Bruce R. Elbert

INTRODUCTION TO
SATELLITE
COMMUNICATION

THIRD EDITION

Introduction to
Satellite Communication

Third Edition

Introduction to
Satellite Communication

Third Edition

Bruce R. Elbert

10 9 8 7 6 5 4 3 2 1

Contents

CHAPTER 6
Spacecraft and Repeater   195

CHAPTER 7
Spacecraft Antennas   237

CHAPTER 8

Spacecraft Mission and Bus Subsystems   257

CHAPTER 9

Earth Stations and Network Technology   291

CHAPTER 10

CHAPTER 11

Preface

Introduction to Satellite Communication is designed to meet the needs of working professionals and students. The first edition was a response to a request by many friends and associates for a basic and clear book that provides newcomers with an accessible way to gain knowledge and become productive. The second edition followed the same approach, updating the older and dated material and adding background in newer systems, particularly the global mobile personal communications systems introduced worldwide. This third edition brings the text in line with industry and technology trends. Whether the reader is technically trained or not, the need exists for an authoritative guidebook to the construction and usage of satellite networks.

This book is designed to give you, the reader, an understanding that should permit you to begin work as a satellite professional or as a user of satellite communication. Sufficient technical information has been included to instill a feeling for how systems are designed and operate. Many categories of professionals and students should profit from a significant portion of the material. The book's explanatory nature and broad coverage make it suitable as a textbook for university programs and internal training in communication systems design and planning. Nontechnical professionals in associated business management, contracts, legal, and financial fields will find the book particularly helpful when they must deal with telecommunication projects and issues.

The book is organized into 12 chapters to correspond to the major areas of commercial satellite communication systems. Chapter 1, "Fundamentals of Satellite Systems," identifies the structure and key features of satellite communication and reviews some of the more basic concepts in a nontechnical style. It is understandable to all readers, including high school students. Likewise, Chapter 2, "Evolution of Satellite Communication," provides an easy-to-understand history of the technology and its applications. It begins with geostationary Earth orbit (GEO) systems, which are the foundation of the industry, and moves into non-GEO systems used in mobile applications. Another purpose of Chapter 2 is to capture in one place the background with which many newcomers otherwise would not be acquainted (e.g., how we got from Arthur C. Clarke's concept in 1945 to the present day). Chapter 3, "Satellite Network Architectures," covers the ways in which satellite links can be applied to practical communication problems. It gives the reader an appreciation for the variety of uses in which satellites have gained a stronghold. Technologists involved with spacecraft or communication systems will find that Chapter 3 explains many of the mysteries surrounding the business of using satellites.

The next two chapters begin the core technical material, focusing on the engineering and design of radio transmissions to and from satellites. Chapter 4, "Microwave Link Engineering," gives the reader a basic understanding of the physics of the radio link between the Earth station and the satellite and covers the factors that are under the designer's control as well as those that are not. It is assumed that the reader has little or no technical training, so the only form of mathematics used is arithmetic. Chapter 5, "Modulation, Multiple Access and Impairments," rounds out the basic theory of communication as it relates to efficient satellite transmission. The chapter is fairly compact and may be helpful for interested readers to supplement their study with a basic textbook on communication engineering. Nontechnical readers can examine Chapters 4 and 5 but not delve deeply into the engineering details (which are more important to technical professionals, who need to understand features and trade-offs).

Chapters 6, 7, and 8 provide a comprehensive review of the functional elements of all communication satellites. The objective here is to aid readers in understanding the key issues in satellites and is not appropriate for the detailed engineering design of satellite components and subsystems. As is customary, the physical piece of hardware is referred to as the spacecraft, which becomes an artificial Earth satellite (or just satellite) when in orbit. The major elements of the spacecraft are the repeater (bent-pipe and digital processor), the antennas, and the spacecraft bus (the supporting vehicle), which are covered in Chapters 6, 7, and 8, respectively. While those chapters are not essential to understanding how to use satellite communications, they will be of general benefit because the actual operation of a spacecraft affects the performance of the services rendered.

The complementary topics for the ground facilities used in conjunction with the satellite are reviewed in Chapter 9, "Earth Stations and Terrestrial Technology." Care has been taken to include only current classes of Earth stations, particularly those used for satellite control, television broadcasting, fixed and mobile very small aperture terminal (VSAT) applications, and mobile satellite services (GEO and non-GEO). Chapter 9 will be useful for those readers who plan to use satellite transmission, since ground facilities fundamentally are under the control of the user rather than the satellite operator.

