrical interference can arise were described in Chapter 10 (section 10.4). It has even been suggested that problems have arisen from dental fillings, where an electrical discharge between different metallic fillings gave rise to artefacts in the EEG. The EEG technician must always be on guard for possible sources of interference.

Particular EEG patterns have been associated with many conditions such as cerebral tumours, epilepsy, haematomas, concussion and vascular lesions. However, no attempt will be made to describe these patterns here. Analysis of EEG signals is not easy because it is difficult to describe the signals. The ECG/EKG can be described in simple terms because there are only about five major components to the waveform, but the EEG is a much more complex signal. The various frequency ranges of the EEG have been arbitrarily assigned Greek letter designations to help describe waveforms. Electroencephalographers do not agree on the exact ranges, but most classify the frequency bands as follows: below 3 Hz, delta rhythm; from 3–7 Hz, theta rhythm; from 8–13 Hz, alpha rhythm; and from 14 Hz upwards, beta rhythm.

Most humans develop EEG patterns in the alpha range when they are relaxed with their eyes closed. The alpha rhythm seems to be the idling frequency of the brain and as soon as the person becomes alert or starts thinking the alpha rhythm disappears. This is the rhythm which is used in biofeedback systems where the subject learns to relax by controlling their own alpha rhythm.

Very many attempts have been made to analyse the EEG using computers, in order to help clinical interpretation. Many EEG machines include a frequency analyser which presents the frequency components of the EEG on the same chart as the EEG. Currently none of the methods of analysis have been found useful routinely and so they will not be described here.

Many EEG departments make EEG evoked response measurements in addition to the background EEG.

16.4 ELECTROMYOGRAPHIC (EMG) SIGNALS

An electromyograph is an instrument for recording the electrical activity of nerves and muscles. *Electro* refers to the electricity, *myo* means muscle and the graph means that the signal is written down. The electrical signals can be taken from the body either by placing needle electrodes in the muscle or by attaching surface electrodes over the muscle. Needle electrodes are used where the clinician wants to investigate neuromuscular disease by looking at the shape of the electromyogram. He may also listen to the signals by playing them through a loudspeaker, as the ear can detect subtle differences between normal and abnormal EMG signals. Surface electrodes are only used where the overall activity of a muscle is to be recorded; they may be used for clinical or physiological research but are not used for diagnosing muscle disease. Both surface electrodes and needle electrodes only detect the potentials which arise from the circulating currents surrounding an active muscle fibre, and do not enable transmembrane potentials to be recorded. Nerves and muscles produce electrical activity when they are working voluntarily, but it is also possible to use an electrical stimulator to cause a muscle to contract and the electrical signal then produced is called an evoked potential. This is the basis for nerve conduction measurements, which allow the speed at which nerves conduct electrical impulses to be measured. This technique can be used to diagnose some neurological diseases and the principles of the method are explained in section 16.5.1.

Needle electrode measurements are almost always performed and interpreted by clinical neurologists although both technical and scientific assistance may be required for the more sophisticated procedures. Nerve conduction measurements can be made as an unambiguous physiological measurement which is then interpreted either by medically or technically qualified staff.

16.4.1 *Signal sizes and electrodes*

The functional unit of a muscle is one motor unit but, as the muscle fibres which make up the unit may be spread through much of the cross-section of the muscle, it is impossible to record an EMG from just one unit. If a concentric needle electrode, of the type shown in figure 9.3, is placed in a weakly contracting muscle, then the EMG obtained will appear as in figure 16.21. Each of the large spike deflections is the summation of the muscle action potentials from the fibres of the motor unit which are closest to the tip of the needle electrode.

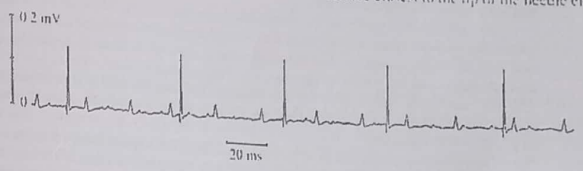


Figure 16.21. *An EMG recorded with a concentric needle electrode in a weakly contracting striated muscle.*

Remember that an upwards deflection represents a negative potential. The largest spikes all come from the same motor unit which is firing repetitively every 50 ms, but many other smaller spikes can be seen from fibres which are further away from the needle. The signal shown in figure 16.21 is a normal needle electrode EMG. The signal has a maximum amplitude of about 500 μV and the frequency content extends from about 10 Hz to 5 kHz. If the strength of contraction of the muscle is increased, then more motor units fire and the spike repetition frequency is increased, but the frequency content of the signal will not change significantly.

If a more localized recording is required then a bipolar concentric needle electrode can be used (figure 16.22). With a monopolar concentric needle electrode the signal is recorded from between the tip of the needle and the shaft, with a bipolar needle, the signal is that which appears between the two exposed faces of platinum at the needle tip, and the shaft is used as the earth or reference electrode. A bipolar needle only records from the tissue within about 1 mm of the tip and the signals obtained are smaller and also have a higher-frequency content than concentric needle recordings.

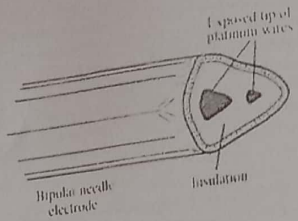


Figure 16.22. *The tip of a bipolar needle electrode.*

Any surface electrode placed over an active muscle can be used to record an EMG. If one electrode is placed over a muscle and the other electrode is placed several centimetres away, then EMG signals will be obtained from all the muscles lying between the electrodes. A more specific recording can be made if smaller electrodes are used and the separation is reduced to a few millimetres, but if the separation is reduced below about 1 mm then the amplitude of the signal falls rapidly. A very convenient and cheap electrode can be made from metal foil which can be cut to the desired size, will conform to the contours of the body and can be attached with adhesive tape. The skin must of course be cleaned and abraded before electrolyte jelly is applied and the electrode attached.

It is not possible, using surface electrodes, to record an EMG from just one muscle without interference from other muscles lying nearby. Even if two small electrodes are placed on the forearm, the EMG obtained will arise from many muscles. Localized recordings can only be made from needle electrodes, but these are uncomfortable and cannot be left in place for long periods of time. There is a type of electrode called a 'fine wire electrode' which can be left in a muscle for long periods: a wire is passed down the centre of a hypodermic needle which is then withdrawn, leaving the wire within the muscle. This can give an excellent long-term EMG recording.

The high-frequency content of surface electrode EMG signals is less than that from needle electrodes because of the volume conductor effects which were described in section 16.1.6. The recording amplifier should have a bandwidth from 10 to 1000 Hz. The amplitude of the signals depends upon the relative position of the electrodes and the muscle, but signals up to about 2 mV are typical.

16.4.2. *EMG equipment*

An EMG machine can be used to record both voluntary signals and evoked potentials. The amplitude of the signals will range from less than 1 μV up to 10 mV, the smallest signals are those produced by nerves and recorded from surface electrodes, whereas the largest are those evoked potentials from large muscles. Figure 16.23 gives a block diagram of an EMG machine.

The pre-amplifier will be a differential amplifier with the following typical specification:

Amplification	100
Input impedance	10 MΩ
Noise with input shorted	2 μV p–p
Common-mode rejection ratio	80 dB
Bandwidth (−3 dB points)	10 Hz–10 kHz

The output from the pre-amplifier is taken to the main amplifier and then to the A–D converter and host computer. The signal is also usually taken to a loudspeaker as EMG signals fall within the audio band and the

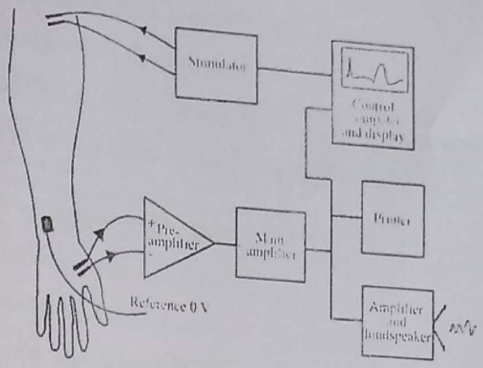


Figure 16.23. *A block diagram of an electromyograph.*

ear is very sensitive to subtle distinctions between signals. The remaining component of the EMG machine is the stimulator which is used for nerve conduction measurements. This will be considered in some more detail in section 16.5.

Equipment testing

The most common fault in all electrophysiological equipment is broken leads. Plugs and leads must be inspected regularly. Surface electrodes will have three connections to the differential amplifier: two inputs and an earth connection. A check on the operation of an EMG amplifier can be made by shorting together the three input connections and setting the amplifier gain to maximum. This may give a display of 10 μV per division on the screen, and if the amplifier is operating correctly a band of noise will be seen on the trace. By increasing the volume control on the loudspeaker amplifier the noise should be heard as a random broad-band signal. The simplest way to check the stimulator is to hold both output connections in *one* hand and to increase the output control slowly. A shock should be felt at an output of about 60 V.

16.4.3. *Normal and abnormal signals*

Clinical electromyography using needle electrodes consists of inserting the sterilized needle into a muscle and then recording a voluntary EMG pattern from several points within the muscle. Samples are taken at several points because a diseased muscle may contain both normal and abnormal fibres. The neurologist will usually listen to the EMG signal, which sounds like intermittent gunfire. The patient will normally be asked to make only a mild contraction of the muscle so that individual spikes can be identified. When a strong contraction is made, a complete interference pattern is obtained sounding rather like an audience clapping. Not all muscles give the same sound although the difference between muscles of the same size is not great. An extreme case is the signals which can be recorded from a fine needle placed in the small muscles which move the eyes. these muscles are very small and the spikes obtained are of very short duration.

An individual action potential or spike when viewed on a screen has a total duration of a few milliseconds and usually contains only two or three deflections. In a myopathic muscle the action potentials are often smaller, may have more than three phases, and are of shorter duration than normal signals.

It is very difficult to distinguish individual spike potentials from a surface electrode recording. The amplitude of the EMG waveform is the instantaneous sum of all the action potentials generated at any given time. Because these action potentials occur in both positive and negative directions at a given pair of electrodes, they sometimes add and sometimes cancel. Thus the EMG pattern appears very much like a random noise waveform with the energy of the signal a function of the amount of muscle activity (figure 16.24).

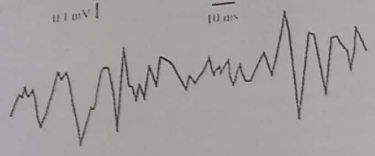


Figure 16.24. *An EMG recorded from surface electrodes*

Signal analysis and clinical uses

Electromyography is used

- in the diagnosis of neuromuscular disorders;
- as a measure of relaxation in the application of biofeedback techniques;
- as an index of muscle activity in physiological studies such as gait analysis.

There are very many clinical uses of electromyography, but it must be said that electromyography is really an extension of the classical methods of clinical examination, and each patient must be studied as an independent exercise in neurology. The skill of the electromyographer is as much in the planning of the examination as in its performance and interpretation. It seems improbable that electromyography will ever become a routine test performed by a technician under remote supervision.

Having made the point of the last paragraph it is of interest to outline very briefly the areas where clinical electromyography is useful. Following damage to a nerve, EMG signals give characteristic patterns of denervation which allow a prediction to be made about recovery. Damaged nerves may recover over periods as long as several years. EMG patterns characteristic of denervation include spontaneous activity such as small fibrillation potentials of short duration, instead of normal voluntarily produced spike potentials. Central neurogenic lesions such as motor neurone disease, poliomyelitis, and also spinal cord compression, cause characteristic EMG patterns which include large spike potentials with many deflections, synchronized motor unit activity, and some spontaneous electrical activity. Various inherited myopathies such as muscular dystrophy also give characteristic EMG patterns where the spike potentials are small, look ragged and contain more high-frequency components than the normal EMG.

Many methods of signal analysis have been tried to quantify EMG patterns, some depend upon measuring the frequency content of the signals. These can be of some use in quantifying the EMG and they have been shown to be helpful in identifying the carriers of muscular dystrophy, but they are not yet applied routinely

3.4.5 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 oder 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

(c) Now suppose a 100 mV signal is applied to the antenna. Calculate the output power. Is this a reasonable answer? Explain. What would actually happen in a real receiver?

19. (a) Draw the block diagram for an FM broadcast receiver. It is to have one RF stage and an IF frequency of 10.7 MHz. The local oscillator will operate above the signal frequency. Indicate on the diagram the frequency or frequencies at which each stage operates when the receiver is receiving a station at 94.5 MHz.

(b) What is the image frequency of the receiver described above?

(c) Give two ways in which the image rejection of the receiver could be improved.

20. An FM receiver is double conversion with a first IF of 10.7 MHz and a second IF of 455 kHz. High-side injection is used in both mixers. The first local oscillator is a VFO, while the second is crystal-controlled. There is one RF stage, one stage of IF amplification at the first IF, and three stages of combined IF amplification and limiting at the second IF. A quadrature detector is used as the demodulator. The receiver is tuned to a signal with a carrier frequency of 160 MHz. Draw a block diagram for this receiver.

21. Consider the transceiver whose block diagram is in Figure 9.20. Let the first I.F. be 16.9 MHz, with low-side injection of the first local-oscillator signal.

(a) What frequencies would the synthesizer produce while transmitting and receiving, respectively, on marine channel 14 (156.7 MHz)?

(b) Describe and explain the technique used for modulation in the transmitter section.

(c) The transmitter section uses a maximum deviation of 5 kHz and operates with modulating frequencies in the range of approximately 300 Hz to 3 kHz. Suggest an appropriate receiver bandwidth.

Cellular Radio

10

Objectives

After studying this chapter, you should be able to:

- Describe the history of personal communication up to the beginning of digital cellular radio.
- Explain the operation and limitations of CB radio and cordless telephones.
- Explain the operation of and perform relevant calculations for North American analog cellular telephone systems.
- Explain the operation of and perform relevant calculations for North American digital cellular telephone systems.

AR

Use of **half-duplex** (push-to-talk) operation for CB radio means only one channel is needed per conversation, and using AM (including its narrower bandwidth variant, SSB) keeps the required bandwidth less than would be needed for FM. The channels are spaced at 10-kHz intervals.

Selection among the 40 available channels is done by the operator, who is supposed to listen and manually switch to a clear channel before transmitting. Since anyone can listen in on any channel, there is no privacy.

The main disadvantages of CB radio are directly related to its simplicity and informality. Lack of privacy and co-channel interference are major problems. So is the lack of any connection to the wireline phone system or any access to repeaters for reliable, long-distance communication. The relatively high power level needed for communication without repeaters causes portable transceivers to be large and heavy. See Figure 10.1 for a typical example. The low frequency, which requires antennas to be large if they are to be efficient, is also a problem for portable transceivers.

FIGURE 10.1
Handheld CB transceiver
(Courtesy of Tandy Corporation)

Some of these problems have been addressed with unlicensed FM transceivers currently being sold for the 46/49-MHz and 460-MHz bands. The latter frequency range is called the *Family Radio Service (FRS)*. These transceivers are more compact but they still suffer from the other limitations mentioned earlier.

$$P = \frac{V^2}{R}$$

$$= \frac{(0.5 \times 10^{-6})^2}{50}$$

$$= 5 \text{ fW}$$

$$= -113 \text{ dBm}$$

We also need to express the 4 W transmitter power in dBm:

$$P_t = 4 \text{ W}$$
$$= 36 \text{ dBm}$$

The antenna gains increase the effective power by 2 dB. Our allowable path loss is then

$$L_p = 36 \text{ dBm} - (-113 \text{ dBm}) + 2 \text{ dB} = 151 \text{ dB}$$

The path loss is given by

$$L_p = 32.44 + 20 \log d + 20 \log f \qquad (10.2)$$

where

L_p = free-space loss in decibels
d = path length in km
f = frequency in MHz

Here we know L_p and f and we need to calculate d. Rearrange Equation (10.2):

$$20 \log d = L_p - 32.44 - 20 \log f$$
$$= 151 - 32.44 - 20 \log 27$$
$$= 89.9$$
$$d = 31378 \text{ km}$$

Therefore, this system is quite obviously limited by the distance to the radio horizon (and possibly by interference from other transmitters nearer to the receiver) and not by transmitter power. In this situation the power level could easily be reduced considerably with no effect on communication quality.

In a practical mobile situation the attenuation might actually be proportional to the fourth power of distance and would vary greatly depending on reflections. Still, range tends to be limited by the horizon, reflections and shadows, and interference, rather than power level.

We can estimate the range of CB communication by making some assumptions based on typical equipment and propagation paths, using the techniques introduced in Chapter 7.

EXAMPLE 10.1

Two handheld CB transceivers a,e held 1 m above flat, level terrain. The transmitter power output is 4 W and the receiver sensitivity is 0.5 μV into 50 Ω. The transmitting and receiving antennas are both loaded vertical monopoles with a gain of 1 dBi. Determine whether the maximum communication range is limited by power or distance. Assume there is no interference and that free-space attenuation applies.

SOLUTION

First we should recognize that our answer is likely to be optimistic. Usually the terrain is not flat and there are reflections from buildings, vehicles, and so forth. Our results might be fairly accurate for transmission over water.

First let us find the possible line-of-sight range, which is limited by the distance to the radio horizon. In Chapter 7 we noted that

$$d = \sqrt{17h_t} + \sqrt{17h_r} \qquad (10.1)$$

where

d = maximum distance in km
h_t = transmitting antenna height in m
h_r = receiving antenna height in m

Applying this equation to our situation gives

$$d = \sqrt{17h_t} + \sqrt{17h_r}$$
$$= \sqrt{17 \times 1} + \sqrt{17 \times 1}$$
$$= 8.2 \text{ km}$$

This is the maximum distance regardless of power level. Now let us see what free-space distance would be possible with the given power level, antenna gains, and receiver sensitivity, ignoring the horizon. We'll ignore transmission line losses, which are probably negligible anyway since the antenna is mounted directly on the transceiver.

First we need to convert the required voltage at the receiver to a power level in dBm.

10.1 Introduction

With this chapter we begin our discussion of several systems that are sometimes grouped together as personal communication systems (PCS). Like many technical terms, this one has several meanings. Specifically, it is used for a particular variant of cellular radio which will be described in the next chapter. But more generally, it can be applied to any form of radio communication between individuals.

In this chapter, after some historical introduction, we look at the common North American cellular telephone system, known as the Advanced Mobile Phone Service (AMPS). The network, the cell sites, and the portable and mobile telephones are described.

AMPS is still the most common cellular radio technology in North America, but a digital variation has become increasingly popular. We look at this system in this chapter as well. *Newer Personal Communication Systems (PCS)*, which use cellular techniques but at a different frequency range and with an assortment of digital modulation schemes, are the subject of the next chapter.

10.2 Historical Overview

The cellular radiotelephone system has its origins in much earlier systems. There has long been a need for portable and mobile communication. Three early concepts, two of which are still in wide use today, show aspects of what is needed. A brief look at each will show why cellular radio was created.

5' Band Radio

This is probably the earliest true personal communication system. Introduced in the United States in the 1960s, *citizens' band (CB) radio* enjoyed great popularity in the 1970s, followed by an almost equally steep decline as CB radio was intended to do some of the same things that are envisaged by more recent personal communication systems. The relatively low frequency of 27 MHz made transceivers affordable when CB radio was introduced, and the absence of any test for a license made it easy for anyone to get involved. The transmitter power limit of four watts for full-carrier AM, or twelve watts peak envelope power for SSB, is designed to reduce interference by restricting the communication range to a few kilometers. The fact that most CB operation is between mobile units with low antenna heights and no repeaters also limits the effective range. The restricted range is necessary to limit interference since there are only 40 channels. This should not be a problem since CB radio is intended for local communication.

Cordless Telephones

Most cordless phones are intended as simple wireless extensions to ordinary telephone service. For best results, a telephone, cordless or otherwise, should operate in full-duplex mode; that is, it should be capable of transmitting and receiving at the same time. Thus a cordless phone needs two radio channels, separated widely enough in frequency to avoid interference between them. Early designs had only a single channel for each direction, so the transmitter power levels and range had to be kept very small to minimize interference. Nonetheless, consumers found that a telephone that could be carried freely throughout a house and its grounds was very useful and cordless phones have been extremely popular. A typical modern example is shown in Figure 10.2.

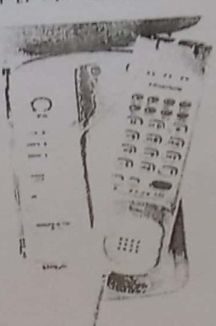


FIGURE 10.2
Cordless phone
(Courtesy of Tandy Corporation)

Most current designs use analog FM in either the 43–49-MHz or 900-MHz bands. Older phones use AM at about 1.7 MHz for the handset, with FM at 49 MHz for the base unit. Some 900-MHz models use analog FM while others use digital spread-spectrum techniques, and there is at least one model (the Panasonic Gigarange™) that uses a 2.4-GHz spread-spectrum signal from base to handset, and 900 MHz from handset to base. Most cordless

phones can automatically scan 10 or 25 channels, choosing a clear channel (assuming one exists) without any user involvement.

Table 10.1 shows the frequencies currently in use for 43–49-MHz cordless phones. Note that the base and portable frequencies must be widely separated, since both are in use simultaneously for full-duplex communication. Both the phone and base units require duplexers to separate the transmit and receive frequencies. Different manufacturers of 900-MHz phones use different channel frequencies and spacings, but the base and handset

TABLE 10.1 Cordless Telephone Frequencies: 43–49 MHz band

Channel	Base	Handset
1	43.720	48.760
2	43.740	48.840
3	43.820	48.860
4	43.840	48.920
5	43.920	49.000
6	43.960	49.080
7	44.120	49.100
8	44.160	49.160
9	44.180	49.200
10	44.200	49.240
11	44.320	49.280
12	44.360	49.360
13	44.400	49.400
14	44.460	49.480
15	44.480	49.500
16	46.610	49.670
17 (B)	46.630	49.845
18 (C)	46.670	49.860
19	46.710	49.770
20 (D)	46.730	49.875
21 (A)	46.770	49.830
22 (F)	46.830	49.890
23	46.870	49.930
24	46.930	49.990
25	46.970	49.970

The letters A through F denote channels that are also used for baby monitors.

frequencies are typically separated by about 20 MHz. The power level is deliberately set very low so that range is power-limited rather than extending to the radio horizon. It is surprising how low the power level for a cordless phone can be and still give reasonable results.

MPLE 10.2

A cordless phone operating at 49 MHz is to have a range of 50 m. Assuming 0 dBi gain for the antennas and the same receiver sensitivity as in Example 10.1, what transmitter power is required?

SOLUTION

Obviously the distance to the horizon is not the limiting factor here. We can use Equation (10.2) to calculate the loss for a path length of 50 m:

$$\begin{aligned} L_{fs} &= 32.44 + 20 \log d + 20 \log f \\ &= 32.44 + 20 \log 0.05 + 20 \log 49 \\ &= 40.2 \text{ dB} \end{aligned}$$

If the required signal strength at the receiver is −113 dBm as before, then the transmitter power must be at least

$$P_t = -113 \text{ dBm} + 40.2 \text{ dB} = -72.8 \text{ dBm} = 52.7 \text{ pW}$$

In practice, the power levels are much higher to cope with fading due to reflections and absorption. Cordless phones in the 46/49-MHz band are restricted to an EIRP of about 30 μW, while 900 MHz digital phones can use about 16 mW EIRP.

Cordless telephones share much of the simplicity of CB radio. There are no license requirements, and there is no official coordination of frequencies. Users, or in most cases the phones themselves, simply try to choose a channel that is not in use. The newer cordless phones use digital access codes to prevent unauthorized persons from dialing the phone and possibly making unauthorized toll calls, but it is still not possible to use two nearby phones on the same channel at the same time. The use of FM does provide some protection from interference: due to the capture effect, the desired signal has only to be a few decibels stronger than the interfering signal in order to reduce interference to a reasonable level.

Privacy is not quite as nonexistent as with CB radio, since the newer phones automatically avoid occupied channels. This reduces accidental privacy violations, but anyone who wants to eavesdrop on an analog phone

can do so by using a scanner, for instance. The digital phones offer much better privacy.

At this time, listening to others' cordless or cellular phone calls is illegal in the U.S.A., though not in Canada. (In Canada, listening to these conversations is legal, but divulging what you heard to another person is not.)

Because of the limited number of channels, cordless phones rely on extremely low transmitter power (microwatts to a few milliwatts depending on the band and the phone) to limit interference. Of course, this also limits their range. These phones certainly provide access to the wireline phone network, but in general, only from the customer's own premises or very nearby. The newer spread-spectrum phones do have more range—up to a kilometer or so under ideal conditions.