Chapter 10, "Launch Vehicles and Services," covers topics that are of great concern to operators and major users of satellites alike. The chapter is a complete review of the alternatives for placing satellites into Earth orbit and emphasizes that particular launch vehicle choices change over time. However, because reliability is based on a consistent experience record, much of the change is evolutionary rather than revolutionary. Chapter 10 also discusses the planning and operation of the mission, which is the sequence of events of launch and placement into operating orbit. Risk management is addressed as well.

Chapter 11, "Satellite Operations and Organization," addresses the special needs of this type of business. In some detail, it reviews the complete satellite control system, the communication network needed to support such a system, and the human resources that are appropriate to those functions. Chapter 12, "Economics of Satellite Systems," provides the underlying characteristics of satellites and Earth stations that are related to the cost of implementing and operating satellite networks. Our perspective is that of a commercial operator who is in

business to make money (or reduce costs). The framework is useful for analyzing the economics of either a complete system or a portion of a system (e.g., one or a few Earth stations). Chapter 12 is expanded to address the overall systems engineering process in satellite communication. This is based on the structured approach found in the aerospace industry adapted to the needs of commercial applications. From this, we move into the topic of engineering economics and overall development of the entire satellite program.

The book can be read sequentially from cover to cover because the material follows a consistent thread. All the elements and uses of spacecraft and Earth stations are covered. Chapters also can be read out of sequence, if necessary, because each chapter explains the concepts relevant to it. References to other chapters are provided throughout. The material is completely current as of the time of publication (circa 2008), but care has been taken to emphasize concepts that are not likely to change quickly. That is the same approach taken in the first edition, a book that remained in print (and in demand) for 12 years. Once read, this book can be used as a reference because most of the terminology in current usage is defined and illustrated. This book is a mate to the author's previous work, *The Satellite Communication Applications Handbook, Second Edition,* and *The Satellite Communication Ground Segment and Earth Station Handbook.*

Acknowledgments

Fundamentals of Satellite Systems

Satellite communications[1], no longer a marvel of human space activity, have evolved into an everyday, commonplace thing. By 2010, satellite systems will represent well over $100 billion of investment and provide an essential ingredient to many businesses and governments worldwide. Nearly all television coverage travels by satellite, today reaching directly to the home from space. Even in the age of wideband fiber optic cables and the Internet, satellites still serve the basic telecommunication needs of a majority of countries around the world. For example, domestic satellites have greatly improved the quality of service of the public telephone system and brought nations more tightly together. Satellites are adapting to developments in multimedia information and personal communication, with the cost of one-way and two-way Earth stations now within the reach of many potential users. A unique benefit has developed in the area of emergency preparedness and response. When a devastating earthquake hit Mexico City in September 1985, the newly launched Morelos 1 satellite maintained reliable television transmission around the nation even though all terrestrial long-distance lines out of the city fell silent. Similarly, communications were restored to facilitate disaster relief on the island of Sumatra after the December 2004 earthquake and subsequent tsunami.

To provide an overview of the field, the remainder of this chapter describes the features of satellite communication networks as well as the principal elements of an overall system. The brief history in Chapter 2 is intended to explain how the satellite industry developed and why things are the way they are. Chapter 2 can be read in sequence or deferred until other chapters have been read, to get a better feel for the technology and terms.

1.1 Basic Characteristics of Satellites

A communications satellite is a microwave repeater station that permits two or more users with appropriate earth stations to deliver or exchange information in various forms. As discussed later in this chapter and again in Chapter 6, a satellite in a geostationary Earth orbit (GEO) revolves around the Earth in the plane of the equator once in 24 hours, maintaining precise synchronization with the Earth's rotation. There are two other classes of 24-hour orbits: the geosynchronous orbit and the highly elliptical synchronous orbit. Both involve satellites that appear to

---

1. *Satellite communication* and *satellite communications* are used interchangeably throughout.

move relative to a fixed point on the Earth. It is well known that a system of three satellites in GEO each separated by 120 degrees of longitude, as shown in Figure 1.1, can receive and send radio signals over almost all the inhabited portions of the globe. (The small regions around the North and South Poles—above 81° NL and below 81° SL—are not covered.) A given GEO satellite has a coverage region, illustrated by the shaded oval, within which Earth stations can communicate with and be linked by the satellite. The range from user to satellite is a minimum of 36,000 km, which makes the design of the microwave link quite stringent in terms of providing adequate received signal power. Also, that distance introduces a propagation delay of about one-quarter of a second for a single hop between a pair of users.