Despite interference problems and severely limited range, cordless phones have been and remain very popular with consumers. Various attempts have been made, particularly in Europe, to devise systems called telepoints that would enable users to take their cordless handsets to public places like malls and office buildings and use them there. However, recent developments in cellular and PCS systems have caused these ideas to lose favor. The cordless phone seems likely to remain popular in its current niche as a low-cost, wireless extension phone. Its low power allows it to have long battery life (weeks of standby, hours of talk time), and especially at 900 MHz, antennas are reasonably small and unobtrusive. Comparing the cordless phone shown in Figure 10.2 with the portable CB transceiver displayed earlier shows that cordless phones are a step in the right direction in terms of convenience.

Improved Mobile Telephone Service (IMTS)

The familiar cellular radiotelephone system has its origins in much earlier systems that used a few widely spaced repeaters. Wide coverage was obtained by using powerful base-station transmitters with antennas mounted as high as possible. The mobile transceivers likewise used relatively high power, on the order of 30 watts. Very similar systems are still widely used in dispatching systems, such as those for taxicabs and ambulances, for example.

The most common type of mobile telephone, from its introduction in the mid-1960s until the coming of cellular radio in the early 1980s (the first commercial cellular system became operational in Chicago in 1983), was known as the Improved Mobile Telephone Service (IMTS). IMTS is a *trunked* system; that is, radio channels are assigned by the system to mobile users as needed, rather than having one channel, or pair of channels, permanently associated with each user. Narrowband FM technology is used. Two frequency ranges, at about 150 and 450 MHz were used for IMTS, with an earlier system called MTS operating at around 40 MHz. The three systems

given region. One of these would be the local wireline telephone company (*telco*). The other would be an independent company called a radio common carrier (RCC). Each carrier was assigned half the channels in each area in an attempt to encourage competition. (It would actually be more efficient in terms of spectrum usage to have only one provider; with the current system it is possible for one provider's channels to be fully loaded while the other still has available frequencies.)

Cellular radio goes a long way toward relieving the congestion described above by essentially reversing the conventional wisdom about radio systems using repeaters. Instead of trying to achieve long range by using high power, cellular repeaters are deliberately restricted in range by using low power, as discussed in Chapter 7, the high path loss associated with mobile propagation actually makes it easier to reduce interference in a cellular system. A reasonable elevation for the base-station antennas is still required to minimize radio shadows (behind buildings, for instance). Similarly, the mobile radios use low power (no more than 4 W ERP for mobiles and 600 mW or less for portable phones), and in fact, the mobile transmitter power is automatically limited by the system to the minimum required for reliable communication.

Instead of one repeater, there are many, located in a grid pattern like that shown in Figure 10.3. Each repeater is responsible for coverage in a small cell. As shown, the cells are hexagons, but of course in a real situation the antenna patterns will not achieve this precision—the cells are more likely to be approximately circular, with some overlap. All the cell sites in a region are connected by copper cable, fiber optics, or microwave link to a central office called a mobile switching center (MSC) or mobile telephone switching office (MTSO), and the MSCs are themselves interconnected so that the system can keep track of its mobile phones. The cellular system is connected at a point of presence to the wireline network, so that cellular customers can speak to wireline customers.

Note that there is no provision for direct mobile-to-mobile radio communication. Even if two cell phones are in the same room, a call from one to the other must go through a cell site and an MSC. Provided both portable phones are connected to the same network (A or B), there would be no need to go through the PSTN.

FIGURE 10.3
Cell boundaries (seven-cell repeating pattern)

Since each transmitter operates at low power, it is possible to reuse frequencies over a relatively short distance. As we saw in Chapter 7, typical mobile propagation conditions allow for a repeating pattern of either seven or twelve cells; the available bandwidth is divided among these cells. The frequencies can then be reused in the next twelve or seven cells, with the lower number possible when directional antennas are used with three sectors using different frequencies per cell.

Carriers and Frequencies

In the current North American system, there are 395 duplex voice channels, each consisting of one channel in each direction for each of the two carriers. There are also 21 control channels for each carrier used to set up calls and administer the system. AMPS uses narrowband analog FM, with a maximum frequency deviation of 12 kHz and a channel spacing of 30 kHz.

Table 10.2 shows how these channels are divided between the two carriers: A represents the non-wireline carrier and B represents the wireline carrier. Note that the frequencies assigned to each carrier are not all contiguous because of the extra frequencies added to the system in 1986. Note also the rather large separation (45 MHz) between base and mobile transmit frequencies. This allows for simple duplexers to separate transmit and

TABLE 10.2 North American Cellular Radio Frequencies

Base Frequencies (forward channels)	Mobile Frequencies (reverse channels)	Type of Channel	Carrier
869.040–879.360	824.040–834.360	Voice	A
879.390–879.990	834.390–834.990	Control	A
880.020–880.620	835.020–835.620	Control	B
880.650–889.980	835.650–844.980	Voice	B
890.010–891.480	845.010–846.480	Voice	A*
891.510–893.970	846.510–848.970	Voice	B*

Table denotes transmit carrier frequencies. Mobile transmits 45 MHz below base.
A = non-wireline carrier (RCC) B = wireline carrier (telco)
* = frequencies added in 1986

combined had only 33 available channels. A few IMTS systems are still in use, mainly in remote locations.

IMTS is capable of assigning channels automatically, by the rather simple means of transmitting a tone from the base station on unoccupied channels. The receiver in the mobile unit scans channels until it detects the tone. IMTS is capable of full-duplex operation using two channels per telephone call. Direct dialing is also possible, so using a mobile phone is almost as simple as using an ordinary telephone at home.

The main problem with IMTS and similar systems is that whatever bandwidth is made available to a single repeater, is tied up for a radius of perhaps 50 km or even more, depending on the height of the antenna and the power of the transmitter at the base station. Any attempt to reuse frequencies within this radius is likely to result in harmful interference. Simple systems like this also suffer from fading and interference near the edges of their coverage areas. For instance, suppose two similar trunked systems with identical repeaters are located 50 km apart. Then, at a location midway between the two, a receiver would receive equally strong signals from each. Communication would be impossible if both repeaters used the same frequencies.

10.3 Introduction to the Advanced Mobile Phone System (AMPS)

As the demand for mobile telephony grew, it became obvious that another way had to be found to accommodate more users. More spectrum near the existing mobile telephone allocations was already being used, but more spectrum had been assigned to this service than was actually channels, probably because of the popularity of cable television for specialty channels. A band approximately 40-MHz wide (increased to about 50 MHz in 1986) was assigned to the new system.

To make the system more efficient, cellular radio technology, based on many repeaters with their range deliberately restricted, was introduced at the same time. See Chapter 7 for a description of the cellular principle. Cellular radio had to wait until enough computing power could be introduced into the system, at reasonable cost, to allow the system to keep track of mobiles as they moved from one cell site to the next and to allow mobile phones to change frequency and power level by remote control from the cell site.

When cellular telephony was introduced in North America, it was [illegible] to allow two different companies, called carriers, to operate in any

receive signals in the phones. The base transmits to the mobile on a forward channel, while transmissions from mobile to base use a reverse channel.

An individual cell doesn't use all these channels, of course. Each cell has only one-seventh or one-twelfth of the total number of channels assigned to a carrier, depending on the system. Contiguous frequencies are not used in order to reduce interference. With a seven-cell repeating pattern, transmitters in the same cell are generally separated by about seven channels or 210 kHz. Each cell in a seven-cell pattern also has three of the 21 control channels.

To further reduce receiver selectivity requirements, adjacent channels are not used in adjoining cells. Therefore, transmitters in adjacent cells are separated in frequency by at least 60 kHz.

Channel Allocation

The control channels are used, among other things, to allocate voice channels to phones. When a user dials a phone number on a mobile phone and presses the *Send* button, the phone scans all the control channel frequencies to find the strongest. This control channel should be associated with the closest cell site. The cell phone transmits on its corresponding control channel, and once the call has been set up, the cell site assigns it a clear voice channel, assuming one is available.

While the conversation continues, the cell sites adjacent to the one in use monitor the signal strength from the mobile. When the strength is greater in one of the adjacent cells, the system transfers the call to that cell. This procedure is called a handoff. Handoffs, of course, require a change in frequency for the mobile phone, under control of the system.

A similar procedure takes place for incoming calls. The mobile periodically identifies itself to the system whenever it is turned on, so the system usually has a good idea of its location. Paging signals are sent out on control channels and the mobile responds, enabling the system to locate it more precisely. In the early days of cell phones it often took some time, a minute or more, to find a mobile, but improved communication within the system has reduced this time to a few seconds in most cases. The phone is instructed to ring, and once it is answered, the system assigns it a voice channel. After that the system follows the phone as it moves from one cell to the next, as explained earlier.

Frequency Reuse

The reason for the complexity of the cellular system is, of course, frequency reuse. Once a mobile has moved out of a cell, the frequency pair it occupied is available for another conversation. By making cells smaller, frequencies

Mobile transmitter power is controlled by the land station in 4 dB increments, with the lowest power level being –22 dBW (6.3 mW) ERP. The idea is to reduce interference by using as little power as possible. Mobile and transportable phones thus have better performance than portable phones only when propagation conditions are bad enough, or cells large enough, that the system needs to increase mobile power past the maximum for a portable phone. Using a portable phone inside a vehicle attenuates the signal considerably, so communication from a portable phone can sometimes be established in marginal areas by simply getting out of a car.

The cellular system has an identifying number called the *system identification number (SID)*. This enables the mobile phone to determine whether it is communicating with its home system or roaming. (Using a "foreign" system usually costs more and the user may disable this ability if desired.) In addition each cell site has a digital color code (DCC). When the mobile detects a change in DCC without a change in frequency, it is an indication that co-channel interference is being received from another base station.

Turning on a Phone

When a cell phone is turned on, it identifies itself to the network. First it scans all the control channels for its designated system (*A* or *B*) and finds the strongest. It looks for the SID from the system to determine whether or not it is roaming. If it does not receive this information within three seconds, it tries the next strongest control channel. After receiving the system information, the mobile tunes to the strongest paging channel. Paging channels are control channels that carry information about calls that the system is trying to place to mobiles. If someone is calling the mobile, its number will be transmitted by the paging channel.

The control channel constantly updates the status of its associated reverse control channel (from mobile to system). Only the system transmits on the forward channel, but any mobile can transmit on the reverse channel. The system tells the mobiles when this channel is busy to reduce the chance of a *collision*, which occurs when two or more mobiles try to use the control channel at the same time. After checking that the reverse channel is free, the newly activated phone transmits its ESN and MIN to the land station so that the system knows the phone is ready for calls and in which cell the phone is located. If the mobile loses the signal and reacquires it or detects that it has moved to a different cell, it identifies itself again. In addition, the system may periodically poll its mobiles to see which are still active.

While turned on but otherwise idle, the mobile phone continues to periodically (at least once every 46.3 ms) check the control channel signal from the cell site. It has to verify that a signal is still available, that it is from the same system, and that there are no calls for the mobile phone.

can be reused at shorter distances. There is no theoretical limit to this, but there are practical limits. As cells become smaller, more cell sites are needed and handoffs occur more frequently, requiring more computing power and faster response both at the system level and in the individual mobile phone. Once the radius drops below about 0.5 km the handoffs occur so frequently that it is difficult to cope with a mobile moving at high speed. The flexibility of cell sizes allows for larger cells in less-developed areas and smaller cells in areas of the greatest traffic.

AMPLE 10.3

A vehicle travels through a cellular system at 100 kilometers per hour. Approximately how often will handoffs occur if the cell radius is:

(a) 10 km?

(b) 500 m?

SOLUTION

The reason for the word "approximately" in the problem statement is that we are not sure how the road crosses the cell boundaries. Let us assume for simplicity that the vehicle drives along a road leading directly from one cell site to the next. Thus, the vehicle will change cells each time it travels a distance equal to the diameter of a cell (twice the radius).

First convert the speed into meters per second.

$$v = \frac{100 \text{ km/hr} \times 1000 \text{ m/km}}{3600 \text{ s/hr}}$$
$$= 27.8 \text{ m/s}$$

Now we can find the time between handoffs.

(a) Let the diameter be $d = 20$ km

$$t = \frac{d}{v}$$
$$= \frac{20 \times 10^3 \text{ m}}{27.8 \text{ m/s}}$$
$$= 719 \text{ s}$$
$$= 12 \text{ min}$$

(b) This time, $d = 1$ km

$$t = \frac{d}{v}$$
$$= \frac{1 \times 10^3 \text{ m}}{27.8 \text{ m/s}}$$
$$= 36 \text{ s}$$

10.4 AMPS Control System

In this section we study in more detail the process by which the AMPS cellular system keeps track of phones and calls. We need to look at the functions of the control channels and also at the control information that is sent over the voice channels.

An effective control system has to do several things. It needs to keep track of mobile phones, knowing which ones are turned on and ready to receive a call and where they are. It needs to keep track of telephone numbers for authentication and billing, and it should have some way to detect and prevent fraudulent use. It must be able to set up calls, both from and to mobile phones and transfer those calls from cell to cell as required. It would be best if all this were transparent to the user, who should only have to dial the phone number or answer the phone, just as with a wireline phone at home. A more advanced system might also be able to send faxes and e-mail, surf the Internet, and so forth, but let's look at the basics first.

First we need to understand the functions of the voice and control channels. You might assume that the voice channels are for talking and the control channels are for control signals, and that is mostly correct. However, there is a problem: cell phones contain only one receiver and one transmitter, so they can't receive both a voice channel and a control channel at the same time. Therefore, any control messages that have to be sent during a conversation must use the voice channel. Some of this is done using in-channel, out-of-band signaling (consisting of tones above the voice frequency range), and the rest is done with blank-and-burst signaling, during which the voice signal is muted for a short time (100 ms) while data is sent over the voice channel.

Digital signals on the control channel and those sent during blank-and-burst signaling on the voice channel are sent in a relatively simple way. They use FSK with 8 kHz deviation (16 kHz total frequency shift) and a channel bit rate of 10 kb/s. The data is transmitted using Manchester code. In

order to reduce the likelihood of errors, the control channel sends each message five times and also uses Hamming error-correction codes. This increases the robustness of the control system but reduces the actual data throughput to 1200 b/s. There is no encryption in the AMPS system: all the data coding information is publicly available. This is a serious oversight that has been remedied in the newer PCS systems to be described in the next chapter.

l Base :ation

Each mobile unit has two unique numbers. The mobile identification number (MIN) is stored in the number assignment module (NAM) in the phone. The MIN is simply the 10-digit phone number for the mobile phone (area code plus 7-digit local number), translated according to a simple algorithm into a 34-bit binary number. The NAM has to be programmable, since it may be necessary to assign a different telephone number to the phone, but it is not supposed to be changeable by the user. In most cases, however, it can be changed from the keypad if the user knows the right procedure. (Check the internet—it took the author less than ten minutes to find the procedure for his own cell phone.)

Usually a cell phone is registered on either the *A* or *B* system and has one MIN. It can operate on the other system as a roamer, if necessary and if there is an agreement between the two systems to allow it. It is also possible for a phone to have two MINs so that it can be used on both *A* and *B* systems without roaming. In that case the user of the phone has two phone numbers (and two bills to pay).

The other identification number is an electronic serial number (ESN), which is a unique 32-bit number assigned to the phone at the factory. It is not supposed to be changeable without rendering the phone inoperable, but in practice it is often stored in an EPROM (erasable programmable read-only memory chip) that can be reprogrammed or replaced by persons with the right equipment and knowledge. The combination of the MIN and the ESN enables the system to ensure proper billing and to check for fraudulent use (for instance, if a registered MIN appears with the wrong ESN the system will not allow the call to go through).

The mobile phone also has a number called the station class mark (SCM), which identifies its maximum transmitter power level. There are three power classes corresponding to phones permanently installed in a vehicle, transportable "bag phones," and handheld phones. The maximum power levels, specified as ERP (effective radiated power with respect to a half-wave dipole) are as follows:

Class I (mobile):	+6 dBW (4 W)
Class II (transportable):	+2 dBW (1.6 W)
Class III (portable):	–2 dBW (600 mW)

g a Call When the user of a mobile phone keys in a phone number and presses *Send*, the mobile unit transmits an origination message on the reverse control channel (after first checking that this channel is available). This message includes the mobile unit's MIN and ESN and the number it is calling. The cell site passes the information on to the mobile switching center for processing. Once authorization is complete, the cell site sends a message to the mobile on the forward control channel, telling it which voice channel to use for the call. It also sends the digital color code which identifies the cell site, and a Control Mobile Attenuation Code (CMAC), which sets the power level to be used. This power level can be changed by the land station as needed during the call by means of a control message on the forward voice channel.

Now both stations switch from the control channel to a voice channel, but the audio is still muted on the phone. The cell site sends a control message on the forward voice channel confirming the channel. It then sends a supervisory audio tone (SAT) on the voice channel to the mobile phone. This is a continuous sine wave, with a frequency above the voice band. There are three possible frequencies: 5970, 6000, and 6030 Hz. The mobile relays the tone back to the cell site. Reception of this tone by the cell site confirms that the correct cell site and mobile are connected. The mobile sends a confirmation message on the reverse voice channel. After this handshaking, the call can begin.

During the call the SAT continues (it is filtered out before the audio reaches the speaker in the phone, of course). Reception of the wrong frequency tone by the base station indicates an interfering signal and interruption of the tone indicates that the connection has been lost, perhaps due to severe fading. If the tone is not resumed within five seconds, the call is terminated.

A 10-kHz *signaling tone* (ST) may also be transmitted on the voice channel during a call. It is used to signal handoffs to another cell and the termination of the call.

g a Call An incoming call is routed by the network to the cell where the mobile last identified itself. (If it has not identified itself to the network, it is assumed to be turned off and a recorded message to that effect is given to the caller.) The land station sends the MIN on the paging channel along with the voice channel number and power level to use. The mobile confirms this message and sends its ESN on the reverse control channel to be matched by the network with the MIN. This is to avoid fraudulent use. The base sends its information again on the forward voice channel along with the digital color code information, then the mobile confirms the information on the reverse voice channel. After this handshaking, the supervisory audio tone is transmitted on the voice channel and the conversation can begin.

Handoffs The network monitors the received power from the mobile at adjacent cell sites during a call. When it detects that its strength is greater at an adjacent cell site than at the site with which it is communicating, it orders a handoff from one cell to the next. This always involves a change in channel, since to avoid co-channel interference the same channels are never used in adjacent cells. The order to do this is sent by the first cell site to the mobile on the forward voice channel using blank-and-burst signaling. The resulting 100 ms interruption in the conversation is barely perceptible. The voice channel must be used, because during a conversation the mobile is not monitoring any of the control channels. The mobile is given the new channel number, new attenuation code, and new SAT frequency. After confirmation on the reverse voice channel, the mobile switches to the new channel, which connects to the new cell site, and the conversation continues. There will probably be an audible disturbance while this occurs.

10.5 Security and Privacy

The AMPS system is not very private. Voices are transmitted using ordinary FM and conversations can be picked up with any FM receiver that will tune to the correct frequency. Base stations often repeat mobile transmissions, so quite often both sides of the conversation can be picked up with one receiver, just as with a cordless phone but from much greater distances (typically a few kilometers). The change in channels as a mobile is handed off does make it hard to follow conversations when the cell phone user is talking from a moving vehicle.

In 1988, in an attempt to increase cell phone privacy, the United States government banned the import or sale of scanners or other receivers that can tune to cellular frequencies. However, these are still legal in many countries (including Canada); there are millions of old scanners around and frequency converters are easy to build. It should therefore be assumed that AMPS voice transmissions are public. The transmission of confidential information, such as credit card numbers for instance, is not advisable with analog cell phones.

Stolen cell phones work only until the owner has the service cancelled. (There is a code to lock the phone, but most people don't bother, or they leave the password at the factory setting, such as "1234." Generally they do this because the password is supposed to be changeable only by the dealer, though in fact it can often be done from the keypad if one knows the correct method.) Even if the phone number is changed by a knowledgeable thief, the Electronic Serial Number will give away the fact that the phone is "hot." However, it is not impossible, at least on some phones, to change the ESN. It

is also possible to acquire valid pairs of MIN and ESN numbers by monitoring the reverse control channels. There is no encryption and the exact specifications for the data fields are publicly available. It is just a matter of acquiring the hardware and software to decode a 10 kb/s FSK data stream—not a very formidable task. Once the numbers are available it is possible to "clone" a cell phone to emulate a valid phone. Calls made on the cloned phone are billed to the unfortunate legitimate subscriber.

Service providers do have some protection. For instance, if the network detects the "same" phone trying to make two calls at once or two calls in quick succession from widely separated locations, it will flag the occurrence and someone will investigate. As the networks become larger and better integrated, this type of fraud becomes a little more difficult.

Another fraud is to use a cloned or stolen phone on another network as a roamer. If the foreign network is not capable of checking the phone's home network in real time, it may accept the call. This is becoming less likely as networks become better connected with each other. In the meantime, networks are becoming less trusting (especially in the United States) and less likely to allow roaming without identification.

10.6 Cellular Telephone Specifications and Operation

In this section we look at the requirements for the mobile or portable phone itself and consider some examples of phone construction. See Figure 10.4 for a block diagram of a typical analog cell phone. Because the system is full-duplex, the transmitter and receiver must operate simultaneously with a single antenna. A duplexer is used to separate the two signals. The wide 45-MHz frequency separation between transmit and receive frequencies makes this relatively easy. The constant frequency separation also simplifies frequency synthesizer design.

Microprocessor control is necessary to allow the phone to switch channels and power levels by remote control from the base station. The processor and its associated memory are also useful for timing calls, storing passwords to unlock the phone, storing lists of frequently-called numbers, and so on. Some cell phones can enter a sleep state between calls to conserve battery life; they must emerge from this state to check the control channel at least once in each 46.3 ms time period.

nitter r and iency In the previous section we noted that cell phones come in three *station classes*. This term refers to the maximum power level produced. The actual transmitted power level is adjusted in 4 dB steps by signals from the cell site. The mobile transmitter must transmit at within 3 dB of the correct power

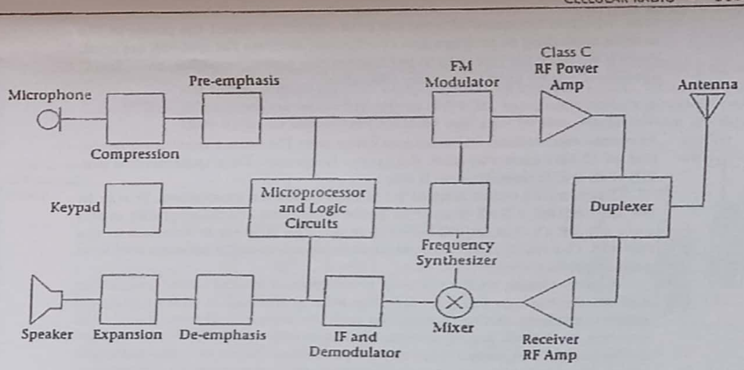


FIGURE 10.4 Block diagram of analog cell phone

level within 2 ms of turning on and must reduce its output to −60 dBm ERP or less within 2 ms of being turned off. The transmitted frequency must be within 1 kHz of the specified channel frequency.

The power levels for mobile, transportable, and portable phones are shown in Table 10.3. The abbreviation *MAC* refers to the *mobile attenuation*

TABLE 10.3 Power Levels for Mobile Phones (EIRP in dBW)

MAC	Class I	Class II	Class III
000	+6	+2	−2
001	+2	+2	−2
010	−2	−2	−2
011	−6	−6	−6
100	−10	−10	−10
101	−14	−14	−14
110	−18	−18	−18
111	−22	−22	−22

There is a range of 28 dB between maximum and minimum mobile power levels with a Class I phone.

code, which is transmitted from the base station to adjust the power of the mobile according to propagation conditions. Because FM and FSK are used, there is no need for linearity in the transmitter power amplifier, and Class C operation can be used for greater efficiency.

Limiter action As mentioned earlier, voice transmission uses FM with a maximum deviation of 12 kHz each way from the carrier frequency. Data transmission uses FSK with 8 kHz deviation each way.

Companding with a ratio of 2:1 is used in voice transmission. That is, in the transmitter, a 2 dB change in audio level from the microphone causes only a 1 dB change in modulation level. The process is reversed in the receiver. The result is an improvement in signal-to-noise ratio for low-level audio signals.