The GEO is the ideal case of the entire class of geosynchronous (or synchronous) orbits, which all have a 24-hour period of revolution but are typically inclined with respect to the equator and/or elliptical in shape. As viewed from the Earth, a synchronous satellite in an inclined orbit appears to drift during a day above and below its normal position in the sky. While ideal, the circular GEO is not a stable arrangement, and inclination naturally increases in time. As discussed in Chapter 8, inclination is controlled by the use of an onboard propulsion system with enough fuel for corrections during the entire lifetime of the satellite. A synchronous satellite not intended for GEO operation can be launched with considerably less auxiliary fuel for that purpose. Orbit inclination of greater than 0.1 degree usually is not acceptable for commercial service unless the Earth station antennas can automatically repoint toward (track) the satellite as it appears to move. Mechanical tracking is the most practical (and cumbersome) approach, but electrical beam steering systems are available for specialized applications such as aeronautical mobile.

Orbits that are below a mean altitude of about 36,000 km have periods of revolution shorter than 24 hours and hence are termed non-GEO. As illustrated
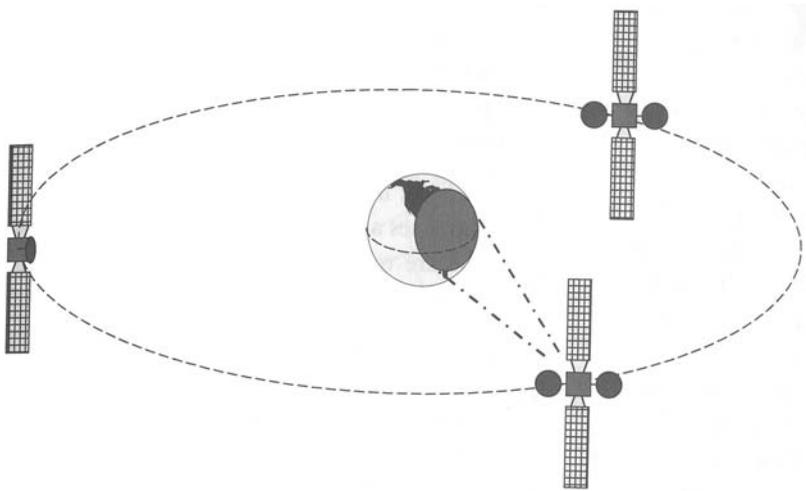
Figure 1.1   A system of three geostationary communication satellites provides nearly worldwide coverage.

in Figure 1.2, the Iridium system uses multiple satellites to provide continuous coverage of a given region of the Earth. That is simply because the satellites appear to move past a point on the Earth. The Iridium mobile satellite system employs low Earth orbit (LEO), in which satellites are at an altitude of approximately 780 km and each passes a given user in only a few minutes. In providing telephone services, users are relatively motionless compared to the satellite they are using. Hence, there is a need to hand off a telephone call while it is in progress. The advantage to using a non-GEO satellite network is that the range to the user is shorter; hence, less radiated power is required and the propagation delay is reduced as well. There is considerable complexity and delay in the processing of telephone calls and data communications due to satellite motion and handoffs.

The key dimension of a GEO satellite is its ability to provide coverage of an entire hemisphere at one time. As shown in Figure 1.3, a large contiguous land area (i.e., a country) as well as offshore locations can simultaneously access a single satellite. If the satellite has a specially designed communications beam focused on those areas, then any receiving antennas within the footprint of the beam (the area of coverage) receives precisely the same transmission. Locations well outside the footprint generally are not able to use the satellite effectively. The typical example in Asia is the JCSat 2a satellite (Figure 1.4), which has 24 C-band and 24 Ku-band transponders. *Transponder* is the term used in the industry to identify one complete microwave channel of transmission from a satellite.

Many satellite networks provide two-way (full-duplex) communications via the same coverage footprint. Terrestrial communications systems, including copper and fiber optic cable and point-to-point microwave radio, offer that capability between fixed points on the ground. Such systems will be around long into the future (as will satellites). Technology is always advancing, and satellite and terrestrial communications will improve in quality, capability, and economy. As shown in Figure 1.5, however, terrestrial systems must spread out over a land mass like a highway network to reach the points of access. They are most effective for connecting major locations, like cities, key suburbs, and industrial parks. The time, difficulty, and expense incurred are extensive, but once established, a terrestrial

·66 satellites
·6 orbits
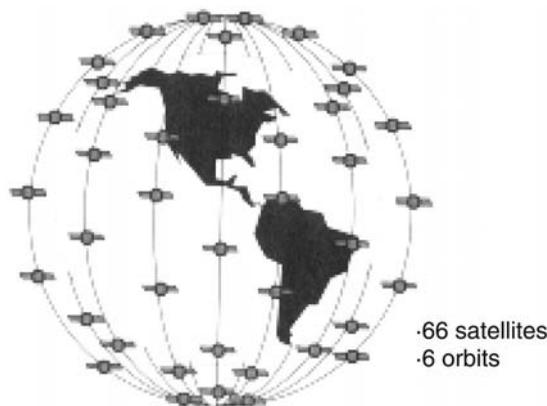
Figure 1.2   The non-GEO satellite constellation used by the Iridium system.
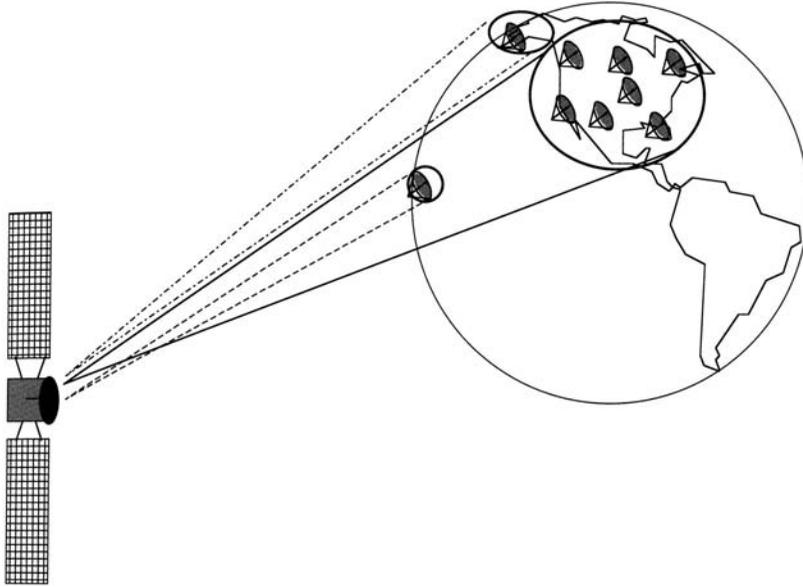
Figure 1.3 Typical footprint of a U.S. domestic communication satellite showing coverage of continental and offshore points.

infrastructure delivers very low unit service costs (e.g., per phone call or data packet) and can last a lifetime. Satellites, on the other hand, are designed to last only about 15 years in orbit, because of the practical inability to service a satellite in GEO and replenish consumables (fuel, battery cells, and degraded or failed components). Non-GEO satellites at altitudes below about 1,500 km are subject to atmospheric drag and a harsh radiation environment and are likely to require replacement after 10 years of operation.

Another way to view a satellite is as a bypass that can step over the existing terrestrial network and thus avoid the installation problems and service delays associated with initiating local telephone or cable TV service. Figure 1.6 depicts the three means of long-haul communications used to connect two user locations. Using the satellite in a duplex mode (i.e., allowing simultaneous two-way interactive communications), a user can employ Earth stations at each end, eliminating any connection with the terrestrial network. In a terrestrial microwave system, radio repeaters must be positioned at intermediate points along the route to maintain line-of-sight contact. It takes about 20 repeater sites to complete a microwave link of 1,000 km. That is because microwave energy, including that on terrestrial and satellite radio links, travels in a straight line with a minimum of bending over or around obstacles.

There are other wireless broadband communication systems which are referred to as nonline-of-sight (NLOS) wireless. The principle here is that a base station antenna located on top of a building, tower or mountain, can ''see'' into urban streets and potentially buildings. In reality, these NLOS systems are restricted to local or at most metropolitan-area coverage and are not suitable for long-haul transmission. On the other hand, NLOS wireless broadband can be combined with

Figure 1.4   The JCSat 2a satellite operated by SkyPerfect JSAT and built by Boeing. (*Source:* Sky Perfect JSAT.)

long-haul satellite links to address the need for broadband services in remote areas that cannot justify an investment in fiber cable or microwave radio.

In the case of a long-distance fiber optic cable system, a different form of repeater is needed to restore the digital bits that convey the information. For that reason, cable systems are probably the most costly to install and maintain. Only providers of local and long-distance telephone services and major users of communications services (e.g., government agencies, multinational corporations, railroads, and utilities) are able to justify the expense of operating their own terrestrial cable or microwave networks.