As with almost all FM systems, pre-emphasis is used in the transmitter and de-emphasis in the receiver. This means the higher audio signals are given more gain in the transmitter and correspondingly less gain in the receiver. In the cell phone system, all frequencies above 300 Hz are boosted at the transmitter, with a slope of 6 dB per octave. Figure 10.5 shows the pre-emphasis curve with the corresponding de-emphasis curve for the receiver.

FIGURE 10.5
Cellular radio pre-emphasis and de-emphasis

Note: This is an idealized (Bode Plot) representation. Real filters have a slight curvature in the response at both ends of the frequency range.

Mobile and Portable Antennas

Since the transmitted power of cellular phones is specified in terms of ERP, the use of more efficient antennas allows transmitter power to be reduced. This is especially important in the case of portable phones, since lower transmitter power leads to longer battery life. On the other hand, more efficient antennas tend to be larger.

Most portable cell phones use a quarter-wave monopole antenna. At 800 MHz, the length of this antenna is about 9.5 cm. The options are wider for mobile antennas. Many of these use a quarter-wave and a half-wave section, separated by an impedance-transforming coil. See Figure 10.6 for examples of typical portable and mobile antennas.

FIGURE 10.6
Portable and mobile antennas

10.7 Cell-Site Equipment

The radio transmitting equipment at the cell site operates at considerably higher power than do the mobile phones, but this power is shared among all the channels that are used at the site. Similarly, there must be receivers for each voice and control channel in use at the site, as well as extra receivers for monitoring the signal strength of mobiles in adjacent cells. Consequently the cell site equipment is much more complex, bulky, and expensive than the individual cell phones. In addition, as already noted, cell sites often need directional antennas to facilitate the division of each cell into sectors. Transmit and receive antennas may be separate or combined at the cell site. Often two receive antennas and one transmit antenna are used per cell, or per sector, in a sectorized configuration. This allows for space diversity in the receiver, to counteract the effects of fading, the receiver monitors the signal from both antennas and chooses the stronger signal. The effects of fading are likely to be greater for one antenna than the other at any given moment. Figure 10.7 is a block diagram of the equipment in a typical cellular base station.

Cellular radio base station

The combination of the mobile cellular phone and the cell-site radio equipment is known as the *air interface*. There is much more to cellular telephony than radio, however. The network must be organized and administered as a whole. This administration includes keeping track of phone locations, billing, setting up and handing off calls, and so on. The substantial computing resources required to do all this are, at least in part, responsible for the delay in the introduction of cellular telephony for many years after the idea was first proposed.

Figure 10.8 shows a typical cellular telephone system. Each cell has several radio transceivers (one per channel), including additional power amplifier is used to provide the transmit power for all channels in a site (or sector, for sectorized systems). The site's radio equipment is operated by a base station controller (BSC). The base station controller takes care of

FIGURE 10.8
Cellular radio system

the air interface: assigning channels and power levels, transmitting signaling tones, and so on. The mobile switching centers (MSCs), also called mobile telephone switching offices (MTSOs), route calls along a private copper, fiber optic, or microwave network operated by the cellular service provider.

Solution

(a) The average traffic is

$$T = 20{,}000 \times \frac{0.5}{24}$$
$$= 416\ E$$

The peak traffic is

$$T = 20{,}000 \times \frac{10}{60}$$
$$= 3333\ E$$

(b) The average traffic per cell is

$$t = \frac{416}{120}$$
$$= 3.47\ E$$

The peak traffic per cell is

$$t = \frac{3333}{120}$$
$$= 27.8\ E$$

(c) *Use the line from Table 10.4 corresponding to 33 channels, since this is the number available in a 12-cell repeating system. For average traffic at 3.47 E, the blocking probability is much less than 1% (in fact it is negligible, since the average number of calls is much less than the number of channels). At peak periods, however, the blocking probability increases to just over 5%.*

■

The other way to increase capacity is to increase the number of cells. The number of channels per cell remains the same as before, but since each cell covers a smaller area, with less potential traffic, the probabilities of cell blocking and call dropping are reduced. The downside of this, of course, is that the expense of the system increases with the number of cell sites, and more frequent handoffs occur, increasing the system overhead.

This reduction of cell size to increase traffic is called *cell-splitting*. Cell-splitting allows the network to begin with large cells throughout, with

Erlang's theory, this number of channels will accommodate about 23 E of traffic for one percent blocking.

The amount of traffic can be increased at the expense of a larger blocking probability. For instance, with 33 channels, a traffic level of 24.6 E can be accommodated with a two percent blocking probability. It might seem at first glance that a 7-cell repeating pattern can allow more traffic, but this is illusory. The cells are each divided into three sectors, using different channels. Therefore, each sector has only $395/21 \equiv 19$ duplex voice channels. Using the 7-cell pattern saves money by reducing the number of cell sites needed, not by increasing the number of channels.

Table 10.4 shows traffic levels in erlangs for 19 and 33 channels, with various blocking probabilities. It also shows traffic levels for larger numbers of channels. We'll use these with digital cell phones shortly.

TABLE 10.4 Cellphone Traffic in Erlangs per Cell or Sector

Number of Channels	Blocking Probability		
	1%	2%	5%
19	11.2	12.3	14.3
33	22.9	24.6	27.7
55	42.4	44.9	49.5
97	81.2	85.1	92.2

EXAMPLE 10.4

A cellular telephone system uses a 12-cell repeating pattern. There are 120 cells in the system and 20,000 subscribers. Each subscriber uses the phone on average 30 minutes per day, but on average 10 of those minutes are used during the peak hour.

Calculate:

(a) the average and peak traffic in erlangs for the whole system
(b) the average and peak traffic in erlangs for one cell, assuming callers are evenly distributed over the system
(c) the approximate average call-blocking probability
(d) the approximate call-blocking probability during the peak hour

Their action is also required in authorizing calls, billing, initiating handoffs, and so on. Sometimes the BSC and MSC are combined. Associated with the MSCs are data banks where the locations of local and roaming mobiles are stored.

At certain points in the system, the cellular network is connected to the public-switched telephone network (PSTN). These gateways allow calls to be made between landline and cellular phones, and between cell phones using different service providers. The cellular system communicates with the PSTN using Signaling System Seven, which was described in Chapter 5.

and
ting

The optimum size of a cell depends on the amount of traffic. Ideally, most of the available radio channels should be in use at peak periods, but situations where all channels are in use should be rare. If all channels in a cell are busy, it is impossible for anyone to place a call to or from that cell. The user has to hang up and try again later. This situation is called call blocking and is obviously undesirable. It causes revenue loss, and if it is frequent, unhappy customers may switch to the competing system. (This is always a possibility with the North American AMPS system since there are two competing systems in each area.) Call blocking takes place on the wireline network as well. For instance, long-distance trunks are sometimes unavailable during peak calling periods. This means that customers are used to blocking and will put up with a small percentage, perhaps one or two percent, of calls being blocked.

A more unpleasant situation occurs when a mobile phone moves into a cell that has all its channels busy. The attempt by the system to hand off the call to the new cell is frustrated by the lack of free channels, and the call must be terminated, or *dropped*. Dropped calls are very inconvenient and sometimes embarrassing, as the effect is the same as if one of the parties had hung up on the other. Call dropping is very rare on the wireline system, so customers are not as tolerant of this problem.

Since call blocking also occurs on the wireline network, which has been in operation for about a century, you might guess that someone has already studied the problem. You'd be correct: A. K. Erlang, a Swedish engineer, studied the problem using statistical analysis early in the twentieth century. He found, not surprisingly, that the more channels there were, the smaller the possibility of blocking for a given amount of traffic. Perhaps less obviously, he found that with more channels, the amount of possible traffic per channel increases for a given blocking probability. This phenomenon is called *trunking gain*, and it is the reason a two-provider system is theoretically less efficient than one using a single provider.

Trunking gain can perhaps be better understood by looking at an everyday situation: customers lining up to use tellers at a bank. Suppose there are two tellers and a separate line for each. Further suppose that the lines are assigned to customers on the basis of the type of accounts they have. Those with checking accounts use the first line, those with savings accounts use the second line.

Now suppose I arrive at the bank. My account is a checking account, but there are several people in line at that window. There is no line at the savings window, but I can't use that one. I am *blocked* and decide to try again later. Of course, after several frustrating attempts, I may notice that the competing bank across the street has shorter lines and change service providers. Now my bank changes its policy. There is only one line, and anyone can use the next available teller. The next time I arrive at the bank I have a much lower probability of being blocked. A similar logic applies to many situations: it is always more efficient to combine channels, and the gains are greater with more channels.

Phone traffic is defined in *erlangs* (E). One erlang is equivalent to one continuous phone conversation. Thus if 1000 customers use the phone ten percent of the time each, they generate 100 E of traffic on average. Mathematically,

$$T = Np \tag{10.3}$$

where

T = traffic in erlangs
N = number of customers
p = probability that a given customer is using the phone

We see immediately that traffic analysis does not coordinate channels directly with the number of customers. Real customers may use the phone continuously for an hour, then not at all for the rest of the day. If enough of them use the phone at the same time (afternoon rush hour, for example, is a peak in cell phone usage in cities), the peak traffic will be much greater than the average. Some cell phone owners use the phone continually for business during the working day; others use it mainly for emergencies and generate very little traffic. Also, the usage patterns may vary, in response to changing rates, for instance (lower rates generate more traffic). Evening and weekend use tends to be light for cell phones, just as it is for wireline long distance, and both wireline and cellular providers commonly provide monetary incentives to customers to shift some of their usage to those periods.

The most obvious way to avoid call blocking and call dropping is, of course, to provide more channels. However, the number of channels for the system is fixed, and so is the number per cell once the repeating pattern, usually seven or twelve cells, has been established. In a 12-cell repeating pattern, each provider has $395/12 \equiv 33$ voice channel pairs per cell. According to

FIGURE 10.9
Microcell site
(Courtesy of Bell Mobility)

the cell size decreasing in high-traffic areas as the traffic increases. Cellular telephone infrastructure is expensive, but it does not all have to be built at once.

Cell-splitting allows spectrum space to be used more efficiently, but it is not particularly cost-effective in terms of equipment. Doubling the number of cells doubles the system capacity with the same bandwidth allocation, but it also doubles the number of cell sites, roughly doubling the system cost.

icrocells, cells and tepeaters

Cell-splitting, as described in the preceding paragraphs, can be used to increase the capacity of a cellular system. At a certain point diminishing returns set in, however. Cell sites are expensive, and the increase in capacity does not justify the increase in cost for very small cells. Another problem is that real estate costs are highest in the areas where demand is greatest, and the use of small cells means less choice in cell-site location and thus higher costs for access to the sites. Finally, there is increased load on the switching system due to the increase in the number of handoffs with small cells.

In high-demand areas, microcells are often used to help relieve the congestion at relatively low cost. A microcell site is a very small unit that can mount on a streetlight pole. The microcell antenna is deliberately mounted lower than the tops of nearby buildings to limit its range. A typical microcell covers about 500 meters of a busy street, but has very little coverage on side streets. See Figure 10.9 for a typical unit.

FIGURE 10.10
Overlay of microcells and macrocells

Because microcells have such small, narrow patterns, it is difficult to obtain general coverage this way. Consequently the original larger cells (*macrocell(s)*) are left in place, so that calls can be handed off between microcells and conventional cells as required. The microcells must use different frequencies than the overarching conventional cells, of course; this is accomplished by assigning to the microcells some of the channels that were formerly used by the macrocells.

A microcell is often under the control of a conventional cell site, with which it usually communicates by microwave radio. The microcell may itself be divided into several zones. Figure 10.10 illustrates how microcells and conventional cells (sometimes called *macrocells*) can work together.

Sometimes the problem is not excessive traffic, but a "hole" in the system coverage caused by propagation difficulties, such as a tall building or hill that casts a radio shadow. Obviously cell sites should be chosen to minimize these problems, but it is not always possible to eliminate them completely. In that case a *repeater* can be used, as shown in Figure 10.12. The repeater simply amplifies signals from the cell site and from the mobiles.

FIGURE 10.12
Cellular repeater

In order to save costs, many microcell sites are not true transceivers but are only amplifiers and frequency translators. Figure 10.11 on page 396 shows the idea. The main cell site up-converts the whole transmitted spectrum to microwave frequencies. The transmitter at the microcell site simply down-converts the block of frequencies to the cellular band and amplifies it. No modulation or channel switching is required at the microcell site. Similarly, the microcell's receiver consists of a low-noise amplifier that amplifies the entire frequency range in use at that site plus an up-converter to convert the signal to microwave frequencies. All demodulation is handled at the main cell site.

Sometimes cellular radio signals are too weak for reliable use indoors or in well-shielded areas like underground concourses. This is especially true in very small cells called *picocells*. When reliable indoor reception is needed, sometimes the PCS systems to be described in the next chapter but are sometimes used with AMPS. These are more common with the outdoor cells in the same area if the attenuation of the structure is sufficient. This is the case in underground malls, for instance.

Indoor picocells can use the same frequencies as the outdoor cells in the

RE 10.11
-cell block
am

It can be connected to the main site by a microwave link, but often the repeater simply receives and transmits at the same frequencies, avoiding feedback by careful location of directional transmit and receive antennas.

10.8 Fax and Data Communication Using Cellular Phones

As can be seen from the foregoing, AMPS is a circuit-switched analog system designed for voice communication, as was the wireline system in its original form. Consequently, the most straightforward way to send data, as with the conventional wireline system, is to use modems at each end of the link. It is also possible to send packet-switched data using AMPS, as will be described shortly.

Cellular Modems

The main differences between wireline phone service and analog cellular phones, for modem use, is that cell phone connections tend to be noisier and are subject to interruption during handoffs and fading. These interruptions, though brief in human terms (usually on the order of 100 ms to about 2 s), result in the loss of a considerable amount of data and possibly in a dropped connection. Consequently, the error-correction schemes in modems for cell phone use should be more robust than is necessary for wireline operation. Of course, this greater robustness results in slower data transmission. In addition, the modem must be set up not to require a dial tone before dialing. The dialing connections must be made separately to the phone; the situation is more complex than just plugging in a phone jack, as for a landline modem.

Many (but not all) modem cards for notebook and laptop computers will work with cell phones. Similarly, many, but not all, cell phones can be used with modems. A proprietary cable is required to connect modem to phone. It is also possible to find cell phone cards which plug into a notebook computer and enable it to send data via cellular radio.

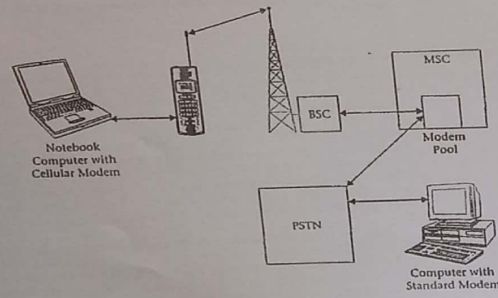
Cellular modems are advertised as having speeds of up to 28.8 kb/s but actual speeds are usually 9600 b/s or less. Performance is improved by operating from a stationary vehicle, as this eliminates handoffs and reduces fading.

An error-correcting protocol called MNP10 is usually used with cellular connections. It must be used at both ends of the connection. MNP10 incorporates some special cellular enhancements. For instance, rather than

trying to connect at the maximum speed and reducing speed if necessary, an MNP10 modem starts at 1200 b/s and gradually increases the speed. Multiple attempts are made to connect, and the packet size starts small and increases provided error rates remain low.

Another protocol, developed by AT&T, is called *Enhanced throughput cellular* (ETC). It has the advantage of being able to work with a conventional wireline modem at the other end of the connection while retaining at least some of its benefits.

In order to allow communication between cellular modems and conventional wireline modems which may lack these more robust protocols, some cellular providers maintain a modem pool. This allows for conversion between cellular and wireline standards. Generally, the cellular user accesses this service by dialing a special prefix before dialing the telephone number of the landline modem. See Figure 10.13 for the idea.



Facsimile transmission is also possible with cell phones. A fax modem and notebook computer can be used, or a conventional fax machine can be used with a special adapter. The adapter allows the cell phone to simulate a conventional two-wire telephone line with dial tone. Fax performance is much better from a stationary vehicle, since no special protocol is used.

Cellular Digital Packet Data (CDPD)

The previous section describes how data can be sent over a cellular voice channel in a manner similar to that used with landline telephony. That procedure is relatively simple and requires no prior arrangement with the cellular service provider, but it tends to be expensive, as full airtime costs, plus long-distance costs if applicable, must be paid for the entire time the call is connected.

Another way exists to send data over the AMPS cellular radio system. The **Cellular Digital Packet Data (CDPD)** system uses packet-switched data and tends to be less expensive than using a cellular modem, especially when data needs to be transmitted in short bursts. On the other hand, a separate account is required, and the cellular system has to be specially configured to use CDPD.

The principle behind CDPD is that at any given moment there are usually some voice channels in an AMPS system that are not in use. The CDPD system monitors the voice channels, using those that are idle to transmit data. When traffic is detected on the voice channel, the data transmissions cease within 40 ms. Since this is less than the setup time for a voice call, the voice customer is not affected. Users, once registered with the CDPD system, can transmit data as required without maintaining a continuous connection and tying up an expensive pair of voice channels.

The bit rate in the RF channel for CDPD is 19.2 kb/s, achieved using Gaussian minimum-shift keying (GMSK), a form of FSK. When overhead is taken into account, the maximum data rate is comparable to that obtained with a 14.4 kb/s modem—slow by current wireline standards, but not too bad for wireless. When the network is busy, the throughput is lower, as packets are stored and forwarded when a channel becomes available.

10.9 Digital Cellular Systems

Until recently, digital voice communication used more bandwidth than analog. The conventional wisdom was that digital was preferred where bandwidth was not an issue, because of its flexibility and immunity to the accumulation of noise, but that analog was better when bandwidth was constrained. For instance, you will recall that wireline telephony uses 64 kb/s for each one-way voice channel. In order to transmit this data rate in a 30-kHz channel, an elaborate modulation scheme would be needed. This would not be robust enough for the mobile radio environment with its noise and fading. Consequently, all first-generation cellular radio systems use analog modulation schemes. FM needs more bandwidth than AM and its variants, but it was found that FM's resistance to noise and interference more than made up for the additional required bandwidth.

still have one of these (the author does). Making the analog phones obsolete would probably obligate the operators to provide new digital phones to their customers at no charge.

For these reasons, the system that evolved in North America uses the same radio frequencies, power levels, and channel bandwidths as AMPS. Some of the existing analog channels were converted to digital, leaving some analog channels in place in every cell for coexistence with analog equipment. The reduction of bandwidth requirements for digital communication allowed those analog channels that were converted to digital to communicate three communication channels on one RF channel using TDMA, as described in the following section. The new system is so similar to AMPS that sometimes it is referred to as *digital AMPS (D-AMPS)*.

The situation in Europe was completely different. Almost every country seemed to have a different analog cell phone scheme. Some had more than one! There were even several different frequency bands in use in different countries. The European Community made the radical decision to scrap the entire analog infrastructure and begin again with a common digital system. This system is called *GSM (Global System for Mobile)*. It is not used in the cellular radio bands in North America but it is one of three systems in use for North American PCS, so GSM will be described in the next chapter.

Other countries varied in the type and extent of analog cell phone penetration. The result is a worldwide mixture of incompatible analog and digital formats, which will probably continue at least until the third generation of wireless telephony.

Conversion of AMPS to TDMA

The compatibility requirements outlined in the previous section constrained the development of a digital cellular radio system for North America. It was decided to combine three digital voice channels into one 30-kHz radio channel using TDMA. This is known as a half-rate TDMA system. The first such system went into operation in 1990. The specifications also encompass a half-rate system with six voice channels in one 30-kHz slot to be implemented at a later date when vocoder technology has improved. The digital system would seem to be able to carry three times as much traffic as the analog system, but, due to trunking gain, the actual increase for a given level of blocking is greater. In the last two lines of Table 10.4, the traffic for various blocking probabilities is shown for both 7-cell sectorized and 12-cell repeating systems. The numbers of voice channels are calculated on the basis that one analog channel, for backward compatibility, is available in each cell or sector. Note that the traffic capacity is more than tripled.

At first, the AMPS control channels were left alone, and only the voice channels were digitized. This digital specification is known as IS-54B. A later

Recent advances in data compression and voice coding, as discussed in Chapter 3, have reversed the conventional wisdom about bandwidth. It is now possible to transmit a digitized voice signal in less bandwidth than is required for an analog FM signal. This removes the last big obstacle to digital communication. Digital systems still tend to be more complex than analog, but large-scale integrated circuits have made it possible to build complex systems at low cost and in small packages. In fact, it appears likely that the analog cellular radio system just described will be the last of the great analog communication systems. Soon there will be no reason to use analog except for compatibility with legacy systems and perhaps for very low-cost communication devices.

Analog cellular radio can be seen as the first generation of wireless communication. Digital systems, which we begin to examine here and continue in the next chapter, are the second. The third generation will undoubtedly be digital. Its exact makeup is being debated at this time. A synopsis of the progress thus far can be found in Chapter 14.

Digital Cellular Radio

The main incentive for converting cellular radio to a digital system was, as suggested earlier, to reduce the bandwidth requirements, allowing more voice channels in a given spectrum allotment. Other reasons also exist. Digital systems have more inherent privacy than analog, being harder to decode with common equipment. They also lend themselves to encryption, if required. Note that analog AMPS already uses digital signaling data. The fact that it is not encrypted is due to oversight: the system designers underestimated the ingenuity of hackers.

Digital communication systems can use error correction to make them less susceptible to noise and signal dropouts. They lend themselves to time- and code-division multiplexing schemes, which can be more flexible than the frequency-division multiplexing used in analog systems. Digital signals are easier to switch: in fact, most of the switching of analog telephone signals, including AMPS cellular telephony, is done digitally after analog-to-digital conversion.

The gradual conversion of the world's cellular radio systems to digital form has been accomplished differently in various parts of the world. In North America, which is the focus of this book, there existed one analog system with a very large installed base. Most of this infrastructure was not yet fully paid for, so operators were understandably reluctant to replace it. The requirement was for a digital system that would allow as much as possible of the analog equipment to be retained.

There were also millions of analog phones in consumers' hands. Many of these also were not yet paid for, especially since operators had given many analog phones to customers in exchange for service contracts. Perhaps you

modification (1991), called IS-136, added high-speed digital control channels for better security and additional features. IS-136 is used both in the 800-MHz cellular radio band and the 1900-MHz PCS band.

IA Voice Channel

Figure 10.14 shows how the RF voice channel is divided in the TDMA system. There are 25 frames per second so each frame is 1/25 s = 40 ms in length. Each frame has 1944 bits so the total bit rate for the RF signal is 1944 × 25 = 48.6 kb/s. Phase-shift keying with four levels (π/4 QPSK) is used, so there are two bits per symbol and the baud rate is 24.3 kbaud. Note that this is a data rate of 48.6/30 = 1.6 b/s per hertz of bandwidth. This is quite conservative, necessarily so because of the radio environment, which is subject to noise, interference, and very deep fading.

FIGURE 10.14
TDMA frame

40 ms
1944 bits

Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6
Voice A	Voice B	Voice C	Voice A	Voice B	Voice C

6.67 ms
324 bits

Each frame has six time slots lasting 40 ms/6 = 6.67 ms and containing 1944/6 = 324 bits each. For full-rate TDMA, each voice signal is assigned to two time slots as shown. Six voice signals, occupying one slot each, can be accommodated with half-rate TDMA. For the full-rate system, speech data corresponding to 40 ms of real time is transmitted in 2 × 6.67 ms = 13.3 ms. As with analog AMPS, the TDMA system uses separate RF channels for transmit and receive.

Speech encoding (see Chapter 3) is used to limit the bit rate to approximately 8 kb/s for each speech channel for the full-rate system. The full-rate system allocates two noncontiguous time slots to each voice channel: slots 1 and 4 for the first, 2 and 5 for the second, and 3 and 6 for the third. In addition, the bits corresponding to each 20 ms of speech are divided among two time slots. Interleaving the data bits in this way reduces the effect of burst errors.

Overhead reduces the number of data bits available per time slot to 260. The data rate available for each voice channel is 260 bits/20 ms = 13 kb/s. The voice is actually encoded at 7.95 kb/s and the remaining bits are used for error correction. The half-rate system will use 4 kb/s for voice coding.

The frames as just shown are similar for both forward (base to mobile) and reverse (mobile to base) channels, but the composition of the time slots differs. The frames are synchronized for the forward and reverse channels, but the timing is offset so that a frame starts 90 bits (1.85 ms) earlier at the mobile. A mobile transmits during two of the six time slots and receives on a different two slots. The remaining two time slots are idle: the phone may use these to check the signal strength in adjacent cells to assist in initiating a handoff. This technique allows the digital cell phone to have only one transmitter and one receiver, just as for AMPS. In fact the RF situation is a little simpler, since there is no need for the mobile to transmit and receive simultaneously. (Since digital cell phones have to work with the analog AMPS system as well, this unfortunately does not really simplify the design, as a duplexer is still needed for analog operation.)