Satellite networks, on the other hand, are well within the reach of much smaller organizations because satellite capacity can be rented from a much larger company or agency. Approximately half of the current commercial GEO satellites are owned and operated by a few major companies that make the investment and sell the service to regional and global markets. As the original global operator and now the largest provider, Intelsat is able to serve users in nearly every country. Other large operators include SES, Eutelsat, Loral Skynet, Telesat, and JSAT. You can also find national operators throughout the world that focus on one or a few countries. Outside of the general provision of satellite capacity to users, there

Figure 1.5 An extensive network of cable and microwave links would be required to provide wide area coverage comparable to that of a satellite network.
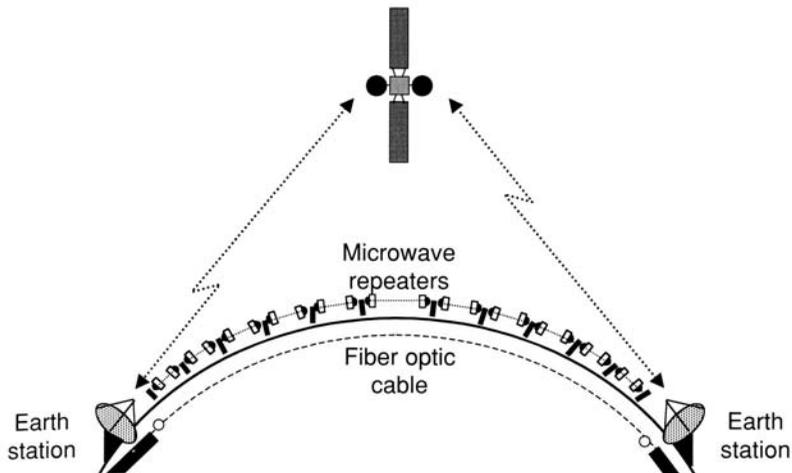
Figure 1.6 Terrestrial microwave and cable systems require multiple hops, while a satellite system provides the same capability with a single hop.

are some service providers who own their own satellites. Direct-to-home satellite broadcasters Echostar and DIRECTV are in this category, as are Inmarsat (mobile communications) and XM and Sirius Satellite Radio (mobile broadcasting). The availability of small, low-cost Earth stations that take advantage of reasonably priced satellite transponder capacity has allowed the smallest potential users to apply satellite bypass networks to achieve economies and save time. The idea of a satellite dish on every rooftop is now possible. The rest of this section reviews important aspects of satellites as related to their proper application.

1.1.1 Advantages of Satellite Communication

Satellites are used extensively for a variety of communication applications as a result of some well-recognized benefits. These derive from the basic physics of the system, the most important of which is that a satellite can "see" a substantial amount of geography at one time. Also, satellites employ microwave radio signals and thereby benefit from the freedom and mobility of wireless connections. Several advantages are interrelated, while others will become more important as the technology and applications evolve.

The point of these benefits is that satellite communication can represent a powerful medium when the developer of the system or service plays to its strengths. Satellite TV networks (which deliver programming directly to subscribers by satellite), national department store chains (which use VSAT networks to overcome the limitations of poor or fragmented terrestrial data communication networks), and ocean shipping lines (which demand reliable ship-to-shore communication) depend on satellites as their life's blood. If you can find that kind of connection through satellite technology, then you have a powerful bond on which to build and extend a business or strategic opportunity.

1.1.1.1 Mobile/Wireless Communication, Independent of Location

Any users with an appropriate Earth station can employ a satellite as long as they are within its footprint. A properly designed service establishes an effective radio link between the satellite and the user. If that is possible, the radio link is said to "close." The user can be stationary (*fixed* in the terminology) or in motion (i.e., *mobile*). As long as the link closes, the service will be satisfactory for the intended purpose. An example is the Hughes satellite modem, designed to use the Inmarsat 4 satellites (Figure 1.7). The phone is oriented so that the back of the modem is pointed toward the satellite position in the equatorial arc.

1.1.1.2 Wide Area Coverage: Country, Continent, or Globe

While governments control radio services within their borders, the coverage footprint of a satellite does not obey provincial, national, or continental boundaries. When designed properly, a satellite can serve any size region that can see it. As mentioned earlier, a GEO satellite can see about one-third of the Earth's surface. Satellites can be colocated at one orbit position to deliver greater quantities of communication channels, as illustrated for the SES Astra series of satellites (Figure 1.8). A non-GEO constellation like Iridium extends coverage beyond what one satellite sees through a web of intersatellite links.