Each slot contains 324 bits for both forward and reverse channels. The allocation of these bits is different for the forward and reverse links, however. See Figure 10.15 for an illustration of the differences.

In particular, the mobile needs time to turn its transmitter on for each transmit time slot, since to avoid interference it must be off when the mobile is not scheduled to transmit. Six bit periods (123 μs) are allocated to this. The base station transmitter is on all the time, since it uses all six time slots to

Synch 28 bits	SACCH 12 bits	Data 130 bits	CDVCC 12 bits	Data 130 bits	Reserved 12 bits

Forward Channel (Base to Mobile)

6.67 ms
324 bits

Guard 6 bits

Guard	Ramp	Data 16 bits	Synch 28 bits	Data 122 bits	SACCH 12 bits	CDVCC 12 bits	Data 122 bits

Ramp 6 bits

Reverse Channel (Mobile to Base)

FIGURE 10.15 TDMA voice time slots

Dr.G.Arumugam M.Sc., B.Ed., M.Phil., Ph.D.

transmit to three mobiles on the same RF channel. The mobile also waits for a guard time (another six bits) to pass before transmitting. This is necessary because different amounts of propagation delay could cause mobile transmissions (on the same RF channel but in different time slots) to overlap when received at the base station.

IPLE 10.5

Calculate the maximum distance between base and mobile that can be accommodated with a guard time of 123 μs.

SOLUTION

The signal from base to mobile is delayed by the propagation time between the two. This causes the mobile's synchronization to be off by that much time, and it will be late in starting its transmission. In addition, the signal is further delayed by the propagation time from mobile to base. Therefore, the 123 μs allowance must include the round-trip propagation time. Since radio waves propagate at the speed of light, the maximum total round-trip distance is:

$$\begin{aligned} d &= ct \\ &= 300 \times 10^6 \text{ m/s} \times 123 \times 10^{-6} \text{ s} \\ &\approx 36.9 \text{ km} \end{aligned}$$

The maximum one-way distance is half that, or about 18.5 km. For even larger cells, the mobile can instruct the base to advance its transmission by up to 30 bit periods (617 μs).

In addition to voice samples and their associated error correction, the digital traffic channels contain synchronizing, equalizer training, and control information.

The *Coded Digital Verification Color Code* (*CDVCC*) provides essentially the same information as the Supervisory Audio Tone (SAT) in AMPS. The Slow Associated Control Channel (SACCH) provides for control signal exchanges during calls and essentially replaces the blank-and-burst signaling in AMPS, though there are provisions to "steal" voice data bits for additional control information as required. These stolen bits form the Fast Associated Control Channel (FACCH), which is used for urgent information such as handoff commands. The *coded digital locator* (*CDL*) field tells the mobile where to find digital control channels, if available.

TDMA Control Channels

As mentioned, there are two "flavors" of TDMA cellular radio in use. The earlier specification, called IS-54B, uses the same control channels and formats as AMPS. These are called *Analog Control Channels* (*ACCH*) because of their association with the analog system, but as noted earlier, they are actually digital, using FSK and a channel data rate of 10 kb/s.

The IS-136 specification incorporates separate control channels for the digital system. These are called *Digital Control Channels* (*DCCH*) to distinguish them from the older type. *Digital control channels consist of pairs of slots on the same RF channels that are used for voice. The DCCH can be assigned to any RF channel; it does not have to be one of the 21 control channels used in the analog system. As with the voice channels, separate forward and reverse channels are needed. Normally there is one DCCH pair per cell, or per sector in a sectorized system. See Figure 10.16.*

FIGURE 10.16
TDMA digital control channel

40 ms
1944 bits

Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6
Voice A	Voice B	Control	Voice A	Voice B	Control

6.67 ms
324 bits

The total bit rate for a DCCH is one-third of the RF channel bit rate, or 44.6/3 = 14.9 kb/s, compared with 10 kb/s for an ACCH. This extra capacity makes the digital control channels useful for many added features, such as call display and short text messages. As with the analog system, digital control channels are used in setting up calls. They cannot be used during a call, since the single receiver in the mobile unit is otherwise occupied.

Privacy and Security in Digital Cellular Radio

Privacy is considerably improved in digital cellular radio compared to the analog system. Ordinary analog scanners can make no sense of the digitized voice signal. Even decoding it from digital to analog is not straightforward, due to the need for a vocoder. However, obviously vocoders are present in all digital cell phones, so a modified cell phone could do the job.

There is some encryption of the authorization information in the TDMA system, enough to make cell phone cloning and impersonation difficult. In general the level of security of the TDMA system is considerably better than with AMPS and is probably adequate for general use.

Dr.G.Arumugam M.Sc., B.Ed., M.Phil., Ph.D.
Assistant Professor of Physics
PG & Research Department of Physics
Annai Vailankanni Arts & Science College
Thanjavur - 7. Cell : 9750953698

11.1 Introduction

As pointed out in Chapter 10, the term *personal communication* is somewhat vague. In general, it refers to direct communication between people rather than between places. To understand the concept, consider Joe, a typical present-day North American. Joe has a home phone and an office phone. The home phone has an answering machine and the office phone has voice mail, but there is no connection between them. Joe also has a cell phone in his car. That system does not have voice mail, though it could if he wanted to pay extra. If you want to talk to Joe, you really call a place rather than a person. You decide if he's more likely to be at home, in the office, or in the car, and place your call accordingly. If he's in none of those places, you leave a message. Actually, you probably leave two messages, one at home and one at work.

Now consider Joan, a slightly more modern North American. Joan also has office and home phones, but her cell phone is a small portable phone which she usually carries with her. However, she has to pay airtime on the cell phone (but not on the others) so she only gives out that number to close friends and important business associates. Those people, trying to prevent Joan from having to pay airtime, usually try her home and office numbers first, then the cellular number. Joan is halfway toward a true PCS.

Move forward a little to the future. Ted has only one phone. It's small enough for him to slip into his shirt pocket, and he carries it with him all day. He has only one phone number. If you want to talk to Ted, you call that number. If he cannot take calls for a while, he turns off the phone and voice mail takes over. Similarly, if Ted wants to talk to someone, he uses his pocket phone—no matter where he is. Ted's phone number is no longer associated with a place; it is associated with a person.

It is possible to live like Ted today, but for most people it is tricky. How does the switchboard at Ted's company find his personal phone? How does he charge business calls to the company? How does he get a listing in the phone book? What happens when he is out of range of his service provider? All these problems are solvable, and there are many people who feel that, eventually, all phones will be wireless. Your pocket phone will connect to a wireless PBX at your office when you're at work, to a telepoint in your house when you're at home, to a cellular or PCS network when you're out, and to a satellite-based telephone system when you are at a remote location.

We are still some distance from true PCS, but the gap is closing. The systems described in this chapter are playing a major role in closing that gap. Perhaps it is appropriate, then, that the name for the larger concept is often applied to the systems described in this chapter. These so-called personal communication systems are derived from the cellular concepts introduced in Chapters 7 and 10, with enhancements to allow the phones to be smaller and lighter, to have improved battery life, and to have extra features not available in first-generation cellular systems.

There are three competing types of PCS in North America. This contrasts vividly with Europe, which simply extended its GSM digital cell phone system to a higher frequency range for PCS. We have already heard of two of the three North American personal communication systems. One is in fact the GSM system, which was mentioned, though not discussed, in Chapter 10. The second is IS-136, the North American digital cell phone standard, which has also been extended for use at PCS frequencies. The third is a direct-sequence spread-spectrum system developed in the United States by Qualcomm and known as IS-95, or by its tradename, CDMAone™. In the next section we look at some features that all these systems have in common and contrast them with conventional analog AMPS. Following that, we look at each of the three systems in more detail.

11.2 Differences Between Cellular Systems and PCS

Though based on the same cellular idea as the first-generation cell phone systems described in Chapter 10, PCS have significant differences which justify the use of a different term. You should realize, however, that many of the differences are transparent, or at least not immediately obvious, to the user. The systems described in this chapter are often called *second generation* personal communication systems; in other words, the analog cell phone system is really the first generation of PCS. The third generation, now being designed, will feature much wider bandwidth for high-speed data communication and it will be discussed in Chapter 14.

Frequency Range

One of the reasons for establishing new PCS was that the cellular frequency bands were becoming crowded, especially in major metropolitan areas. There was no room for expansion in the 800-MHz band, so the new service was established in the 1900-MHz band (1800 MHz in Europe). This has advantages in terms of portable antenna size. A few years ago, electronics for this frequency range would have been prohibitively expensive, but advances in integrated circuit design have reduced the cost penalties.

In North America the broadband PCS band consists of 120 MHz in the 1900-MHz region. The term *broadband* here is relative. It refers to bandwidth sufficient for voice communication and distinguishes this service from such narrowband services as paging, which will be discussed later in this book. Sometimes the term *broadband communication* is used to refer to video and high-speed data; that is *not* the sense in which it is used here.

See Table 11.1 for the PCS band plan. Note that there are six frequency allocations, so up to six licenses can be awarded in any given area. There are three 30-MHz and three 10-MHz allocations. The reverse channel or uplink (mobile to base) is 80 MHz above the forward channel or downlink (base to mobile) frequency. Reverse and forward channel allocations are separated by a 20-MHz band, from 1910 to 1930 MHz, which is allocated for unlicensed services like short-range voice communication. In the United States the frequencies have been assigned by auction; in Canada licenses were allocated after public hearings. Some PCS carriers are established cellular providers with 800-MHz licenses; others are new to the field of wireless communication.

TABLE 11.1 Broadband PCS Band Plan

Allocation	Base Transmit (Forward Channel or Downlink)*	Mobile Transmit (Reverse Channel or Uplink)*
A	1850–1865	1930–1945
B	1870–1885	1950–1965
C	1895–1910	1975–1990
D	1865–1870	1945–1950
E	1885–1890	1965–1970
F	1890–1895	1970–1975

*Frequencies are in MHz

Cell Size Cellular telephony was originally conceived as a mobile radio system, with phones permanently mounted in vehicles. These phones use efficient external antennas on the roof of the vehicle and have a maximum ERP of 4 W. However, in recent years portable cell phones have outsold mobile phones by a considerable margin. This has implications for the system, as portable phones have lower power and are often in difficult locations—from a propagation point of view—such as inside vehicles and buildings. For these reasons, portable AMPS cell phones may not work reliably near the edges of the larger cells. PCS, on the other hand, were designed from the beginning with handheld phones in mind. At first it was thought that most PCS users would be on foot, but it is now quite obvious that subscribers expect to use the phone wherever they are: outdoors, indoors, in an underground shopping mall, or in their cars.

PCS cells are typically smaller than AMPS cells to accommodate more traffic and low-power handheld phones. They must hand off calls very quickly to handle users in moving cars.

All-Digital System Because of technical constraints, all first-generation cellular systems are analog, though some progress has been made in converting them to digital technology. In fact, some providers have marketed 800-MHz digital cell phones as "PCS" systems.

Current digital technology is more efficient than analog FM in its use of bandwidth. It also allows lower power consumption in the portable phone and more advanced data communication and calling features. Security and privacy are inherently better with any digital system, since ordinary scanners cannot be used to intercept calls, and digital coding schemes can also incorporate encryption as required.

There is one major problem with North American digital PCS. Whereas first-generation cellular systems in North America all use the same analog technology, there are three incompatible digital systems in North America. This makes roaming more difficult with PCS than with cellular phones. Many providers and phone manufacturers have solved this problem by offering dual-band, dual-mode phones that are capable of both PCS and analog cellular operation. The solution is not ideal, because it results in phones that are larger and more expensive than they would otherwise have to be.

Extra Features AMPS systems were designed with POTS (*plain old telephone service*) in mind. Even features commonly found on wireline phones, such as call display, present problems in AMPS. Digital systems allow a substantial amount of data transmission in their control channels, making all sorts of enhancements possible. In addition to obvious features like call display, digital systems can allow short printed messages, and even e-mail and limited web browsing are possible without additional modems and computers. The features available and the way they are implemented vary with the type of PCS, and we will look further into this later in this chapter.

Coverage At least at present, the coverage for any PCS is much less universal than it is for AMPS cell phones. This will undoubtedly change in the future, as the systems acquire more customers and build more infrastructure. In the meantime PCS users have to pay more attention to local coverage areas than do analog cell phone users.

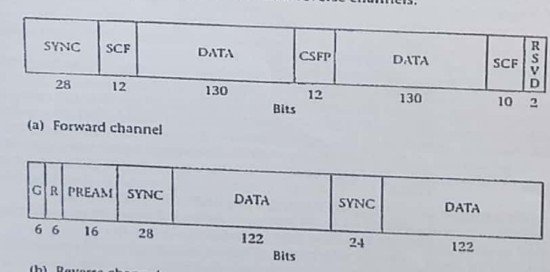
Structure One of the arguments for PCS is that they should be less expensive than analog cellular radio. The utilization of spectrum space is more efficient, for example. In practice, rates tend to be set by a combination of market forces. The analog systems have a head start in paying for their infrastructure and have been able to lower prices to match PCS in many cases.

11.3 IS-136 (TDMA) PCS

We looked at IS-136, the North American Digital Cellular standard, in Chapter 10. Most people just refer to it as *TDMA* (*time division multiple access*) when they are talking about PCS, though GSM is also a TDMA system. The most important difference between the 800-MHz and 1900-MHz versions of TDMA is that there are no analog control channels in the PCS bands. Rather than go over the ground already covered in Chapter 10, in this chapter we will take a closer look at the digital control channel and consider how enhanced services are provided. Much of this material is very similar for the GSM system, described next.

MA Digital l Channel Recall from Chapter 10 that the digital control channel (DCCH) uses two of the six time slots in a TMDA frame (slots 1 and 4, to be precise). Normally only one DCCH is required per cell or sector. Figure 11.1 shows how the time slot is divided up for both forward and reverse channels.

IRE 11.1
A digital
rol channel



Let us look at the forward channel first. The *SYNC* (*synchronizing*) bits have the same function as for the voice channels, allowing the mobile receiver to lock on the beginning of the transmission. The *SCF* (*shared channel feedback*) bits perform several functions. They provide acknowledgement of messages from mobiles and inform the mobiles of the status of the reverse control channel. Just as in analog AMPS, the forward channel is under the control of the base station, but many mobiles share a single reverse control channel. By monitoring the status of the reverse channel as reported by the base in the SCF field, the mobiles can reduce the possibility of collisions. These still occur occasionally; however, in that case, the message from the mobile will not be acknowledged by the base, and the mobile will try again after a random delay time.

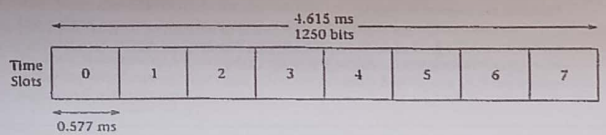
The *CSFP* (*coded superframe phase*) bits identify the location of this time slot in a larger frame that extends over 16 TDMA frames or 32 blocks of control-channel data, representing a time period of 640 ms. Each block is designated as containing broadcast, paging, messaging, or access response information. Each of these types of data can be considered a separate logical channel of data, time-division multiplexed with the other types and with voice as well, since four of every six timeslots on the RF channel that carries the control channel are still used for voice. The number of control-channel blocks assigned to each type of use can be varied within limits. Table 11.2 on page 424 summarizes the logical channels, and a brief description of each is given below.

Two superframes comprise a *hyperframe*. The hyperframe structure allows data to be repeated. This means that a mobile receiver can check the signal strength on other channels, without missing data. It also provides redundancy: if the mobile misses some data because of a burst error, it gets a second chance.

The broadcast channel contains information intended for all mobiles. It is divided into two components. The *fast broadcast channel* (*F-BCCH*) is used to transmit system parameters to all the mobiles. These include the structure of the superframe itself, the system identification, and registration and access parameters. All of these must be communicated to the mobile before it can place a call, so all of this information is transmitted at the beginning of each superframe. The *extended broadcast channel* (*E-BCCH*) has less critical information, such as lists of the channels used in neighboring cells. This information can be transmitted over the course of several superframes.

The *short message service, paging, and access channel* (*SPACH*) is used for control messages to individual telephones and for short paging-type messages to be displayed on the phone. It is not necessary for every phone to monitor all these messages; the phone is told which block to monitor and can go into an idle or *sleep* mode the rest of the time while it waits for a call. This helps to extend battery life.

.2



to use a whole RF channel for this. Instead, one of the eight time slots on one RF channel in each cell or sector is designated as a control channel. The broadcast information is transmitted first, followed by paging information. See Figure 11.3 for an illustration.

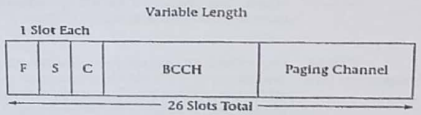


FIGURE 11.3
GSM control channel

F = Frequency correction (sine-wave carrier)
S = Synchronizing
C = Contents list
BCCH = Broadcast control channel

The BCCH and PCH are forward channels only. The corresponding reverse channel is called the *random-access channel* (*RACH*) and is used by the mobiles to communicate with the base. Mobiles transmit on this channel whenever they have information; if a collision occurs, the mobile waits a random time and tries again. Transmissions are shorter than the duration of the slot to prevent interference caused by the propagation delay between mobile and base. The delay problem is avoided on the traffic channels, because the base instructs the mobile to advance or retard the timing of its transmissions to compensate for the changes in propagation delay as it moves about in the cell.

Just as with TDMA, it is also necessary to send control information on the traffic channels. This is because the mobile has only one receiver; it cannot count on receiving the broadcast channel during a call, because both channels may use the same time slot. Also as with TDMA, there are two control channels associated with the traffic channel. The *Slow Associated Control Channel* (*SACCH*) uses one of every 26 bursts on the voice channel. It is used to inform the base of power measurements made by the mobile of signal strength in adjacent cells. The *Fast Associated Control Channel* (*FACCH*) "steals" bits from the voice signal and is used for urgent messages from the base, such as handoff instructions.

Voice Transmission

Each voice transmission is coded at 13 kb/s. A linear predictive coder, which models the way sounds are produced in the human throat, mouth, and tongue, is used. Such coding allows the bit rate to be greatly reduced compared with straightforward PCM. In the future, it is planned to use more advanced voice coders (vocoders) to allow the bit rate to be reduced to 6.5 kb/s, doubling the capacity of the GSM system. Note the similarity with full- and half-rate TDMA, which code voice at 8 and 4 kb/s, respectively. See Chapter 3 for a discussion of vocoders.

The bits from the vocoder are grouped according to their importance, with the most significant bits getting the most error correction and the least significant bits getting none. Then the data is spread over several frames by interleaving it so that the loss of a frame due to noise or interference will have a less serious effect.

Each voice transmission is allocated one time slot per frame. A frame lasts 4.615 ms so each time slot is approximately 577 µs in duration. To allow time for transmitters to turn on and off, the useful portion of the time slot is 542.8 µs, which allows time for 147 bits. This gives a raw data rate of 31.8 kb/s per voice channel. The timing for mobile transmissions is critical so that each arrives at the base station in the correct time slot. Since the propagation time varies with the distance of the mobile from the base, the mobile has to advance its timing as it gets farther from the base. It does this by monitoring a timing signal sent from the base on a broadcast channel. Although the time slots used by a mobile for receiving and transmitting have the same number, they are actually separated in time by a period equal to three time slots (uplink lags downlink). This means that the mobile unit, unlike analog systems, does not have to receive and transmit at the same time. When neither receiving nor transmitting on the voice channels, the mobile monitors the broadcast channels of adjacent cell sites and reports their signal strengths to the network to help it determine when to order a handoff. See Figure 11.4 for the structure of a voice channel.

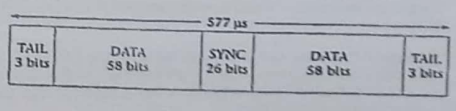


FIGURE 11.4
GSM voice channel

(F)requency (Hoppin)g in GSM When multipath fading is a problem, the GSM system allows for frequency hopping, a type of spread-spectrum communication that was discussed in Chapter 4. This can often solve the problem, since multipath fading is highly frequency-dependent. All GSM mobiles are capable of frequency hopping, but only those cells that are located in areas of severe fading are designated as hopping cells. The system can hop only among the frequencies that are assigned to the cell, so there wil' be only a few hopping possibilities (on the order of three frequencies). Thus GSM is not really a true spread-spectrum system, but rather a TDM/FDM system with some spread-spectrum capability added on. This feature is unique to GSM; IS-136 TDMA has nothing like it.

(Su)bscriber (ID) Module The subscriber ID module (SIM) is unique to the GSM system. It is a smart card with eight kilobytes of memory that can be plugged into any GSM phone. SIMs come in two sizes: one is the size of a credit card, the other is about postage-stamp size. The SIM contains all subscriber information including telephone number (called the International Mobile Subscriber Identification (IMSI) in GSM), a list of networks and countries where the user is entitled to service, and other user-specified information such as memories and speed dial numbers. The card allows a subscriber to use any GSM phone, anywhere. For instance, since the PCS frequencies are different in Europe and North America, there is no point in a North American traveling in Europe with a PCS phone. If a traveler takes the SIM, however, it will work with any phone rented or purchased in Europe, as long as the subscriber has first contacted his or her North American GSM service provider to arrange for authorization.

The SIM also offers some protection against fraudulent use. A GSM phone is useless without a SIM; if the user removes the card when leaving the phone in a car, for example, the phone cannot be used unless the thief has a valid SIM. Unfortunately, the cards can be stolen too. The SIM can be set up to require the user to enter a *personal identification number* (*PIN*) whenever the phone is turned on to provide some security in case the card is lost or stolen.

Once a subscriber has a SIM, buying a new GSM phone is easy. No setup or programming by the dealer is required. Similarly, a user can have a permanently-installed mobile phone and a portable with the same phone number, provided that only one is used at a time. However, the TDMA system also makes purchasing a new phone fairly easy; it allows a phone to be activated and programmed over the air, using the control channel.

Privacy (and S)ecurity The GSM SIM just discussed is only a part of the effort that has gone into securing this system. Both the data used in authorizing calls, such as the subscriber's identifying numbers, and the digitized voice signal itself, are

usually encrypted. (It is possible to weaken or turn off the voice encryption, if a government requires it.) The security in GSM is better than in IS-136 and much better than in analog AMPS.

11.5 IS-95 CDMA PCS

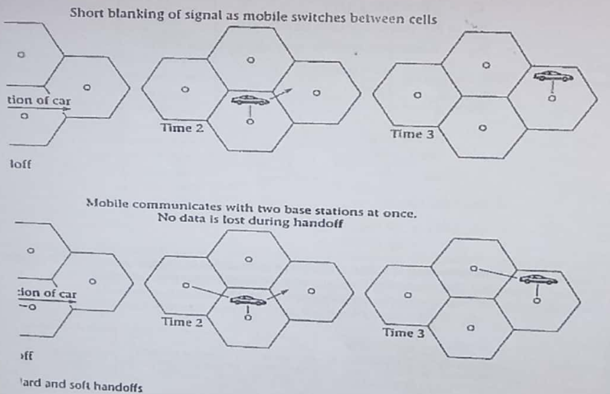
This United States-designed system has an air interface that is radically different from either of the others, though its control and messaging structure is quite similar to GSM. CDMA is used to a limited extent on the 800-MHz band, but is much more common in the 1900-MHz PCS band. It uses code-division multiple access by means of direct-sequence spread-spectrum modulation. See Chapter 4 for an introduction to spread-spectrum radio and code-division multiple access (CDMA).

CDMA Frequency Use One CDMA RF channel has a bandwidth of 1.25 MHz, using a single carrier modulated by a 1.2288 Mb/s bitstream using QPSK. CDMA allows the use of all frequencies in all cells (not one-seventh or one-twelfth of the frequencies in each cell, as required by other systems). This gives a considerable increase in system capacity. Because of the spread-spectrum system, co-channel interference simply increases the background noise level, and a considerable amount of such interference can be tolerated. As with the other personal communication systems, base and mobile stations transmit on separate channels separated by 80 MHz.

Frequency diversity is inherent in any spread-spectrum system. This is especially beneficial in a mobile environment subject to multipath propagation. The GSM system discussed earlier can use a limited amount of frequency diversity by hopping among several (typically three) discrete channels. The CDMA system, on the other hand, uses the full 1.25-MHz bandwidth for all voice channels on a given RF channel. If a small portion of this spectrum suffers a deep fade due to reflections, the only effect will be a slight increase in the error rate, which should be compensated for by the error correction built into the coding of the voice and control signals.