1.1.1.3 Wide Bandwidth Available Throughout

Frequency spectrum availability for satellites is quite good, and satellite users have enjoyed ample bandwidth, principally for fixed services. Bandwidth is the measure of communication capacity in terms of hertz (Hz), either for the amount of radio spectrum used or for the input information that is delivered to the distant end of the link. Most of the satellites in GEO employ microwave frequencies generally

Figure 1.7   The Hughes satellite modem for use with the Inmarsat 4 GEO satellites. (*Source:* Hughes Communications, Inc.)

between 3.5 and 6.5 GHz (C-band) and between 10.5 and 14.5 GHz (Ku-band). It is common to also refer to the actual transfer rate of user data as bandwidth, although this is really measured in bits per second and not Hz. There is a direct relationship between the two that depends on the particular transmission technology, covered in detail in Chapter 5. That bandwidth can be made available to users across the footprint and can be multiplied through a variety of frequency-reuse techniques. (The bands are reviewed in Section 1.4.)

1.1.1.4   Independence From Terrestrial Infrastructure

By providing a repeater station in space, a satellite creates an independent microwave relay for ground-based radio stations. Installing Earth stations directly at the point of application allows users to communicate without external connections. That can be attractive in places where the terrestrial infrastructure is poor or expensive to install or employ. For example, small telephone Earth stations in China extend reliable voice and data services to remote western regions of the country.

1.1.1.5   Rapid Installation of Ground Networks

Once the satellite (or satellite constellation, in the case of a non-GEO system) is operational, individual Earth stations in the ground segment can be activated quickly in response to demand for services. Each Earth station can be installed and tested in a short time frame, depending on the degree of difficulty associated with the particular site. That is much simpler than for a terrestrial infrastructure, which requires an extensive ground construction program, including securing rights-of-
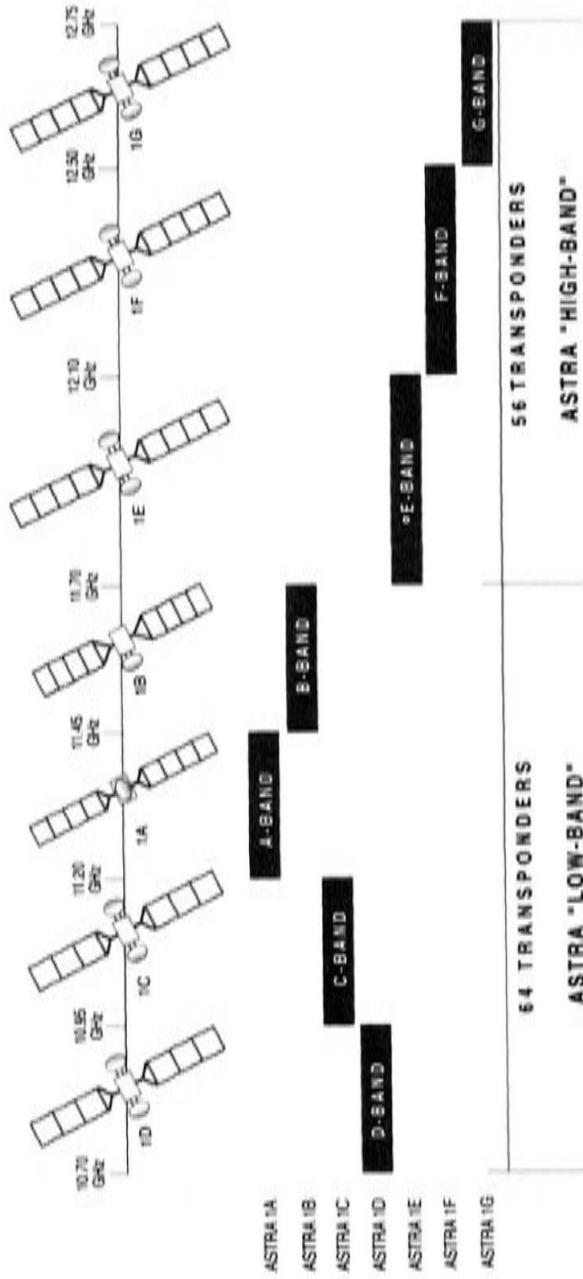
Figure 1.8 Television services are delivered throughout Europe on the SES Astra GEO satellites, operated at 19.2° WL by Société Européenne des Satéllites (satellite designations are for illustration only).

way along cable routes or for towers in the case of microwave or other terrestrial wireless systems. Maintenance of a ground-based infrastructure also is substantially more expensive and complex, due to the greater quantity of working elements and the opportunities for failure. Satellite communication plays a unique and vital role in restoring basic communications following a disaster. This was illustrated in the United States following the hurricanes of 2005. Installations like that shown in Figure 1.9 can be transported to a disaster site by truck, helicopter, or airplane, and erected in minutes. Normal broadband applications like high-speed Internet, telephone, and video are provided with this type of installation. Satellite communication is often the only way to deliver these services when the existing terrestrial infrastructure is destroyed.