Space diversity is also built into a spread-spectrum system. Other cellular systems and PCS typically employ two receiving antennas per cell or sector at the base station to provide some space diversity, but they use only one antenna at the mobile location. Multiple receiving antennas are also used with CDMA; but since all frequencies are used in all cells, it is possible to receive the mobile at two or more base stations. Similarly, a mobile can receive signals from more than one base station. Each can make a decision about the strongest signal and can, in fact, combine signals to obtain an even stronger

one. Since there is no need for the mobile to change frequency on handoff, the CDMA system can use a *soft handoff*, in which a mobile communicates with two or more cells at the same time, rather than having to switch abruptly from one to another. This gives the ultimate in space diversity, with receiving antennas up to several kilometers apart. See Figure 11.5 for a comparison of soft and hard handoffs.



(... H)ard and soft handoffs

(Cha)nnels Each RF channel at a base station supports up to 64 orthogonal CDMA channels, using direct-sequence spread-spectrum, as follows:

- 1 pilot channel, which carries the phase reference for the other channels
- 1 sync channel, which carries accurate timing information (synchronized to the GPS satellite system) that allows mobiles to decode the other channels

- 7 paging channels, equivalent to the control and paging channels i(n) TDMA and GSM
- 55 traffic channels

CDMA thus uses a bandwidth of 1.25 MHz for 55 voice channels, whic(h) works out to about 22.7 kHz per channel. This is similar to GSM and, at firs(t) glance, not as efficient as TDMA. However, the fact that all channels can be used in all sectors of all cells makes CDMA more efficient in terms of spectrum than any of the other systems. Since CDMA degrades gracefully with increasing traffic, it is difficult to arrive at a definite maximum for its capacity. Proponents of CDMA claim spectrum efficiencies ten to twenty times as great as for GSM; those using other systems dispute this and put the gain nearer two. Once there is a large body of data from all the PCS schemes, it will be easier to get at the truth.

Along with the other personal communication systems discussed in this chapter, the CDMA system also uses FDMA. Each PCS carrier has a spectrum allotment of either 5 MHz or 15 MHz in each direction (refer back to Table 11.1), so a cell site can have more than one RF channel.

Forward Channel The forward and reverse channels are quite different in the CDMA system. Let us look at the forward channel first. We already know that sync, paging, and speech channels are combined on the same physical RF channel using CDMA. We learned in Chapter 4 that the direct-sequence form of CDMA is created by combining each of the baseband signals to be multiplexed with a *pseudo-random noise* (*PN*) sequence at a much higher data rate. Each of the signals to be multiplexed should use a different PN sequence. In fact, it can be shown that if the various sequences are mathematically *orthogonal*, the individual baseband signals can, at least in theory, be recovered exactly without any mutual interference. The math involved in proving this is beyond the scope of this text, but we should note that the number of possible orthogonal sequences is limited and depends on the length of the sequence. If the PN sequences are not orthogonal, CDMA is still possible, but there will be some mutual interference between the signals. The effect of this will be an increased noise level for all signals; eventually, as the number of non-orthogonal signals increases, the signal-to-noise ratio becomes too low and the bit-error rate too high for proper operation of the system. However, at no time do we hear audible crosstalk, as we do with two analog signals on the same frequency.

From the foregoing it would seem that using orthogonal PN sequences for CDMA is highly desirable, and this is what is done at the base station. A class of PN sequence called a Walsh code is used. The base station uses

Recall from Chapter 4 that the processing gain can be found as follows:

$$G_p = \frac{B_{RF}}{B_{BB}} \tag{11.1}$$

where

G_p = processing gain
B_{RF} = RF (transmitted) bandwidth
B_{BB} = baseband (before spreading) bandwidth

Here,

$$G_p = \frac{B_{RF}}{B_{BB}}$$
$$= \frac{1.2288 \times 10^6}{19.2 \times 10^3}$$
$$= 64$$

In decibels, this is

$$G_p \text{ (dB)} = 10 \log 64$$
$$= 18.06 \text{ dB}$$

If we consider that the error-correction codes are a form of spreading as well, since they increase the data rate, the total spreading becomes

$$G_p = \frac{B_{RF}}{B_{BB}}$$
$$= \frac{1.2288 \times 10^6}{9.6 \times 10^3}$$
$$= 128$$
$$= 21.1 \text{ dB}$$

A signal-to-noise ratio of about 7 dB is required at the receiver output for a reasonable bit-error rate. This means that the signal-to-noise ratio in the RF channel can be about −14 dB for satisfactory operation; that is, the signal power can be 14 dB less than the noise power. This takes a little getting used to, but is typical of spread-spectrum systems.

The 64 orthogonal channels are transmitted on one RF carrier by summing them, as in Figure 11.7, and using QPSK to modulate them on a single carrier.

Dr.G.Arumugam
M.Sc., B.Ed., M.Phil., Ph.D.
Assistant Professor of Physics
PG & Research Department of Physics
Arignar Anna Arts & Science College
Thenkarai - 7 Cell : 9751955888

GURE 11.7 ultiplexing of DMA channels

Channel The mobile units cannot use truly orthogonal channels because they lack a phase-coherent pilot channel. Each mobile would need its own pilot channel, which would use too much bandwidth. Therefore, they use a more robust error-control system. It outputs data at three times the input data rate. Follow Figure 11.8 to see what happens to the signal.

E 11.8 CDMA reverse voice channel

64 orthogonal Walsh codes; each repeats after 64 bits. This allows for 64 independent logical channels per RF channel, as mentioned earlier. Walsh code 0 is used for the pilot channel to keep mobile receivers phase-aligned with the base station. This is a requirement for coherent demodulation, which is the only way to avoid interference among channels using the same carrier frequency.

In addition to the Walsh codes, two other codes are in use at a CDMA base station: a *short code* for synchronizing, and a *long code*, which is used for encryption of both voice and control-system data and is not used for spreading.

Figure 11.6 shows how the spreading works for one voice signal. The vocoder produces a voice signal with a maximum bit rate of 9.6 kb/s. Error-correction bits are added, and the samples are interleaved over time, just as they are for TDMA and GSM. This process increases the bit rate for one voice channel to 19.2 kb/s. Next the signal is exclusive-or'd with the long code. This code repeats only after $2^{42} - 1$ bits and is used for encryption, not spreading. The signal remains at 19.2 kb/s after this process.

.6 CDMA forward voice channel

Spreading occurs when the 19.2 kb/s baseband data stream is multiplied by one of the 64 Walsh codes. Each of the Walsh codes has a bit rate of 1.2288 Mb/s. The multiplication works as follows: when the data bit is zero, the Walsh code bits are transmitted unchanged; when the data bit is one, all the Walsh code bits are inverted. The output bit stream is at 1.2288 Mb/s, which is 64 times as great a data rate as for the baseband signal at 19.2 kb/s. Therefore, the transmitted signal bandwidth is 64 times as great as it would be for the original signal, assuming the same modulation scheme for each.

$$f_s = 3 \times 9.6 \text{ kb/s}$$
$$= 28.8 \text{ kb/s}$$

The 28.8 kb/s signal is combined with one of the 64 Walsh codes and a long code to reach the full data rate of 1.2288 Mb/s. However, the purpose of each of these codes is different on the reverse channel. Here the long code is used to distinguish one mobile from another, as each uses a unique (though not necessarily orthogonal) long code. The Walsh codes are used to help the base station decode the message in the presence of interference. Each block of six information bits (64 different possible combinations) is associated with one of the 64 Walsh codes, and that code, rather than the actual data bits, is transmitted. Since each Walsh code is 64 bits long, this in itself does some spreading of the signal: the bit rate is increased by a factor of 64/6. The Walsh code mapping thus increases the data rate as follows:

$$f_s = 28.8 \text{ kb/s} \times \frac{64}{6}$$
$$= 307.2 \text{ kb/s}$$

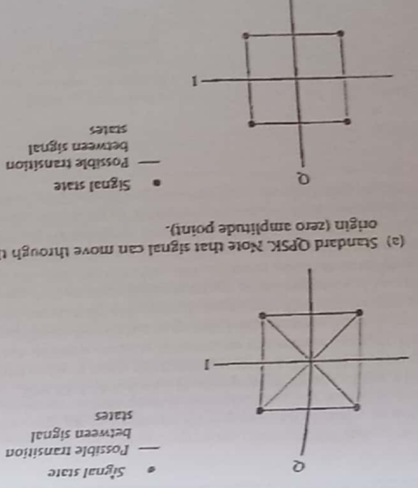
The long code is now multiplied with the data stream to produce a reverse channel bit rate of 1.2288 Mb/s, the same as for the forward channel. Each mobile transmits at the same rate to produce the spread-spectrum signal received at the base.

The modulation scheme is also slightly different on the reverse and forward channels. Both use a form of quadrature phase-shift keying (QPSK). The base station uses conventional QPSK. With this system the transmitter power has to go through zero during certain transitions. See Figure 11.9(a) on page 436.

The mobiles delay the quadrature signal by one-half a bit period to produce offset QPSK, which has the advantage that the transmitter power never goes through zero, though the amplitude does change somewhat. Linear amplifiers are still required in the mobile transmitter, but the linearity requirements are not as strict for offset QPSK as they are for conventional QPSK. See Figure 11.9(b).

Offset QPSK would have no advantage for the base station because a single transmitter is used for all the multiplexed signals. The summing of a large number of signals would result in a signal that still went through the zero-amplitude point at the origin.

Voice Coding CDMA uses a variable rate vocoder. Four different bit rates are possible: 9600, 4800, 2400 and 1200 b/s. The full rate of 9600 b/s is used when the user is talking. During pauses, the bit rate is reduced to 1200 b/s. The other two rates are also in the specifications but are seldom used.



'E 11.9
ord and offset

(a) Standard QPSK. Note that signal can move through the origin (zero amplitude point).

(b) Offset QPSK. ½-bit delay in Q channel means that signal moves along I, then along Q axis. It never goes to zero amplitude.

For many years it has been realized that each user typically talks less than fifty percent of the time during a conversation. Theoretically, the bandwidth allocated to that customer can be reassigned during the pauses while the other person is talking. However, until CDMA PCS came along there were at least two problems with this. The first was that it does not sound natural to have the voice channel go dead when someone stops talking. It sounds as if the phone has been disconnected. The reason is that there is always background noise, even in a quiet room. The CDMA system transmits this noise, but codes it at a lower rate (1200 b/s) because it is not important that it be rendered accurately.

The other problem, with either FDMA or TDMA, was finding a use for the vacated channel or time slot. The slot is usually available for only a few seconds or less, and the amount of time is not known in advance. In CDMA the

reduced amount of information to be sent can be translated directly into reduced interference to other transmissions on the same frequency, which automatically increases the capacity of the system. The way in which this is done is different for the forward and reverse channels.

On the forward channel, data bits are repeated when the coder is running at less than the maximum rate of 9.6 kb/s. For instance, if the coder operates at 1.2 kb/s, as it does during pauses in speech, each block of data is transmitted eight times. Because the error rate at the receiver depends on the energy per received data bit, the power in the transmit channel can be reduced under these circumstances.

The mobile transmitter handles this situation differently. Rather than reduce power, it simply transmits only one-eighth of the time, reducing interference and increasing battery life.

Mobile Power Control Controlling the power of the mobile stations is even more important with CDMA than with other schemes. The power received at the base station from all mobiles must be equal, within 1 dB, for the system to work properly. The power level is first set approximately by the mobile, and then tightly controlled by the base. When first turned on, the mobile measures the received power from the base, assumes that the losses on the forward and reverse channels are equal, and sets the transmitter power accordingly. This is called *open-loop* power setting. The mobile usually works with the equation:

$$P_T = -76 \text{ dB} - P_R \tag{11.2}$$

where

P_T = transmitted power in dBm

P_R = received power in dBm

EXAMPLE 11.1

A CDMA mobile measures the signal strength from the base as −100 dBm. What should the mobile transmitter power be set to as a first approximation?

SOLUTION

From Equation (11.2):

$$\begin{aligned} P_T &= -76 \text{ dB} - P_R \\ &= -76 \text{ dB} - (-100 \text{ dBm}) \\ &= 24 \text{ dBm} \\ &= 250 \text{ mW} \end{aligned}$$

Dr.G.Arumugam
M.Sc., B.Ed., M.Phil., Ph.D.,
Assistant Professor of Physics
PG & Research Department of Physics
Arul Valliammal Arts & Science College
Thanjavur - 7 Cell: 9751953588

The mobile begins by transmitting at the power determined by Equation (11.2) and increases power if it does not receive acknowledgement from the base. This could happen if a substantial amount of the received power at the mobile is actually from adjacent cells. We should also remember, that just as for the other systems, the forward and reverse channels are at different frequencies, so the amount of fading may be different.

Once a call is established, the open-loop power setting is adjusted in 1 dB increments every 1.25 ms by commands from the base station, to keep the received power from all mobiles at the same level. This *closed-loop* power control is required; for CDMA to work properly, all the received signals must have equal power. Otherwise the system suffers from the near/far effect, in which the weaker signals are lost in the noise created by the stronger ones. Careful power control has the added benefit of reducing battery drain in the portable unit, as the transmitted power is always the minimum required for proper operation of the system.

Rake Receivers and Handoffs One of the advantages of the CDMA system is that multipath interference can be reduced by combining direct and reflected signals in the receiver. The receivers used are called rake receivers; the reason can be seen in the diagram in Figure 11.10, which somewhat resembles a rake with several teeth for the reception of signals having different amounts of delay.

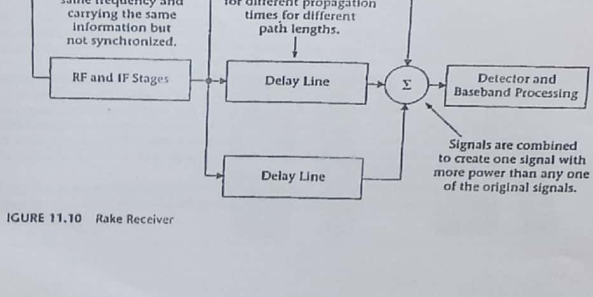


FIGURE 11.10 Rake Receiver

The mobile unit can combine three RF signals, delaying two of them to match the third. One of these signals can be assumed to be the base station in the current cell. The other two may be reflections or neighboring base stations. The base-station receiver can combine four signals: the direct signal from the mobile and three reflections.

In addition, two base stations may receive a signal from the same mobile. The base stations each send their signals to the MSC, which uses the higher-quality signal. Decisions about quality are made on a frame-by-frame basis every 20 ms. It is possible to have two base stations communicating with the same mobile indefinitely in what is referred to as a soft handoff. This avoids the dropping of calls that sometimes occurs when a handoff is unsuccessful in other systems, perhaps because there are no available channels in the new cell. The disadvantage is a considerably increased load on the base stations and the switching network.

CDMA Security CDMA offers excellent security. A casual listener with a scanner will hear only noise on a CDMA channel. In order to decode a call it is not only necessary to have a spread-spectrum receiver, but also to have the correct despreading code. Since this so-called "long code" is $2^{42} - 1$ bits long before it repeats and is newly generated for each call, the chances of eavesdropping are small. Identification is done using private-key encryption, as for GSM.

11.6 Comparison of Modulation Schemes

All of the North American PCS have advantages. TDMA is compatible with much existing North American cell-site equipment. GSM has a long history and a large installed base, which tends to lead to lower prices. It also has more advanced features than IS-136. CDMA is the most sophisticated technically, offers the best security, and makes the best use of system bandwidth, at least in theory. All three are in wide use in the United States and Canada. Table 11.3 on page 440 compares the three systems under several headings.

Compatibility Issues: Multi-Mode Phones From Table 11.3 there appears to be an obvious compatibility problem in PCS. The three systems have only their frequency range in common; none of the systems is compatible with either of the others. Consequently, roaming in the PCS band is possible only among providers that use the same system. At this writing, one or more of these systems is available in most populous areas, but not all areas have all three. Eventually the problem may disappear, as PCS coverage becomes ubiquitous with all three systems, but in the meantime, there is substantially less roaming capability with PCS than

Comparison of North American PCS

perty	IS-136 TDMA	GSM	IS-95 CDMAone
l Width	30 kHz	200 kHz	1.25 MHz
Channels nnel	3	8	64, including 1 sync and 7 control channels
ng Type	TDMA	TDMA plus limited frequency-hopping	Direct-sequence spread-spectrum
ing Rate	8 kb/s full rate 4 kb/s half rate	13 kb/s full rate 6.5 kb/s half rate	Variable 9600 b/s max. 1200 b/s min.
r RF Channel	48.6 kb/s	270.833 kb/s	1.2288 Mb/s
nannel	2 time slots of 6 in one 30-kHz RF channel	1 time slot of 8 in one 200-kHz RF channel	7 of 64 orthogonal codes in one 1.25-MHz RF channel

with 800-MHz analog AMPS, which still has by far the widest distribution of any system in North America.

There is an obvious, though rather unwieldy, solution to the compatibility problem. This is to manufacture dual-band, dual-mode phones, which work with analog, 800-MHz AMPS as well as with one of the 1900-MHz personal communication systems. Dual-mode phones are currently available for all three of the PCS. Those PCS providers who do not also have an 800-MHz license often form alliances with a cellular provider to allow seamless roaming with only one monthly bill. Figure 11.11 shows examples of dual-mode phones.

FIGURE 11.11
Dual-mode phones
(Courtesy of Nokia, Inc.)

11.7 Data Communication with PCS

When we studied the TDMA cellular system, we observed that data communication can actually be more complex with a digital than with an analog system. This is because vocoders will not work properly with modems, so that the classic technique of connecting a modem to an analog voice channel does not work. We saw that at 800 MHz it is common for a digital phone to revert to analog mode for circuit-switched data communication and to use the CDPD system for packet-switched data.

The above techniques are still possible with a dual-mode PCS phone, but each of the three personal communication systems has developed its own techniques for data communication. This can be expected to become more important as a new generation of "smart phones" incorporating larger displays, (some even including web browsers) and portable computers incorporating RF communication modules are introduced.

At present the most popular use for PCS data seems to be short paging-type messages, followed by electronic mail. Worldwide web access is gaining in importance, but is currently limited by slow connection speeds and the limited graphics capability of PCS phone displays. Let us see how data transmission is handled with each of the three PCS.

TDMA Data Communication

The TDMA PCS standard allows for short messages and packet-switched data to be sent on the digital control channels (DCCH) or the digital traffic channels (DTC). Circuit-switched data is possible on the digital traffic channels.

The digital control and traffic channels support two main types of packet-switched data communication. A format called *cellular messaging teleservice (CMT)* is employed for a short messaging service (SMS). This allows for brief paging-type messages and short e-mail messages (up to 239 characters), which can be read on the phone's display and entered using the keypad. For longer messages and extended services like web browsing, the *Generic UDP Transport Service (GUTS)* protocol is used. The acronym-within-an-acronym *UDP* stands for *User Datagram Protocol*.

Both of these services require extra equipment in the PCS network to translate between wireline protocols and those used with the radio link. With GUTS, the user connects to a network server that relays messages to and from the internet. The CMT system also requires the servers in the PCS network to assemble messages and interconnect with other services such as the user's e-mail service. See Figure 11.12 on page 442 for an illustration of packet-switched PCS data.

Circuit-switched data communication is accomplished on the digital traffic channels. The vocoder is bypassed and data is coded and sent directly

APTER 11

Comparison of North American PCS

perty	IS-136 TDMA	GSM	IS-95 CDMAone
l Width	30 kHz	200 kHz	1.25 MHz
Channels nnel	3	8	64, including 1 sync and 7 control channels
ng Type	TDMA	TDMA plus limited frequency-hopping	Direct-sequence spread-spectrum
ing Rate	8 kb/s full rate 4 kb/s half rate	13 kb/s full rate 6.5 kb/s half rate	Variable 9600 b/s max. 1200 b/s min.
r RF Channel	48.6 kb/s	270.833 kb/s	1.2288 Mb/s
nannel	2 time slots of 6 in one 30-kHz RF channel	1 time slot of 8 in one 200-kHz RF channel	7 of 64 orthogonal codes in one 1.25-MHz RF channel

with 800-MHz analog AMPS, which still has by far the widest distribution of any system in North America.

There is an obvious, though rather unwieldy, solution to the compatibility problem. This is to manufacture dual-band, dual-mode phones, which work with analog, 800-MHz AMPS as well as with one of the 1900-MHz personal communication systems. Dual-mode phones are currently available for all three of the PCS. Those PCS providers who do not also have an 800-MHz license often form alliances with a cellular provider to allow seamless roaming with only one monthly bill. Figure 11.11 shows examples of dual-mode phones.

FIGURE 11.11
Dual-mode phones
(Courtesy of Nokia, Inc.)

11.7 Data Communication with PCS

When we studied the TDMA cellular system, we observed that data communication can actually be more complex with a digital than with an analog system. This is because vocoders will not work properly with modems, so that the classic technique of connecting a modem to an analog voice channel does not work. We saw that at 800 MHz it is common for a digital phone to revert to analog mode for circuit-switched data communication and to use the CDPD system for packet-switched data.

The above techniques are still possible with a dual-mode PCS phone, but each of the three personal communication systems has developed its own techniques for data communication. This can be expected to become more important as a new generation of "smart phones" incorporating larger displays, (some even including web browsers) and portable computers incorporating RF communication modules are introduced.

At present the most popular use for PCS data seems to be short paging-type messages, followed by electronic mail. Worldwide web access is gaining in importance, but is currently limited by slow connection speeds and the limited graphics capability of PCS phone displays. Let us see how data transmission is handled with each of the three PCS.

TDMA Data Communication

The TDMA PCS standard allows for short messages and packet-switched data to be sent on the digital control channels (DCCH) or the digital traffic channels (DTC). Circuit-switched data is possible on the digital traffic channels.

The digital control and traffic channels support two main types of packet-switched data communication. A format called *cellular messaging teleservice (CMT)* is employed for a short messaging service (SMS). This allows for brief paging-type messages and short e-mail messages (up to 239 characters), which can be read on the phone's display and entered using the keypad. For longer messages and extended services like web browsing, the *Generic UDP Transport Service (GUTS)* protocol is used. The acronym-within-an-acronym *UDP* stands for *User Datagram Protocol*.

Both of these services require extra equipment in the PCS network to translate between wireline protocols and those used with the radio link. With GUTS, the user connects to a network server that relays messages to and from the internet. The CMT system also requires the servers in the PCS network to assemble messages and interconnect with other services such as the user's e-mail service. See Figure 11.12 on page 442 for an illustration of packet-switched PCS data.

Circuit-switched data communication is accomplished on the digital traffic channels. The vocoder is bypassed and data is coded and sent directly

reless Web Browsing

Any of the PCS schemes just described can be used to access content on the World Wide Web. There are three major problems with all of them however: the data rate is low, even in comparison with ordinary telephone modems; the on-board computing power is low compared with a personal computer; and the handheld devices have very small, low-resolution displays. Many of these displays are not suitable for graphics. A typical web page would take a long time to load and when loaded would be almost, if not completely, unusable.

Third-generation wireless systems, which are described in Chapter 14, will help to solve the first problem, and perhaps make a start on the second. The third is more intractable: large displays and pocket-sized devices are simply not compatible. Therefore, even with third-generation systems, there will be a need for a means to display web pages on the small screens of PCS devices.

Until recently there have been many proprietary standards for displaying web content on wireless devices. Each worked only with a small number of specially created sites. Many of the major wireless manufacturers, including Ericsson, Nokia, and Motorola, have now combined to create a set of de facto standards for creating this content, known as the *Wireless Application Protocol* (*WAP*). The idea is to include a small program called a microbrowser in the wireless device, with most of the required computing done on network servers. These servers have access to specially modified pages on web sites and can also attempt to translate conventional sites so that they can be used by wireless devices. The pages have minimal graphics and condensed text so that they can be used with portable devices.

WAP is compatible with all of the current (second generation) systems and will be compatible with all third-generation systems as well. As more sites begin to provide pages compatible with WAP, the web should become quite accessible to portable wireless devices.

11.8 Testing Cellular Systems and PCS

We saw in Chapter 7 that all calculations of signal strength in a mobile environment are necessarily approximate, as there are too many variables for even a computer analysis to be accurate. Nonetheless, these predictions are generally accurate enough to locate cell sites and repeaters. Once these have been built, it is necessary to go into the field with a receiver to verify that the signal strength is satisfactory. The transmitter used for preliminary tests is often a portable model that puts out a carrier only. This simplifies the measurements, particularly when the system will eventually be CDMA with all channels active in all cells. Figure 11.15 shows a typical portable transmitter and receiver suitable for such testing.

(a) Transmitter (b) Receiver

FIGURE 11.15 Test transmitter and receiver
(Courtesy of Berkeley Vantronics Systems)

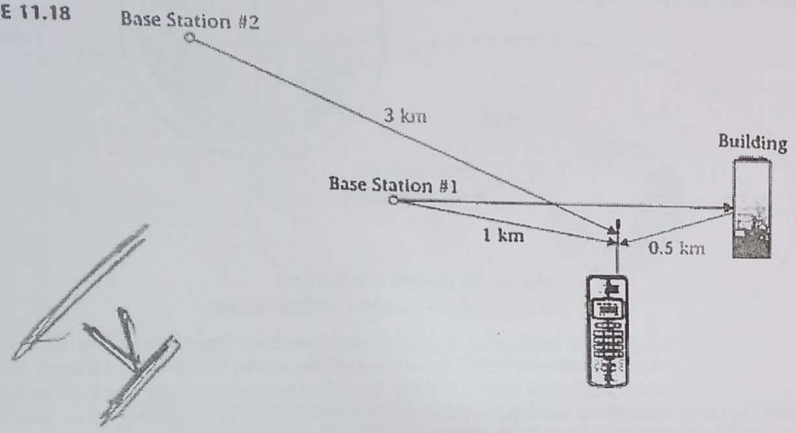
In addition to the RF equipment, it is desirable to have the test vehicle equipped with a global positioning system (GPS) receiver and a computer. It is then possible to plot signal strengths on a map and then to draw signal contour lines. In this way, any problem areas can be identified.

Once the system is in use, signal measurements can still be taken, but the process is more complex. This is especially true for CDMA, because all cells transmit on all frequencies with low power spread over a wide bandwidth. A conventional receiver would not be able to distinguish signals from noise, or one cell from another. A specialized receiver capable of extracting the pilot channel from a CDMA transmission is required. Each cell transmits a pilot with a slightly different delay added to the short code, and thus it can be identified if an accurate timing reference is available. Again, GPS is used by the CDMA system for this function. Figure 11.16 shows a specialized instrument designed for this purpose.

Testing of individual phones presents a problem, especially since it must be done quickly and economically; otherwise it will be less costly to

9. A rake receiver in a mobile receives a direct signal from a base 1 km away and a reflected signal from a building 0.5 km behind the mobile. It also receives a signal from another base station 3 km away. See Figure 11.18 for the situation. Calculate the amount of time delay each "finger" of the receiver needs to apply.

IGURE 11.18



12 Satellite-Based Wireless Systems

Objectives

After studying this chapter, you should be able to:

- Calculate velocity and period for artificial satellites in circular orbits.
- Distinguish between low-, medium-, and geostationary earth orbits, and explain the advantages and disadvantages of each for communication.
- Explain how elliptical orbits can be used in communication satellites.
- Explain the need for satellite tracking, and describe how it is done.
- Describe the types of satellite transponders used for wireless communication.
- Perform signal, noise, and signal-to-noise ratio calculations with satellite links.
- Describe several current and projected projects in the field of wireless communication by satellite, and discuss the merits of each.

12.1 Introduction

Cellular and personal communication systems work very well in populous areas, but they do require extensive infrastructure. Such systems are impractical for use in remote areas or at sea. Until quite recently the only means of portable communication in such areas was high-frequency (HF) radio. Over-the-horizon propagation is possible at frequencies in the range from about 3–30 MHz, due to reflection from the ionosphere; but HF propagation is unreliable, noisy, and suitable only for low-bandwidth applications. The antenna sizes required for wavelengths in the 10 to 100 meter range are awkward for mobile and all but impossible for handheld use.

A better alternative is to use satellites to relay signals from mobile or portable users to the public switched telephone network. Problems of cost, time delay, and equipment size exist, however, and will be addressed as we proceed. It seems unlikely that satellite wireless communication systems will ever be as popular as terrestrial systems, but they will fill an important role where terrestrial systems are not practical.

12.2 Satellite Orbits

We looked very briefly at satellite orbits in Chapter 7. In theory, a satellite can orbit at any altitude, but air resistance makes satellites impractical below about 300 km. A satellite can have any elliptical orbit but most (not all) of those used for communication have orbits that are at least approximately circular.

Satellite orbits can circle the equator (equatorial orbit); they can pass over both poles in a **polar orbit**, which has a 90° angle with respect to an equatorial orbit; or they can have any angle between these. Figure 12.1 shows some examples of possible circular orbits.

Orbital Calculations

Satellites are held in orbit by their momentum. Gravity continually bends a satellite's path toward the earth, but the satellite's momentum is sufficient to prevent its trajectory from reaching the earth. This phenomenon is commonly called centrifugal force, though technically there is no such thing.

Any satellite orbiting the earth must satisfy this equation:

$$v = \sqrt{\frac{4 \times 10^{14}}{d + 6400}} \qquad (12.1)$$

where

v = velocity in meters per second

d = distance above earth's surface in km

Several important things can be seen from Equation (12.1). First, the farther a satellite is from the surface of the earth, the slower it travels. Second, since a satellite that is farther from the earth obviously has farther to go to complete an orbit, and since it also travels more slowly than one closer to earth, then the orbital period of a distant satellite must be longer than the period of one closer to the earth. An example will illustrate this.

FIGURE 12.1
Circular orbits

EXAMPLE 12.1

Find the velocity and the orbital period of a satellite in a circular orbit

(a) 500 km above the earth's surface

(b) 36,000 km above the earth's surface

SOLUTION

(a) From Equation (12.1):

$$v = \sqrt{\frac{4 \times 10^{14}}{(d + 6400)}}$$

$$= \sqrt{\frac{4 \times 10^{14}}{(500 + 6400)}}$$

$$= 7.6 \text{ km/s}$$

al Orbits

Technically all satellite orbits are elliptical. A circle is a special case of ellipse where the maximum distance from the earth (apogee) is equal to the minimum distance (perigee).

When a satellite has an elliptical orbit in which the earth is at one focus, it spends more time in that part of the orbit that takes it farthest from earth. This can be seen from Kepler's second law, which states that such a satellite sweeps out equal areas in space in equal times. See Figure 12.3 for the idea.

FIGURE 12.3
Elliptical orbit

Such orbits can be used to advantage. The Russian Molniya series of communication satellites, for instance, uses highly elliptical polar orbits arranged so that the satellites spend about 11 hours of a 12-hour orbit over the northern hemisphere. This facilitates communication in regions near the North Pole. Geostationary satellites, which must be in a circular orbit above the equator, are near the horizon in these regions. The proposed Ellipso satellite system, described later, will also use elliptical orbits.

12.3 Use of Satellites for Communication

The traditional way to communicate with a satellite in non-geostationary orbit is to use a movable, directional antenna and point it at the satellite. The approximate *azimuth* (angle in the horizontal plane) and *elevation* (vertical angle measured from the horizon) can be calculated from orbital data, and fine position adjustments can be made for the strongest signal. The antenna must then follow, or track, the position of the satellite as it moves. At some point, of course, the satellite goes below the horizon and communication is lost.

There are two obvious problems with this for mobile and portable communication. The first is that rapidly changing the orientation of a highly directional antenna on a moving vehicle is difficult. This can be done on ships where changes in direction are relatively slow and there is room for a dish antenna, but it is very difficult on cars and just about impossible for a portable, hand-carried unit. Moving vehicles require antennas that either have a broad beam or are steered electronically using phased elements.

The disappearance of satellites over the horizon is another serious problem. If real-time voice communication is to be possible, at least one satellite must be visible at all times. Two solutions to the problem are explored below.

Using Satellites in Geostationary Orbit

There is an orbit in which a satellite appears to be stationary above a particular spot at the equator. This is very useful, because such a satellite never goes below the horizon and the antenna position with respect to the earth, once found, never needs to be changed. In fact this orbit is so convenient that until recently most communication satellites used it. It is still the method of choice for point-to-point service and broadcasting.

The geostationary orbit does leave something to be desired for portable and mobile use, however. Being able to point the antenna always in the same direction is not as useful on a car or with a handheld device as it is for a fixed installation. It is the orientation of the antenna with respect to the vehicle that must be able to change rapidly as the vehicle changes direction.

Another problem is the extreme path loss at the distance needed for a geostationary satellite. We have seen that if we require a geostationary orbit, but the altitude of 36,000 km, though greater than we would like, is set by the laws of physics. The large path loss usually means that a high-gain directional antenna is needed, and this is problematic for portable/mobile operation.

One of the main reasons for using satellites for wireless communication is to provide coverage in remote areas, which are typically out of reach of terrestrial cellular systems. Many of these areas are in high latitudes (northern Canada and Alaska, for example). Unfortunately as the earth curves toward the poles, the geostationary satellites are closer to the horizon. This makes it harder to achieve a direct line of sight to the satellite, which is free from obstructions such as buildings or hills. Thus the geostationary satellite, which must be above the equator, become less useful in extreme northern or southern latitudes. Another glance at Figure 12.2 should make this clear.

Yet another problem, though less important than those already mentioned, is propagation time. We saw in Chapter 7 that the propagation delay is about 0.25 s for a round trip to a geostationary satellite. This delay, while not fatal, is annoying in real-time conversations. It also causes delays in data

When real-time communication is required, the only way to address the second problem is to use a constellation containing more than one satellite. The closer the satellites are to earth, the more of them are needed in order to have at least one satellite visible from a given point at all times. This can make the system quite complex and expensive, even allowing for the fact that it is less expensive to put satellites in lower orbits than in geostationary orbit.

The Doppler effect requires careful receiver design for both satellite and ground stations, so that the receiver can lock onto an incoming signal and track its frequency changes.

Satellite orbits are usually divided into three ranges. **Low-earth-orbit (LEO) satellites** range from about 300 to 1500 km above the earth. **Medium-earth-orbit (MEO) satellites** are about 8,000 to 20,000 km in altitude. The gap between LEO and MEO orbits is there to avoid the lower of the two Van Allen radiation belts that surround the earth; this radiation can damage satellites. These radiation belts extend from 1,500 to 5,000, and from 13,000 to 20,000 km above the earth's surface. MEO satellites are typically near the lower end of the MEO range to avoid the upper Van Allen belt. Finally, of course, there is the geostationary earth orbit (GEO) already mentioned. It would, of course, be possible to orbit satellites at still greater distances from the earth, but since the GEO is already farther from Earth than we would like, there is no point in using more distant satellites for communicating between points on Earth. Figure 12.5 shows the LEO, MEO, and GEO ranges, along with the Van Allen belts.

FIGURE 12.5
LEO, MEO and GEO orbits (not to scale)

12.4 Satellites and Transponders

The satellite as a spacecraft, with its attendant guidance systems and po tioning jets, is outside the scope of this book. We are concerned, howev with the satellite as a radio repeater. We need to know what the satelli looks like from the ground.

The traditional way to build a communication satellite is to design it as frequency-shifting repeater or collection of repeaters (for some reason, a re peater on a satellite is called a **transponder**.) One satellite may have many transponders, each of which has a block diagram that looks like Figure 12.6. A range of frequencies is received from the ground via the *uplink*, amplified and shifted in frequency, and retransmitted on the *downlink*. No signal processing other than amplification and frequency shifting is done on the satellite. For obvious reasons, this type of transponder has what is known as the **bent-pipe configuration**.

FIGURE 12.6
Bent-pipe transponder

This transponder design is extremely versatile. A bent-pipe transponder can be used for anything from broadcast television using analog wideband FM to digital telephony using either time-division or frequency-division multiplexing or both. When used for one wideband FM signal, the satellite power amplifier can be operated in a saturated mode, much like a Class C amplifier, for greatest efficiency. If required to handle amplitude-varying signals, the amplifier can be "backed off" to a lower-power linear mode by remote control from the ground. Such a transponder is ready for just about any signal, or combination of signals, that will fit into its bandwidth.

It is also possible to design satellite transponders for specific applications. Some are designed to store digital information and retransmit it at a later time. Using this **store-and-forward** technique, data can be communicated using a low-earth orbit satellite that may not be visible to the transmitting and receiving stations at the same time. Also, satellites can be designed to communicate with each other. Such **crosslinks** can improve the efficiency of communication between earth stations, as indicated in Figure 12.7.

Professor of Physics
Department of Physics
Arts & Science College
Thanjavur - 7

The circumference of the orbit can be found from its radius, which is that of the earth (6400 km) plus the distance of the satellite from the earth. In this case the total distance is

$$r = 6400 \text{ km} + 500 \text{ km} = 6900 \text{ km}$$

and the circumference of the orbit is

$$\begin{aligned} C &= 2\pi r \\ &= 2\pi \times 6900 \text{ km} \\ &= 43.4 \text{ Mm} \end{aligned}$$

The period of the orbit can be found by dividing the circumference by the orbital velocity,

$$T = \frac{C}{v}$$

$$= \frac{43.4 \times 10^6 \text{ m}}{7.6 \times 10^3 \text{ m/s}}$$

$$= 5.71 \times 10^3 \text{ s}$$

$$= 1.6 \text{ hours}$$

(b) Again, start with Equation (12.1):

$$v = \sqrt{\frac{4 \times 10^{14}}{(d + 6400)}}$$

$$= \sqrt{\frac{4 \times 10^{14}}{(36{,}000 + 6400)}}$$

$$= 3.07 \text{ km/s}$$

Note that the speed is less than before. The new radius is

$$r = 6400 \text{ km} + 36{,}000 \text{ km}$$

$$= 42.4 \text{ Mm}$$

The circumference of the orbit is, as before,

$$\begin{aligned} C &= 2\pi r \\ &= 2\pi \times 42.4 \text{ Mm} \\ &= 266.4 \text{ Mm} \end{aligned}$$

Now we can find the period of the orbit.

$$T = \frac{C}{v}$$

$$= \frac{266.4 \times 10^6 \text{ m}}{3.07 \times 10^3 \text{ m/s}}$$

$$= 86.8 \times 10^3 \text{ s}$$

$$\equiv 24 \text{ hours}$$

Geostationary Orbit

The satellite in Example 12.1(b) has a particularly interesting orbit. It is, at least approximately, geosynchronous; that is, the satellite orbits the earth in the same amount of time it takes the earth to rotate once on its axis. If the orbit is circular and above the equator and the satellite travels in the same direction as the earth's rotation, it will also be geostationary; that is, it will appear stationary from the ground, because it rotates at the same rate and in the same direction as the earth. Though it is theoretically possible for a satellite to be geosynchronous without being geostationary by orbiting the earth in the direction opposite to the earth's rotation, in practice this is never done, and the two terms are used interchangeably. The orbit for a geostationary satellite has a radius more than five times as large as that of the earth. See Figure 12.2 for an illustration of the geostationary orbit.

FIGURE 12.2
Geostationary orbit
(not to scale)

FIGURE 12.4 Footprint for Anik E-1 satellite (Courtesy of Telesat Canada)

transmission whenever a protocol requires acknowledgment from the receiving station before the transmission can continue.

Stationary Satellite Beams and Footprints

A geostationary satellite can be "seen" from almost half the earth's surface. Therefore, three such satellites should be sufficient to cover the entire earth, except for the polar regions, with some overlap. A satellite designed for such wide coverage has an antenna with a relatively large beamwidth called a hemispheric beam. Because of its wide beam, such an antenna necessarily has relatively low gain.

Many geostationary satellites are not intended to cover an entire hemisphere. They use much more directional antennas, producing spot beams to cover populated regions. Such antennas have higher gain and can deliver a stronger signal on earth for the same transmitter power. Similarly, when used for receiving, such antennas can achieve a better signal-to-noise ratio at the satellite for a given transmitter power at the ground. This is very important in portable and mobile communication where ground-station transmitter power and antenna gain are strictly limited.

It is certainly possible for a satellite to have a combination of hemispheric and spot beams in order to provide basic service over a wide area and the ability to use smaller earth stations in particular areas.

When designing receiving installations on the ground, it is the EIRP of a satellite and not simply its transmitter power that is important. EIRP depends on the gain of the transmitting antenna in a particular direction, as well as the transmitter power, so it is different at different points on earth. Satellite operators publish maps showing the footprint of each geostationary satellite on the earth. These show the effective EIRP (in dBW) for the satellite at each point on the earth where reception is possible. See Figure 12.4 for an example of a satellite footprint.

Low- and Medium-Earth Orbits

Geostationary satellites are usable for wireless communication, but as we have seen, there are reasons to prefer satellites with lower orbits. The two main problems with such satellites are (1) their position in space is not fixed with respect to a ground station and (2) the annoying tendency of such satellites to disappear below the horizon. (Another smaller problem is the Doppler effect, which causes frequencies to change. Transmitted frequencies are shifted higher as the satellite approaches a point on the ground and lower as the satellite recedes.)

The first problem is less important than it might seem. Shorter range results in much less propagation loss and removes the requirement for highly directional antennas. This makes antenna tracking less critical. In any case, if the antenna is mounted on a moving vehicle or person, the direction to the satellite is constantly changing, even with geostationary satellites.

where

T_{eq} = equivalent noise temperature in kelvins
NF = noise figure as a ratio (not in dB)

EXAMPLE 12.2

A receiver has an equivalent noise figure of 2 dB. Calculate its equivalent noise temperature.

SOLUTION

First convert the noise figure to a ratio.

$$NF = \text{antilog}\left(\frac{NF_{dB}}{10}\right)$$

$$= \text{antilog}\left(\frac{2}{10}\right)$$

$$= 1.58$$

Now use Equation (12.5) to find the equivalent noise temperature:

$$T_{eq} = 290(NF - 1)$$
$$= 290(1.58 - 1)$$
$$= 170 \text{ K}$$

To find the equivalent noise temperature of the receiving installation, we need to add the antenna noise temperature, as modified by the feedline, to the receiver equivalent noise temperature. That is,

$$T = T_{eq} + T_a \tag{12.6}$$

where

T = system noise temperature in K
T_{eq} = receiver equivalent noise temperature in K
T_a = antenna noise temperature in K

The antenna noise temperature results from thermal noise picked up from objects in the beam of the antenna. This depends on the angle of elevation of the antenna: when the antenna beam includes the ground, the noise level increases because of radiation from the ground itself. Luckily, this is seldom the case with ground stations used with geostationary-satellites, except in very high latitudes where the satellite is just above the horizon. This means that noise entering the antenna originates mainly from extraterrestrial sources (stars, for instance) and from the atmosphere. Occasionally the sun passes through the main lobe of the antenna pattern; the sun is a very powerful noise source and makes communication impossible for the few minutes it takes to pass through the antenna beam. Otherwise the sky noise temperature for an earth-station receiving antenna is quite low, typically 20 K or less. The situation is very different for less directional antennas such as those used with portable phones in LEO systems. The noise temperature of these antennas may be about the same as the ambient temperature because of noise picked up from the surroundings; signals are much stronger at these antennas as well.

Losses in the antenna system contribute to its noise temperature. The noise temperature of an antenna system, at the far end of a feedline, is given by

$$T_a = \frac{(L - 1)290 + T_{sky}}{L} \tag{12.7}$$

where

T_a = effective noise temperature of antenna and feedline, referenced to receiver antenna input, in kelvins
L = loss in feedline and antenna as a ratio of input to output power (not in decibels)
T_{sky} = effective sky temperature, in kelvins

EXAMPLE 12.3

An earth station for use with a geostationary satellite has a dish antenna which sees a sky temperature of 25 K. It is connected to the receiver with a feedline having 1 dB loss. The receiver equivalent noise temperature is 15 K. Calculate the noise temperature for the system.

SOLUTION

First convert the feedline loss to a power ratio:

$$L = \text{antilog}\left(\frac{L_{dB}}{10}\right)$$

$$= \text{antilog}\left(\frac{1}{10}\right)$$

$$= 1.26$$

The signal strength is determined by the factors noted in Chapter 7. The important parameters are the EIRP of the transmitting installation, the distance to the receiver, and the gain of the receiving antenna, less any losses between antenna and receiver. Free-space propagation can usually be assumed in satellite communication, though for frequencies much above 10 GHz, losses in the atmosphere should be taken into account.

Noise Temperature As discussed in Chapter 1, the noise level is determined by the bandwidth of the system and its equivalent noise temperature. To review,

$$P_N = kTB \tag{12.2}$$

where
P_N = noise power in watts
k = Boltzmann's constant, 1.38×10^{-23} joules/kelvin (J/K)
T = system noise temperature in kelvins (K)
B = noise power bandwidth in hertz

The system noise temperature is not necessarily the actual temperature at which it operates as measured with a thermometer. It depends on the contribution of the various electronic components, as well as any noise picked up by the antenna or contributed by the feedline from antenna to receiver.

The total noise temperature of a system can be found by adding together the noise temperatures of its various components, provided all are referenced to the same point. Usually the receiver input is used as a reference point. Receiver noise temperatures are specified at their input. Sometimes, *noise figure*, rather than noise temperature is given. The noise figure is a measure of how much an electronic system degrades the signal-to-noise ratio of a signal at its input; that is,

$$NF = \frac{(S/N)_i}{(S/N)_o} \tag{12.3}$$

where
$(S/N)_i$ = signal-to-noise ratio at the input
$(S/N)_o$ = signal-to-noise ratio at the output

The noise figure, expressed this way, is a dimensionless ratio. In practice, it is always specified in dB, where

$$NF_{dB} = 10 \log NF \tag{12.4}$$

It is simple to convert noise figure to equivalent noise temperature using the equation:

$$T_e = 290(NF - 1) \tag{12.5}$$

The calculation of T is a little more difficult. We have to convert the feedhorn loss into a ratio, as follows:

$$L = \text{antilog}\left[\frac{0.4}{10}\right] = 1.096$$

Substituting into Equation (12.7), we get

$$T_s = \frac{(L-1)290 + T_{ant}}{L}$$
$$= \frac{(1.096-1)290 + 15}{1.096}$$
$$= 39 \text{ K}$$

The receiver noise temperature is given with respect to the chosen reference point, so it can be used directly. Therefore,

$$G/T(\text{dB}) = G_A(\text{dB}) - 10 \log (T_r + T_{ex})$$
$$= 39.6 - 10 \log (39 + 40)$$
$$= 20.6 \text{ dB}$$

to-Noise sing G/T Once G/T has been found, it can be used to help calculate the carrier-to-noise ratio for a system. We can use the following equation:

$$\frac{C}{N}(\text{dB}) = EIRP(\text{dBW}) - L_f(\text{dB}) + G/T - k(\text{dBW}) \tag{12.9}$$

where
C/N(dB) = carrier-to-noise ratio in decibels
$EIRP$(dBW) = effective isotropic radiated power in dBW
L_f(dB) = free-space loss in decibels
G/T = figure of merit as given in Equation (12.8)
k(dBW) = Boltzmann's constant expressed in dBW (−228.6 dBW)

The free-space loss for this application is as found in Chapter 7, that is

$$L_f = 32.44 + 20 \log d + 20 \log f \tag{12.10}$$

where
L_f = free space loss in decibels
d = path length in km
f = frequency in MHz

EXAMPLE 12.6

The receiving installation whose G/T was found in Example 12.2 is used as a ground terminal to receive a signal from a satellite at a distance of 38,000 km. The satellite has a transmitter power of 50 watts and an antenna gain of 30 dBi. Assume losses between the satellite transmitter and its antenna are negligible. The frequency is 12 GHz. Calculate the carrier-to-noise ratio at the receiver.

SOLUTION

The earth station was found to have $G/T = 20.6$ dB. The satellite transmitter power, in dBW, is

$$P_T(\text{dBW}) = 10 \log 50 = 17 \text{ dBW}$$

The EIRP in dBW is just the transmitter power in dBW plus the antenna gain in dBi, less any feedline losses, which are negligible here. Here, we have

$$EIRP(\text{dBW}) = 17 \text{ dBW} + 30 \text{ dBi} = 47 \text{ dBW}$$

Next we find L_f(dB). Using Equation (12.10),

$$L_f = 32.44 + 20 \log d + 20 \log f$$
$$= 32.44 + 20 \log 38{,}000 + 20 \log 12{,}000$$
$$= 205.6 \text{ dB}$$

Now we can find C/N. From Equation (12.9),

$$C/N(\text{dB}) = EIRP(\text{dBW}) - L_f(\text{dB}) + G/T - k(\text{dBW})$$
$$= 47 \text{ dBW} - 205.6 \text{ dB} + 20.6 \text{ dB} + 228.6 \text{ dBW}$$
$$= 90.6 \text{ dB}$$

JRE 12.7 sllinks

It is possible to turn a satellite transponder into a switching center so that a signal from a transmitter on the ground can be relayed to one of a variety of ground- or satellite-based receivers as required. However, the gain in efficiency is counterbalanced by an increase in satellite cost and complexity. It should be remembered that it is much easier to build and maintain complex equipment on the ground than in space. When we look at practical examples of wireless satellite communication later in this chapter, we'll see examples of both philosophies. Figure 12.8 shows a satellite with beam switching.