1.1.1.6   Low Cost Per Added Site

Along with rapid installation, the cost of constructing a single site can be quite modest. Simple DTH receivers are consumer items that even homeowners can install. Earth stations for two-way communication may require more care in preparation for secure mounting of the antenna and selection of a suitable location for indoor equipment. Alternatively, hand-held mobile satellite terminals can be purchased for as little as $1,000 each. In general, the cost per site is low enough to keep the particular application within the budget of the particular user.

1.1.1.7   Uniform Service Characteristics

The footprint of the satellite defines the service area, and within that area services are delivered in precisely the same form. That overcomes some of the fragmentation

Figure 1.9   Highly transportable Ku-band Earth station for disaster recovery and temporary applications.

of service that would result from connecting network segments from various terrestrial telecommunication operators. An example is WalMart's satellite network in the United States that interconnects its stores located in small towns and villages to the Arkansas headquarters.

1.1.1.8 Total Service from a Single Provider

A satellite network can be operated by a single company or by a government agency. The customer, which also could be a nationwide organization, can then deal with a single entity to arrange service. A satellite-based service can be provided nationally or regionally, without regard to fragmentation imposed by the local market situation. An example is satellite radio from XM Satellite Radio (Figure 1.10).

1.1.2 Use of Microwave Frequencies

Satellites were not the first to employ microwave frequencies, which generally extend from about 1,000 MHz (1 GHz) to 30 GHz. That range was first developed during World War II for radar defense, then applied to various terrestrial communication systems between 1950 and 1980. As a result, the technology base existed at the time that satellite repeaters came into vogue. The microwave link generally is characterized by the following properties (explained later in detail):

- Line-of-sight propagation through space and the atmosphere;
- Blockage by dense media, like hills, tree trunks, solid buildings, metal walls, and at higher frequencies, heavy rain;
- Wide bandwidths, compared to the following lower frequency bands: high frequency (HF) shortwave radio, very high frequency (VHF) TV, frequency-modulation (FM) radio, and ultrahigh frequency (UHF) land mobile radio;

Figure 1.10 The XM Xpress RC receiver has the ability to store the most recent 60 minutes of listening. (*Source:* XM Satellite Radio.)

- Compact antennas typically using metal reflectors, which can focus energy in a desired direction (thus providing gain over an antenna with broad or omnidirectional radiation properties);
- Transmission through metal waveguide structures as opposed to wire conductors;
- Somewhat reduced efficiency of power amplification, which is the ratio of radio frequency (RF) power out divided by the direct current (dc) power in.

The analysis of a microwave communication link is carried out using the so-called link budget approach (discussed in Chapter 4). The behavior of the microwave link between a satellite and the ground is more predictable than any other type of radio transmission system. Engineers are able to design systems and services with a high degree of confidence, owing to the many years of experience with such systems and with the stable propagation medium at microwave frequencies. Transmission is not as reliable as for a high-quality fiber optic link; however, techniques like automatic power control, space and angle diversity, forward error correction (FEC), and dynamic modulation and bandwidth allocation tend to overcome the statistical nature of radio wave propagation, rendering a service that is attractive in a commercial sense.

1.1.3   Digital Transmission, Compression, and Routing

Digital communication and processing technologies have made a tremendous impact on satellites and their applications. From the time that satellites were introduced commercially in 1965 (discussed in Chapter 2), engineers foresaw the benefits of using digital approaches in lieu of transmitting information in analog form. Through developments in digital cellular systems, computing, image processing, and data communication, the foundation for greater exploitation of satellite networks exists today. Moore's law of semiconductor technology, which states that there is a doubling of the density of transistors on a chip (and hence the power of computing and capacity for data storage) every 18 months, propels all aspects of information-based high technology. Earth stations that once were housed in buildings were reduced to the size of a clock radio. Such "smarts" now are contained in hand-held units for voice communications.

Along with the hardware technology is a body of standards that defines how information can be digitized, compressed down to a small fraction of the number of bits previously demanded, and transferred in effective bundles for a complete end-user service. The most notable is the Motion Picture Experts Group (MPEG) series of standards geared toward television and multimedia. Other digital standards facilitate voice compression, two-way digital video teleconferencing, and highly interoperable information networks. Notable among terrestrial protocols are the Internet standard, Transmission Control Protocol/Internet Protocol (TCP/IP), and Asynchronous Transfer Mode (ATM). Satellite networks support those protocols that are the foundation of modern telecommunications.