JRE 12.8
n switching

12.5 Signal and Noise Calculations

Because of the very weak received signals, satellite systems require low-noise receivers and relatively high-gain antennas for both the satellite and the earth station, especially when geostationary satellites are used. This is a different situation from the terrestrial systems we studied earlier, where interference or distance to the horizon was more likely than thermal noise to limit the range. This is, therefore, a good place to consider in more detail the factors that determine signal-to-noise ratio.

Now use Equation (12.7) to find the antenna noise temperature:

$$T_a = \frac{(L - 1)290 - T_{ar}}{L}$$

$$= \frac{0.26 \times 290 - 25}{126}$$

$$= 80 \text{ K}$$

Now add this to the receiver equivalent noise temperature to find the system temperature, as in Equation (12.6).

$$T = T_a + T_r$$
$$= 15 + 80$$
$$= 95 \text{ K}$$

to-Noise Ratio Once the system temperature is known, it is easy to calculate the noise power in any given bandwidth.

MPLE 12.4

Calculate the noise power for the system in the previous example, if the bandwidth is 2 MHz.

SOLUTION

From Equation (12.3),

$$P_N = kTB$$
$$= 1.38 \times 10^{-23} \times 95 \times 2 \times 10^6$$
$$= 2.62 \text{ fW}$$
$$= -116 \text{ dBm}$$

Now, if we know the signal strength, we can easily find the signal-to-noise ratio. Usually the satellite power is specified as an EIRP in dBW, so all we need is the path loss and antenna gain, less any feedline losses, to find the received signal strength. We'll review free-space path loss after we look at another way to specify the noise performance of a receiving installation.

G/T A figure of merit called *G/T* has evolved to measure the combination of antenna gain and equivalent noise temperature for a receiving installation. *G/T* is defined as:

$$G/T(\text{dB}) = G_A(\text{dBi}) - 10 \log (T_r + T_a) \quad (12.8)$$

where

$G/T(\text{dB})$ = figure of merit for the receiving system
$G_A(\text{dBi})$ = receiving antenna gain in dBi
T_r = the noise temperature of the antenna
T_a = the equivalent noise temperature of the receiver

The gain and noise temperatures should all be taken at the same reference point. The gain required is the antenna gain less any losses up to the reference point. As before, the reference point is usually the receiver input. See Figure 12.9. In a satellite receiver, the first stage is often located separately from the rest of the receiver and very close to the antenna. This stage is called the *low-noise amplifier (LNA)*.

FIGURE 12.9

EXAMPLE 12.5

A receiving antenna has a gain of 40 dBi and looks at a sky with a noise temperature of 15 K. The loss between the antenna and the LNA input, due to the feed horn, is 0.4 dB, and the LNA has a noise temperature of 40 K. Calculate *G/T*.

SOLUTION

First, we find G in dB. This is simply the antenna gain less any losses up to the reference point.

$$G = 40 \text{ dBi} - 0.4 \text{ dB}$$
$$= 39.6 \text{ dBi}$$

TABLE 12.1 Geostationary Satellite Systems

System	Inmarsat-3	MSAT
Date of operation	1996–1998	1996
Major uses and coverage	Voice, data, especially for ships Worldwide except polar regions	Voice, data, mainly for land mob Western hemisphere
Number of satellites in active service	4	1
Uplink frequency (mobile to satellite), GHz	1.6	1.6315–1.6605
Downlink frequency (satellite to mobile), GHz	1.5	1.53–1.559
Uplink frequency (gateway to satellite), GHz	6.4	10.75–10.95
Downlink frequency (satellite to gateway), GHz	3.6	13–13.15, 13.2–13.25
Satellite EIRP, dBW	48	57.3

FIGURE 12.12 MSAT terminals

FIGURE 12.13 Mobile telephone network using geostationary satellites

arison of Table 12.1 shows a comparison of some of the important features of Inmarsat-3 and MSAT. Both use bent-pipe transponders connecting mobile users to a gateway that is in turn connected to the PSTN. Figure 12.13 shows the relatively simple network that results.

12.6 Systems Using Geostationary Satellites

In spite of their disadvantages, the relative simplicity of geostationary systems has made them attractive for the first generation of mobile systems. Global coverage can be achieved with only three GEO satellites, and all of North and South America can be covered with one. Let us look first, then, at two of the pioneers in satellite wireless communication. Both of them are still very much alive, and both use geostationary satellites.

Inmarsat *Inmarsat (International Maritime Satellite Organization)* was established in 1979 as an intergovernmental treaty organization (it has since been privatized) and began service in 1982. It is now into its third generation of satellites, known as Inmarsat-3. Originally Inmarsat's mandate was to provide voice and data services to ships at sea, supplementing and eventually partly supplanting high-frequency radio. Since then its services have expanded to include land and aeronautical mobile communication.

Inmarsat uses a total of nine GEO satellites (four are in service, the rest are spares or leased out), covering the whole world except for the polar regions. Each satellite has a hemispherical beam and five spot beams with a maximum EIRP of 48 dBW in the spot beams. Power and bandwidth can be dynamically allocated among the beams. The satellites operate in the L-band (1.5/1.6 GHz).

Inmarsat offers several different types of telephone services. Inmarsat-A is the original analog telephone service. Inmarsat-B is a more modern digital service with similar capabilities. Both of these are intended mainly for communication with ships and need dish antennas with diameters of about 80 cm.

The Inmarsat mini-M service is more closely related to the theme of this book. Using the satellite spot beams, it is designed mainly for operation on land and in coastal waters. A typical portable transceiver, including antenna, is about the size of a laptop computer. It cannot be carried about when in use, but must be set up where its antenna can be aimed at one of Inmarsat's geostationary satellites. See Figure 12.10.

Handheld Inmarsat terminals are available, but only for text messaging services. Inmarsat is a major partner in ICO, a proposed MEO system for use with handheld phones.

FIGURE 12.10
Inmarsat portable terminal
(Courtesy of Inmarsat)

FIGURE 12.11
MSAT footprint

IT (Mobile Satellite) This joint Canadian and United States project uses one GEO satellite to provide coverage for North and Central America, the Caribbean and Hawaii (via a spot beam), and the surrounding coastal waters. See Figure 12.11 for the satellite footprint.

MSAT's satellite is about ten times more powerful than those used by Inmarsat and has an EIRP of at least 57.3 dBW in its coverage area so antennas can be smaller. Mobile terminals use a reasonably compact roof-mounted antenna and portable terminals are about the size of a notebook computer and have a lid-mounted antenna. See Figure 12.12(a) for a typical mobile installation. The geostationary system does not allow for portable telephones that can be carried when in use.

The use of a single geostationary satellite allows MSAT to use a relatively simple network. Only one ground station with an 11 m dish antenna is needed. All calls are relayed through the satellite to the single ground station and from there to the PSTN. See Figure 12.12(b).

TABLE 12.2 "Big LEO" Systems

System	Iridium	Globalstar	Teledesic
Date of operation	1998	1999	2000
Major uses	Voice, paging, low-speed data	Voice, paging, low-speed data	High-speed data, voice
Number of satellites	66	48	288
Satellite altitude (km)	780	1414	1375
Uplink frequency (mobile to satellite), GHz	1.616–1.6265	2.4835–2.500	28.6–29.1
Downlink frequency (satellite to mobile), GHz	1.616–1.6265 (same range as uplink, TDM used)	1.610–1.6265	18.8–19.3
Uplink frequency (gateway to satellite), GHz	29.1–29.3	5.025–5.225	27.6–28.4
Downlink frequency (satellite to gateway), GHz	19.4–19.6	6.875–7.075	17.8–18.6
Intersatellite crosslink frequency, GHz	23.18–23.38	(No crosslinks)	65–71

LEO One LEO one is a proposed "little LEO" system with a similar structure to that of ORBCOMM. It will be designed to use 48 satellites at an altitude of 950 km for paging and short messaging. It is expected to be operational in 2002.

E-Sat E-Sat, expected to begin this year (2000), is an interesting "special case" LEO system. Using only six satellites, orbiting at an altitude of 1260 km, and using CDMA, it is intended to concentrate on the niche market of remote meter reading, especially in rural areas. Since meter reading is not urgent, store-and-forward technology is very appropriate for this type of system.

Table 12.3 summarizes the specifications of the little LEO satellite systems discussed in this chapter.

Dr. Anuradha Sugam
M.Sc., B.Ed., M.Phil., Ph.D.
Assistant Professor of Physics
PG & Research Department of Physics
Arnad Vaishnavam Arts & Science College
Thanjavur - 7. Cell : 9751853666

ttle LEOs"

In addition to the huge projects sometimes referred to as "big LEOs," there are a number of more modest schemes, both existing and proposed, that exist only to provide low-data-rate digital services such as paging, short messaging, and vehicle tracking for trucking companies. These operations are called "little LEOs" because the systems are simpler and smaller. It is not necessary that there be a satellite in view of all stations at all times for these services because messages can be stored briefly and forwarded when a satellite becomes available (typically within a few minutes). Here are some examples of "little LEO" systems.

ORBCOMM The ORBCOMM system went into operation in 1998. It uses 28 LEO satellites and plans to increase this to 48. The system is used for short messaging, e-mail, and vehicle tracking. Unlike the big LEOs, little LEOs typically use frequencies in the VHF range to communicate with customer earth stations. ORBCOMM's uplink from mobile to satellite is at 148–150.05 MHz, with downlink at 137–138 MHz.

angle to the equator, thus providing global coverage. In fact, most of the time a given location will be in view of two satellites to provide diversity. In case the view of one is blocked (by a tall building, for instance), the other satellite would be able to support 4,500 telephone calls using TDMA technology. The satellites are to use a simple bent-pipe configuration. The frequency ranges are projected to be similar to those for Ellipso, in the area of 1.6 GHz for satellite-to-mobile links.

nmary

The main points to remember from this chapter are:

- Satellites are especially useful for telecommunication in remote areas where terrestrial cellular systems are prohibitively expensive or impossible to build.
- The orbital period of a satellite in a circular orbit depends on its distance from the earth, with satellites farther from the earth having a longer period.
- Satellite orbits are classified according to their distance from earth, as low earth orbit (LEO), medium earth orbit (MEO), and geostationary earth orbit (GEO). Geostationary satellites appear stationary at a point above the equator.
- Systems using satellites in lower orbits have lower path loss and shorter propagation times, but require more satellites for real-time coverage.
- Satellites can act as simple repeaters or can contain elaborate switching systems to route calls. They can also store data for later forwarding.
- Current satellite systems for mobile communication use either GEO or LEO satellites. Only the latter systems allow handheld transceivers. There are MEO systems in the planning and construction stages.

on List

$$v = \sqrt{\frac{4 \times 10^{14}}{(d + 6400)}} \quad (12.1)$$

$$P_N = kTB \quad (12.2)$$

$$NF = \frac{(S/N)_i}{(S/N)_o} \quad (12.3)$$

$$NF_{dB} = 10 \log NF \quad (12.4)$$

$$T_a = 290(NF - 1) \quad (12.5)$$

$$T = T_{ra} + T_s \quad (12.6)$$

$$T_s = \frac{(L_r - 1)290 + T_{ar}}{L_r} \quad (12.7)$$

$$G/T(\text{dB}) = G_r(\text{dBi}) - 10 \log (T_s + T_{ra}) \quad (12.8)$$

$$\frac{C}{N}(\text{dB}) = EIRP(\text{dBW}) - L_{fs}(\text{dB}) + G/T - k(\text{dBW}) \quad (12.9)$$

$$L_{fs} = 32.44 + 20 \log d + 20 \log f \quad (12.10)$$

Key Terms

apogee point in a satellite orbit that is farthest from the earth
bent-pipe configuration a design of satellite transponder in which signals are amplified and shifted in frequency but do not undergo any other processing
constellation in satellite telephony, a group of satellites coordinated in such a way as to provide continuous communication
crosslink a radio or optical connection directly between satellites, without going through an earth station
equatorial orbit a satellite orbit that is entirely above the equator
footprint depiction of the signal strength contours from a satellite transmitter on the earth
geostationary orbit satellite orbit in which the satellite appears to remain stationary at a point above the equator
geosynchronous orbit satellite orbit in which the satellite's period of revolution is equal to the period of rotation of the earth
hemispheric beam antenna beam on a geostationary satellite that is adjusted to cover the whole earth
low-earth-orbit (LEO) satellite an artificial satellite orbiting the earth at an altitude less than about 1500 kilometers
medium-earth-orbit (MEO) satellite a satellite in orbit at a distance above the earth's surface of approximately 8,000 to 20,000 km
perigee the point in a satellite orbit that is closest to the earth
polar orbit a satellite orbit passing over the north and south poles

FIGURE 12.14 Mobile telephone network using LEO satellite and crosslinks

authorities forbade it), users generally preferred to use their local cellular system where possible to save on airtime costs. Airtime for Iridium can several dollars per minute, compared to a few cents per minute for most terrestrial cellular systems. The high cost of airtime, and of the phones themselves, may have been contributing factors to Iridium's financial problems.

Globalstar

The Globalstar system began commercial operation in 1999, and by April 2000 service was available in more than 100 countries.

This system is slightly less ambitious than Iridium. It uses a constellation of 48 LEO satellites (plus four spares) at an altitude of 1414 km. The satellites use simple "bent pipe" transponders but have high power (about 1 kW per satellite). The satellites are in eight orbital planes of six satellites each, inclined at 52 degrees with respect to the equator. This allows the system to provide service from 70 degrees North latitude to 70 degrees South, which includes most of the Earth except for the polar regions.

CDMA is used, allowing a ground user to access two or more satellites simultaneously, provided they are in view, and utilizes the soft handoff techniques introduced for CDMA PCS services.

Because there is no switching on the satellites, communication is possible only when at least one satellite is visible simultaneously from the mobile phone and a ground station. This will require at least 36 ground stations, called *gateways*, for worldwide coverage. By April 2000 there were 18 gateways in operation.

12.7 Systems Using Low-Earth-Orbit Satellites

LEO satellite systems are very attractive, especially for use with handheld portable phones. The short distance to the satellite allows transmitter power and antenna gain requirements to be relaxed. This permits the use of portable phones that are only somewhat larger than a conventional cell phone. On the other hand, such a system requires many satellites (on the order of 40 to 70) and a complex network.

LEO systems are the most complex and expensive wireless communication systems yet devised. Several proposed systems have been cancelled before becoming operational due to a shortage of funds. As of April 2000, one LEO wireless telephony system (Globalstar) was operating, and another (Iridium) had been forced to discontinue service after only a year in operation, due to bankruptcy. If their technical advantages can be coupled with the necessary financial resources, LEO systems look like the ultimate in satellite telephony.

Iridium

Iridium began service in November 1998—and applied for bankruptcy protection in August 1999. As of April 2000, many attempts to restructure and continue to operate the network had failed, all commercial operation had ceased, and it seemed very likely that the satellites would be destroyed by deliberately taking them out of orbit. Nonetheless, a brief look at the Iridium system will give useful insights into the possibilities of LEO satellite telephony. Technically, the system was a success, and if the economic obstacles could be overcome, a similar system could be built in the future.

The Iridium system comprises 66 LEO satellites in a complex constellation, such that at least one satellite is visible from any location on earth at all times. The satellites are crosslinked; telephone calls can traverse the network from one satellite to another before being relayed to a ground station. This means that it is not necessary for every satellite to be in range of a ground station at all times and reduces the number of ground stations required. See Figure 12.14 for an illustration of this idea.

Iridium uses digital modulation, with a combination of FDMA and TDMA to assign channels.

Because Iridium's satellites are powerful and close to the ground, portable phones are usable with this system. The phone weighs about half a kilogram and has a fairly awkward helical antenna which must be positioned vertically and with a clear view of the sky for reliable operation. For operation where terrestrial cellular radio is available, a *cellular cassette* with its own antenna can be installed in the back of the phone. Although the Iridium satellite system was available nearly everywhere (in a few countries local

4. A receiving installation has a G/T of 30 dB. If its antenna gain is 50 dBi and the antenna noise temperature is 25 K, what is the equivalent noise temperature of the receiver?
5. How could the G/T of the system described in the preceding problem be improved to 35 dB? Give two ways, and perform calculations for each.
6. Find the noise temperature of a receiver with a noise figure of 1 dB.
7. Find the noise temperature of an antenna on a satellite if it looks at the earth, giving it a "sky" noise temperature of 290 K, and is coupled to the reference plane by a waveguide with a loss of 0.3 dB.
8. Compare the time delay for a signal transmitted by geostationary satellite to that for a signal transmitted via terrestrial microwave radio link over a distance of 2000 km. Assume the distance from ground to satellite is 37,000 km at both ends of the path. Ignore any time delay in the electronics for both systems.
9. Compare the signal strengths from two satellites as received on the ground. One is geostationary, with a path length of 40,000 km; the other is in low earth orbit, with a path length of 500 km. Assume all other factors are equal. Which signal is stronger, and by how many decibels?
10. It is possible to communicate between points on earth by using the moon as a passive reflector. In fact, it is done routinely by radio amateurs. Calculate the time delay and round-trip path loss at 1 GHz using the moon as a reflector. (The distance to the moon is calculated in problem 2.) Note that your path loss calculation ignores losses in the reflection at the moon's surface. What can you conclude about the likely commercial possibilities of this idea?
11. (a) Calculate the path loss for a satellite in the Iridium system on the link from satellite to mobile, assuming that the satellite is directly overhead.
    (b) Calculate the round-trip time delay for the Iridium system under the same conditions as in part (a).
12. Repeat the previous problem, but use the ICO MEO system. Compare the results and draw some conclusions about the differences between LEO and MEO systems.
13. Calculate the velocity and orbital period for each of the following types of satellite:
    (a) Iridium
    (b) Globalstar
    (c) Teledesic
    (d) ORBCOMM

13

Paging and Wireless Data Networking

Objectives

After studying this chapter, you should be able to:

- Describe and explain the operation of several systems used for one- and two-way paging.
- Compare paging systems with respect to capabilities and complexity.
- Describe the operation of voice paging systems.
- Describe the operation of wired Ethernet LANs.
- Explain the need for wireless LAN equipment.
- Discuss the IEEE 802.11 and Bluetooth standards and suggest which would be preferred for given applications.
- Explain the need for and the operation of wireless Ethernet bridges and modems.
- Describe the operation of infrared LANs and compare them to wired LANs and wireless LANs using radio.
- Describe and compare public packet-data networks and compare them with other kinds of wireless data communication.

Ellipso Ellipso uses an interesting combination of elliptical and circular orbits. Its constellation is based on the fact that there is far more land mass and far greater population in high northern latitudes than in similar southern latitudes. A glance at any globe will confirm that most of the world's land mass is north of 40° south latitude.

The Ellipso system is designed to take advantage of this asymmetry. It will initially include six satellites, later increasing to ten, in a circular orbit about 8000 km above the equator for worldwide coverage at low latitudes. These are to be complemented by eight active satellites (plus two spares) in inclined elliptical orbits, designed so that they spend most of their time above the northern hemisphere. The elliptical orbits have a maximum height of approximately 7800 km and a minimum height of approximately 520 km. The elliptical-orbit satellites will have an orbital period of about three hours. See Figure 12.15.

FIGURE 12.15

Ellipso's coverage is scheduled to be phased in gradually, beginning with the deployment of the circular-orbit satellites in 2002. These satellites can provide coverage to about 80% of the world's population.

The main focus of the Ellipso system is expected to be voice communication using portable and mobile terminals that are projected to be roughly the size of conventional cell phones but are likely to be considerably larger. CDMA is used, with the uplinks of all satellites receiving on the same frequency bands. This allows use of the soft-handoff feature of CDMA and avoids any necessity for portable phones to switch frequencies. The satellites will relay signals directly to ground-station gateways, with no on-satellite processing or intersatellite links.

ICO *ICO* stands for *Intermediate Circular Orbit*. The plan, initiated by Inmarsat but since spun off and privatized, is to launch ten operational satellites and two spares in two orthogonal planes at an altitude of 10,353 km, each at a 45°

"Little LEO" Systems

em	ORBCOMM	LEO One	E-Sat
ation	1998	2002	2000
	Paging, short messaging/ e-mail, vehicle location	Paging, short messaging/ e-mail, vehicle location	Remote meter reading
satellites	28	48	6
tude (km)	785	950	1360
iency ateilite),	148–150.05	148–150.05	148–148.905
equency mobile),	137–138, 400	137–138	137.0725–137.9725
iency satellite),	149-61	148–150.05	148–148.905
equency gateway),	137–138	400.15–401	137.0725–137.9725

12.8 Systems Using Medium-Earth-Orbit Satellites

Satellites in medium earth orbit are a compromise between the LEO and GEO systems. More satellites are needed than for GEO (on the order of 6 to 20 for real-time communication), but fewer than for LEO. Delay and propagation loss are much less than for GEO, but greater than LEO. Portable phones are possible with MEO systems, but they must be heavier and bulkier than for LEO systems.

The main advantage of using MEO rather than LEO satellites is financial. These systems promise rates for airtime that are at least on the same order of those for terrestrial cellular systems, unlike LEO systems.

At this printing no MEO systems are up and running, but some proposed systems appear likely to become operational in the near future.

LAN Topologies

LAN physical topology defines the geographical arrangement of networking devices.
Topologies are driven fundamentally by two network connection types:
A point-to-point connection is a direct link between two devices. For example, when you attach your computer to a printer, you have created a point-to-point link. In networking terms, most of the today's point-to-point connections are associated with modems and PSTN (Public Switched Telephone Network) communications because only two devices share point-to-point connections, it defeats the purpose of a shared network.

A multipoint connection, on the other hand, is a link between three or more devices. Historically, multipoint connections were used to attach central CPUs to distributed dumb terminals.
In today's LAN environments, multipoint connections link many network devices in various configurations.

The major topologies of LAN are:

1. Bus Topology
2. Ring Topology
3. Star Topology
4. Mesh Topology
5. Cellular Topology
6. Hybrid Topology

Bus Topology

The physical bus topology is the simplest and most widely used of the network designs. It consists of one continuous length of cabling (trunk) and a terminating resistor (terminator) at each end. The data communications message travels along the bus in both directions until it is picked up by a workstation or server NIC.

If the message is missed or not recognized, it reaches the end of the cabling and dissipates at the terminator. All nodes in the bus topology have equal access to the trunk – no discriminating here. This is accomplished using short drop cables or direct T-connectors.

12. Why is it necessary to use multiple satellites for real-time coverage with LEO and MEO systems?
13. How can the use of multiple satellites be avoided for data communication with MEO and LEO satellites?
14. What advantages and disadvantages do satellites in low earth orbit have compared with geostationary satellites for mobile communication systems?
15. What is meant by a "bent-pipe" satellite transponder?
16. What is meant by the term *backoff* with respect to a transponder power amplifier, and when is it necessary to use it?
17. What are crosslinks and how are they useful in a satellite communication system?
18. What satellite communication system is especially designed for use by ships at sea?
19. Compare the Inmarsat mini-M and MSAT services in terms of coverage and convenience.
20. Compare Iridium with MSAT in terms of coverage and convenience.
21. Compare Iridium with Globalstar in terms of the way the networks are organized.
22. How does the Teledesic system differ from the other LEO systems described in this chapter?
23. What are the differences between "big LEO" and "little LEO" systems?
24. Why are store-and-forward techniques unsuitable for voice communication?
25. Explain how the Ellipso system can achieve worldwide coverage with fewer satellites than the other systems discussed in this chapter.
26. Compare the LEO and MEO concepts for voice communication. What advantages does each have?

Problems

1. Find the orbital velocity and period for a satellite that is 1000 km above the earth's surface.
2. The moon orbits the earth with a period of approximately 28 days. How far away is it?
3. What velocity would a satellite need to have to orbit just above the surface of the earth? Why would such an orbit be impossible in practice?

spot beam in a satellite system, a focused beam of energy that covers a relatively small area on the earth, produced by a high-gain antenna on the satellite

store-and-forward technique in digital communication, the use of a device (repeater) to receive one or more data packets, store them, and retransmit them at a later time

tracking in satellite communication, continuously adjusting the position of a directional antenna on the ground, so that it always points at the satellite

transponder in satellite communication, a repeater located on a satellite

ions

1. Compare the following in terms of cost and practical communication distance:
   (a) repeaters on towers
   (b) satellites in low earth orbit
   (c) geostationary satellites
2. How does the orbital period of a satellite change as it moves farther from the earth?
3. Sketch the earth and the orbit of a geostationary satellite, approximately to scale.
4. Why do all geostationary satellites orbit the earth at the same distance and above the equator?
5. Why are geostationary satellites unusable from earth stations in the polar regions?
6. What are the Van Allen belts and what effect do they have on the placement of satellites?
7. Explain the advantages of the use of elliptical orbits for satellite communication.
8. Why are spot beams used with geostationary satellites?
9. How many geostationary satellites are necessary for a system to have worldwide coverage (except in the polar regions)?
10. What is Doppler shift? Why does the effect increase as the height of a satellite's orbit decreases?
11. Why is it necessary for mobile systems to use antenna tracking, even with geostationary satellites?

This design is easy to install because the backbone trunk traverses the LAN as one cable segment. This minimizes the amount of transmission media required. Also, the number of devices and length of the trunk can be easily expanded.

Advantages of Bus Topology:

1. It uses established standards and it is relatively easy to install.
2. Requires fewer media than other topologies.

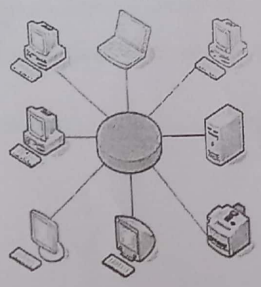
Disadvantages of Bus Topology:

1. The bus networks are difficult to reconfigure, especially when the acceptable number of connections or maximum distances have been reached.
2. They are also difficult to troubleshoot because everything happens on a single media segment. This can have dangerous consequences because any break in the cabling brings the network to its knees.

Ring Topology

As its name implies, the physical ring topology is a circular loop of point-to-point links. Each device connects directly or indirectly to the ring through an interface device or drop cable. Messages travel around the ring from node to node in very organized manner. Each workstation checks the messages for a matching destination address.

RING Topology



If the address doesn't match, the node simply regenerates the message and sends it on its way. If the address matches, the node accepts the message and sends a reply to the originating sender. Initially, ring topologies are moderately simple to install; however, they require more media than bus systems because the loop must be closed.

Once your ring has been installed, it's a bit more difficult to reconfigure. Ring segments must be divided or replaced every time they're changed. Moreover, any break in the loop can affect all devices on the network.

Advantages of Ring Topology:

1. They are very easy to troubleshoot because each device incorporates a repeater.
2. A special internal feature called becoming, allows the troubled workstation to identify themselves quickly.

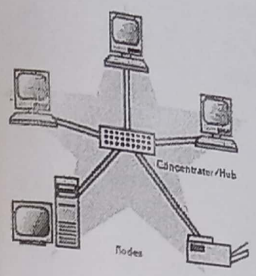
Disadvantages of Ring Topology:

1. It is considerably difficult to install and reconfigure ring topology.
2. Media failure on unidirectional or single loop causes complete network failure.

Star Topology

The Physical star topology uses a central controlling hub with dedicated legs pointing in all directions – like points of a star. Each network devices has a dedicated point-to-point link to the central hub. This strategy prevents troublesome collisions and keeps the line of communication open and free of traffic.

Star Topology



Star topologies are somewhat difficult to install because each device gets its own dedicated segment. Obviously, they require a great deal of cabling. This design provides an excellent platform for reconfiguration and troubleshooting.

Changes to the network are as simple as plugging another segment into the hub. In addition, a break in the LAN is easy to isolate and doesn't affect the rest of the network.

Advantages of Star Topology:

1. Relatively easy to configure.
2. Easy to troubleshoot.
3. Media faults are automatically isolated to the failed segment.

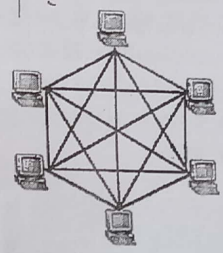
Disadvantages of Star Topology:

1. Requires more cable than most topologies.
2. Moderately difficult to install.'

Mesh Topology

The mesh topology is the only true point-to-point design. It uses a dedicated link between every device on the network. This design is not very practical because of its excessive waste of transmission media. This topology is difficult to install and reconfigure.

Mesh Topology



Moreover, as the number of devices increases geometrically, the speed of communication also become slow. ATM (Asynchronous Transfer Mode) and switched Hubs are the example of high-speed Mesh implementation.

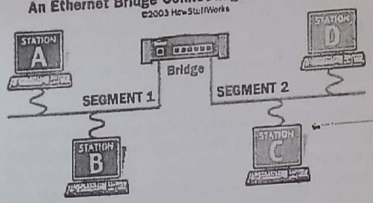
Advantages of Mesh Topology:

1. Easy to troubleshoot because each link is independent of all others.
2. You can easily identify faults and isolate the affected links. Because of the high number of redundant paths, multiple links can fail before the failure affects any network device.

Bridges

PREV NEXT

An Ethernet Bridge Connecting Two Segments



To alleviate problems with segmentation, Ethernet networks implemented **bridges.** Bridges connect two or more network segments, increasing the network diameter as a repeater does, but bridges also help **regulate traffic.** They can send and receive transmissions just like any other node, but they do not function the same as a normal node. The bridge does not originate any traffic of its own; like a repeater, it only echoes what it hears from other stations. (That last statement is not entirely accurate: Bridges do create a special Ethernet frame that allows them to communicate with other bridges, but that is outside the scope of this article.)

Remember how the multiple access and shared medium of Ethernet meant that every station on the wire received every transmission, whether it was the intended recipient or not? Bridges make use of this feature to relay traffic between segments. In the figure above, the bridge connects segments 1 and 2. If station A or B were to transmit, the bridge would also receive the transmission on segment 1. How should the bridge respond to this traffic? It could automatically transmit the frame onto segment 2, like a repeater, but that would not relieve congestion, as the network would behave like one long segment.

Ethernet bridges

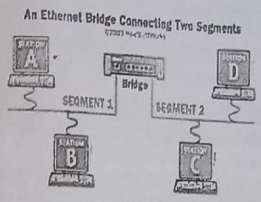
To alleviate problems with segmentation, Ethernet networks implemented **bridges.** Bridges connect two or more network segments, increasing the network diameter as a repeater does, but bridges also help **regulate traffic.** They can send and receive transmissions just like any other node, but they do not function the same as a normal node. The bridge does not originate any traffic of its own; like a repeater, it only echoes what it hears from other stations. (That last statement is not entirely accurate: Bridges do create a special Ethernet frame that allows them to communicate with other bridges, but that is outside the scope of this article.)

Remember how the multiple access and shared medium of Ethernet meant that every station on the wire received every transmission, whether it was the intended recipient or not? Bridges make use of this feature to relay traffic between segments. In the figure above, the bridge

connects segments 1 and 2. If station A or B were to transmit, the bridge would also receive the transmission on segment 1. How should the bridge respond to this traffic? It could automatically transmit the frame onto segment 2, like a repeater, but that would not relieve congestion, as the network would behave like one long segment.

Bridges

An Ethernet Bridge Connecting Two Segments



One goal of the bridge is to **reduce unnecessary traffic** on both segments. It does this by examining the destination address of the frame before deciding how to handle it. If the destination address is that of station A or B, then there is no need for the frame to appear on segment 2. In this case, the bridge does nothing. We can say that the bridge **filters** or drops the frame. If the destination address is that of station C or D, or if it is the broadcast address, then the bridge will transmit, or **forward** the frame on to segment 2. By forwarding packets, the bridge allows any of the four devices in the figure to communicate. Additionally, by filtering packets when appropriate, the bridge makes it possible for station A to transmit to station B at the same time that station C transmits to station D, allowing two conversations to occur simultaneously!

Switches are the modern counterparts of bridges, functionally equivalent but offering a **dedicated segment** for every node on the network

2 WLAN

Stands for "Wireless Local Area Network." A WLAN, or wireless LAN, is a network that allows devices to connect and communicate wirelessly. Unlike a traditional wired LAN, in which devices communicate over Ethernet cables, devices on a WLAN communicate via Wi-Fi.

While a WLAN may look different than a traditional LAN, it functions the same way. New devices are typically added and configured using DHCP. They can communicate with other devices on the network the same way they would on a wired network. The primary difference is how the data is transmitted. In a LAN, data is transmitted over physical cables in a series of Ethernet packets containing. In a WLAN, data is transmitted over the air using one of Wi-Fi 802.11 protocols.

As wireless devices have grown in popularity, so have WLANs. In fact, most routers sold are now wireless routers. A wireless router serves as a base station, providing wireless connections to any Wi-Fi-enabled devices within range of the router's wireless signal. This includes laptops, tablets, smartphones, and other wireless devices, such as smart appliances

and smart home controllers. Wireless routers often connect to a cable modem or other Internet-connected device to provide Internet access to connected devices. LANs and WLANs can be merged together using a bridge that connects the two networks. Many wireless routers also include Ethernet ports, providing connections for a limited number of wireless devices. In most cases, wireless routers act as a bridge, merging the Ethernet and Wi-Fi-connected devices into the same network. This allows wired and wireless devices to communicate with each other through a single router.

Advantages of WLANs

The most obvious advantage of a WLAN is that devices can connect wirelessly, eliminating the need for cables. This allows homes and businesses to create local networks without wiring the building with Ethernet. It also provides a way for small devices, such as smartphones and tablets, to connect to the network. WLANs are not limited by the number of physical ports on the router and therefore can support dozens or even hundreds of devices. The range of a WLAN can easily be extended by adding one or more repeaters. Finally, a WLAN can be easily upgraded by replacing routers with new versions — a much easier and cheaper solution than upgrading old Ethernet cables.

Disadvantages of WLANs

Wireless networks are naturally less secure than wired networks. Any wireless device can attempt to connect to a WLAN, so it is important to limit access to the network if security is a concern. This is typically done using wireless authentication such as WEP or WPA, which encrypts the communication. Additionally, wireless networks are more susceptible to interference from other signals or physical barriers, such as concrete walls. Since LANs offer the highest performance and security, they are still used for many corporate and government networks.

Radio LAN

Radio Local Area Networks (RLANs) are intended to cover smaller geographic areas like homes, offices and to a certain extent buildings being adjacent to each other. Radio LANs are also known as Wireless LANs (WLANs).

A popular deployment of Radio LANs is providing broadband connectivity at public locations like airports, railway stations, conference centres, hotels and street cafés. Even on trains and aboard aircraft Radio LANs are or will become available for providing network access. Radio LANs are also rather popular at home and at the office enabling the users to connect all equipment wirelessly.

Currently, the frequency bands 2.4 GHz and 5 GHz are mainly used by Radio LANs and in many cases, the deployed technology is based on the IEEE 802.11 standards family. However, other technologies such as LTE-LAA are deployed in those frequency bands as well.

A radio LAN is composed of the following components:

- **Access point(s),**
  A radio LAN can be built around one or more access points. The typical range of an access point is in the order of 10-20 m (30-60 ft.). Above this range the througput will dramatically drop off. For home applications, this range is usually enough. For a home network the wireless access point can be integrated with a router and/or an ADSL or cable modem. For a corporate network a number of access points might be needed.
- **Wireless LAN PC card,**
  For each computer that is to be connected to the wireless network an wireless adapter is needed. This can be an internal PC card or an external USB or PCMCIA card.

Most wireless LANs operate in the unlicensed 2.4 GHz band using the IEEE 802.11b or 802.11g standard, more commonly known as Wi-Fi. In this band , the radio LANs have to share the spectrum with a lot of other applications. There is also a more exclusive band available for RLANs, ranging from 5150-5350 and 5470 - 5725 MHz. Expectations are that the wireless LANs will gradually migrate to the 5 GHz range.

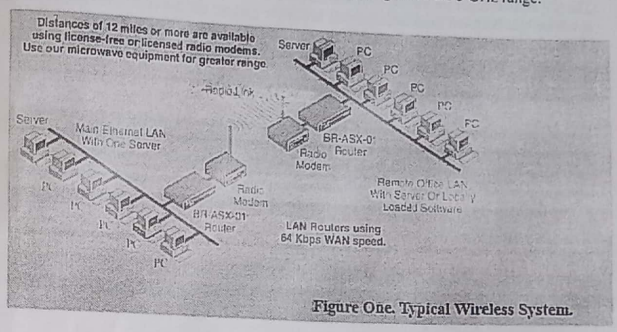


Figure One. Typical Wireless System.

Bluetooth

Bluetooth technology is a short-range wireless communications technology to replace the cables connecting electronic devices, allowing a person to have a phone conversation via a headset, use a wireless mouse and synchronize information from a mobile phone to a PC, all using the same core system.

The Bluetooth RF transceiver (or physical layer) operates in the unlicensed ISM band centered at 2.4 gigahertz (the same range of frequencies used by microwaves and Wi-Fi). The core system employs a frequency-hopping transceiver to combat interference and fading.

Bluetooth devices are managed using an RF topology known as a "star topology." A group of devices synchronized in this fashion forms a piconet, which may contain one master and up to seven active slaves, with additional slaves that are not actively participating in the network. (A given device may also be part of one or more piconets, either as a master or as a slave.) In a piconet, the physical radio channel is shared by a group of devices that are synchronized to a common clock and frequency-hopping pattern, with the master device providing the synchronization references.

Let's say the master device is your mobile phone. All of the other devices in your piconet are known as slaves. This could include your headset, GPS receiver, MP3 player, car stereo, and so on.

Devices in a piconet use a specific frequency-hopping pattern, which is algorithmically determined by the master device. The basic hopping pattern is a pseudorandom ordering of the 79 frequencies in the ISM band. The hopping pattern may be adapted to exclude a portion of the frequencies that are used by interfering devices. The adaptive hopping technique improves Bluetooth technology's coexistence with static (nonhopping) ISM systems, such as Wi-Fi networks, when these are located in the vicinity of a piconet.

The physical channel (or the wireless link) is subdivided into time units known as slots. Data is transmitted between Bluetooth-enabled devices in packets that are positioned in these slots. Frequency hopping takes place between the transmission or reception of packets, so the packets that make up one transmission may be sent over different frequencies within the ISM band.

The physical channel is also used as a transport for one or more logical links that support synchronous and asynchronous traffic as well as broadcast traffic. Each type of link has a specific use. For instance, synchronous traffic is used to carry hands-free audio data, while asynchronous traffic may carry other forms of data that can withstand more variability in the timing for delivery, such as printing a file or synchronizing your calendar between your phone and computer.

One of the complexities often associated with wireless technology is the process of connecting wireless devices. Users have become accustomed to the process of connecting wired devices by plugging one end of a cable into one device and the other end into the complementary device.

Bluetooth technology uses the principles of device "inquiry" and "inquiry scan." Scanning devices listen in on known frequencies for devices that are actively inquiring. When an inquiry is received, the scanning device sends a response with the information needed for the inquiring device to determine and display the nature of the device that has recognized its signal.

Let's say you want to wirelessly print a picture from your mobile phone to a nearby printer. In this case, you go to the picture on your phone and select print as an option for sending that picture. The phone would begin searching for devices in the area. The printer (the scanning device) would respond to the inquiry and, as a result, would appear on the phone as an available printing device. By responding, the printer is ready to accept the connection. When you select the Bluetooth wireless printer, the printing process kicks off by establishing connections at successively higher layers of the Bluetooth protocol stack that, in this case, control the printing function.

Like any successful technology, all of this complexity goes on without the user being aware of anything more than the task he or she is trying to complete, like connecting devices and talking hands-free or listening to high-quality stereo music on wireless headphones.

Wireless Bridge

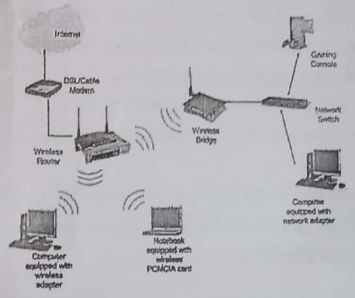
In its most basic form a wireless bridge is created by linking 2 access points together. One access point adopts the role of being an "access point" and the other the role of being a "client" or "station". The client access point connects to the other access point in a very

everyday access point. The difference with a point-to-point wireless bridge is that the connection is an exclusive one between the 2 devices.

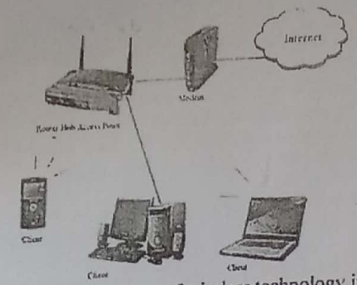
Another key difference is that wireless bridges are very directional. In most access points the RF energy is transmitted in a 360degree coverage pattern. This is useful in wireless networking as usually the client devices are either mobile or there are multiple of them. This means that the access point needs to be able to connect to them wherever they may be in relation to it. With a wireless bridge however, in most cases both points are fixed so any RF energy not directed towards the other device forming the bridge is effectively wasted.

Wireless bridges are generally required to provide links over long distances. In order to do this the devices must focus the RF energy. To explain why this is the case, think about what you do when you want to shout at someone who is quite far away from you… generally you would cup your hands around your mouth. This allows you to throw your voice further and be heard at greater distances. On the flip side if you are struggling to hear someone who is quite far away from you, then you may choose to cup your hands around your ear in the direction of the person speaking. This allows you to block out other sounds and focus on the person talking to you. This is exactly what a wireless bridge does.

Typically a wireless bridge is a Layer 2 Connection between 2 wired ports. If a link is Layer 2 it means that it uses switching rather than a combination of switching and routing in order for the packets to get to their destination. Both ends of the link will exist within the same Subnet. You should effectively treat most wireless bridges as if they were simply a long ethernet cable. In fact, it actually doesn't matter what IP addresses you give the radio units at each end of the link, the link will still work, you may however not be able to log into them to manage the link unless they are addressed correctly.



Connections using infrared wireless modems



IR wireless is the use of wireless technology in devices or systems that convey data through infrared (IR) radiation. Infrared is electromagnetic energy at a wavelength or wavelengths somewhat longer than those of red light. The shortest-wavelength IR borders visiblered in the electromagnetic radiation spectrum;the longest-wavelength IR borders radio waves.

Some engineers consider IR technology to be a sub-specialty of optical technology. The hardware is similar, and the two forms of energy behave in much the same way. But strictly speaking, "optical" refers to *visible*electromagnetic radiation, while "infrared" is *invisible* to the unaidedeye. To compound the confusion, IR is sometimes called "infrared light."

IR wireless is used for short- and medium-range communications andcontrol. Some systems operate in *line-of-sight mode*; this means that theremust be a visually unobstructed straight line through space between the transmitter(source) and receiver (destination). Other systems operate in *diffuse mode*,also called *scatter mode*. This type of system can function when the sourceand destination are not directly visible to each other. An example is a televisionremote-control box. The box does not have to be pointed directly at the set,although the box must be in the same room as the set, or just outside the room with thedoor open.

IR wireless technology is used in intrusion detectors; home-entertainment control units; robot control systems; medium-range, line-of-sight laser communications; cordless microphones, headsets, modems, and printers and other peripherals.

(10)

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.

5.5. 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.

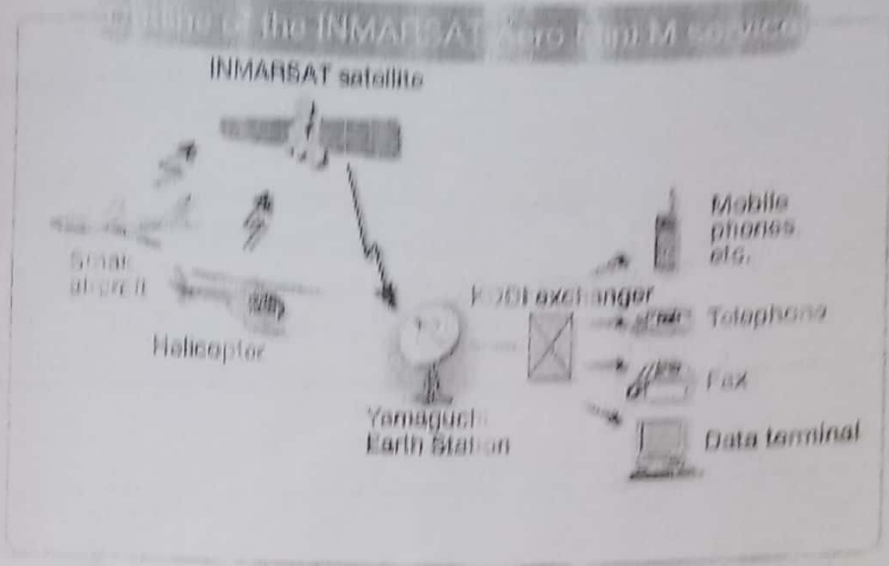


Figure 5.7 INMARSAT Satellite Service

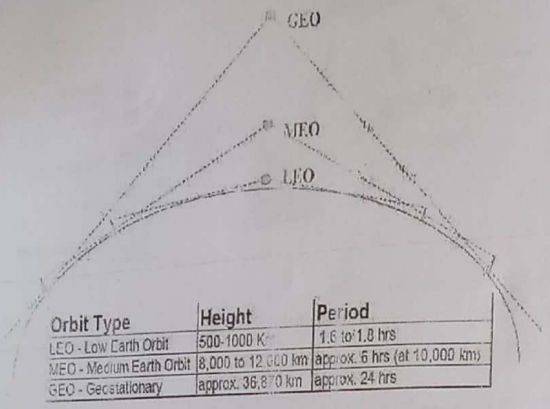
GMDSS services continue to be provided by:

- 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

5.6 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 K	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

5.7 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

5.8 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)

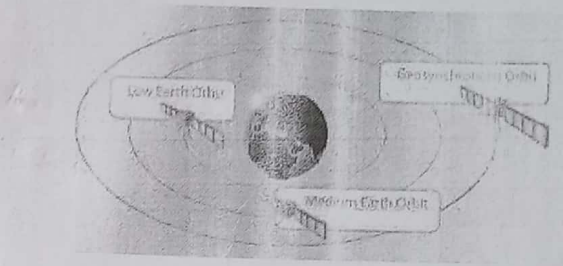


Figure 5.9 LEO, MEO & GEO Orbits

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.
	TUPSO	24	4	6	A:7800km P:520km	63.4
	MOLNIYA	4	1	4	A:39863km P:504km	63.4
	ARCHIMEDES	4	4	1	A:39417km P:926km	63.4

Figure 5.10 Diff b/w LEO, MEO & GEO Orbits

EO: 35,786 km above the earth, MEO: 8,000-20,000 km above the earth & LEO: 500-
000 km above the earth.

Satellite Navigational System:

Benefits:

- Enhanced Safety
- Increased Capacity
- Reduced Delays

Advantage:

- Increased Flight Efficiencies
- Increased Schedule Predictability
- Environmentally Beneficial Procedures

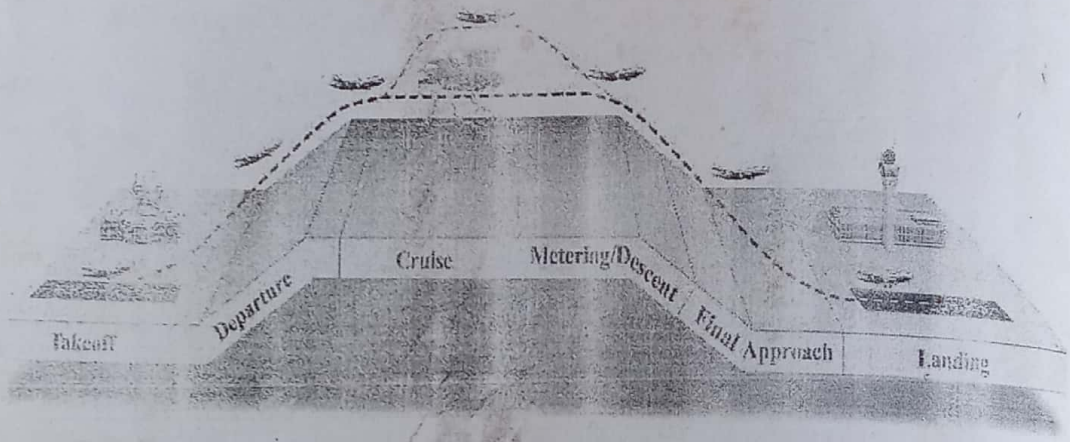


Figure 5.11 LEO, MEO & GEO Orbits

- 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