1.1.4   Improved Space Platforms and Launching Systems

Satellites were once relatively small, while the Earth stations that employed them were correspondingly large. The first edition of this book focused on satellite designs that provided consistent performance, like the dual-spin configuration pioneered by Hughes and the three-axis designs first introduced in Europe by Aerospatiale and Matra and in the United States by RCA Astro and Ford Aerospace. In time, the three-axis design became the standard of the industry because of its greater capability to lift large communication payloads and to power them at a high level for 15 years or more of service. Those platforms now demonstrate service reliability that is nearly comparable to the previous generation of spinners, although with greater complexity in terms of operational procedure and software requirements.

We are now moving into new technologies to further enhance the potential to deliver services that are more useful and economic. Applications like GEO mobile services and two-way multimedia and digital audio broadcasting are pushing the platforms to go beyond what satellites flying in 1998 could deliver. Power levels up to 20 kW and beam pointing down to less than ±0.025 become the norm. An example of such a spacecraft is the Boeing 702, depicted in Figure 1.11. In contrast, the requirements of non-GEO systems have pushed spacecraft manufacturers to devise more cost-effective platforms. That is because of the vastly greater number
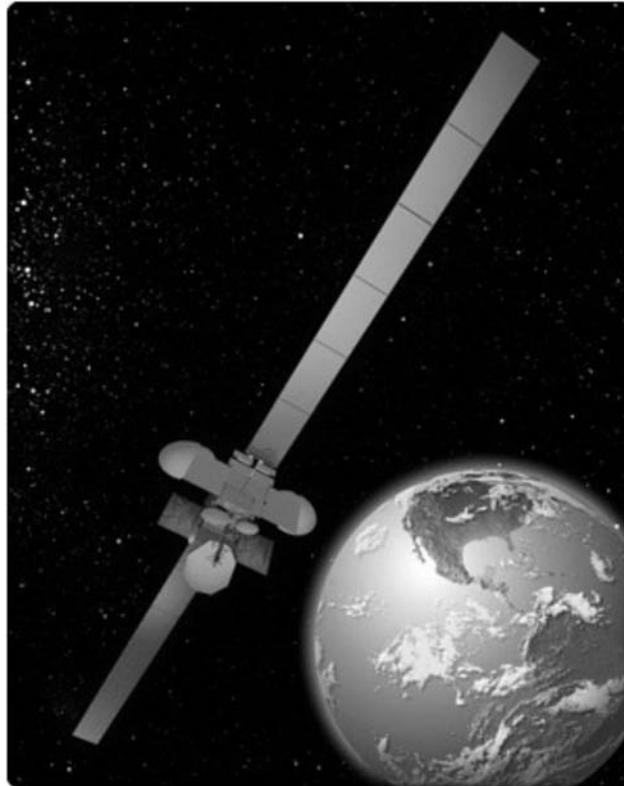
Figure 1.11   The Boeing 702 spacecraft is exemplified by Galaxy IIIC, with 87 transponders. (*Source:* Boeing.)

of satellites required coupled with the need for quicker replacement. Figure 1.12 illustrates an Iridium satellite from Motorola and Lockheed Martin, which is indicative of this class of spacecraft.

Launch vehicles have had to keep pace with the demands of spacecraft builders and buyers. As discussed in Chapter 10, launch vehicle agencies have extended their systems to lift the largest class of GEO satellite into transfer orbit. Multiple launch systems were introduced so that non-GEO constellations can be created economically. Competition is the other factor working to the benefit of satellite users, as new rocket systems and launch services companies enter the world market.

1.1.5 Integration with Terrestrial Wired and Wireless Networks

To a great degree, satellite networks are connected to and interoperate with primary terrestrial networks. When the first commercial satellites went into service, this integration found itself in the form of transferring telephone calls between the gateways of different countries. Thus, satellite voice networks had to work transparently with domestic and international telephone systems. Likewise, TV signals are handed off between terrestrial microwave and fiber systems to take advantage of the reach and broadcast capability of repeaters in space. More recently, Internet services and applications have been developed that move data seamlessly through terrestrial and satellite connections.

In coming years, satellites could be integrated with terrestrial wireless services to deliver a single service to a variety of user terminal devices. This is a natural outgrowth of the move to third generation cellular (3G) technology that provides broadband data over a wide area. New systems will take advantage of the fact that the satellite spectrum used for mobile satellite service (MSS) is within the range of frequencies common to 2G and 3G cellular technologies.
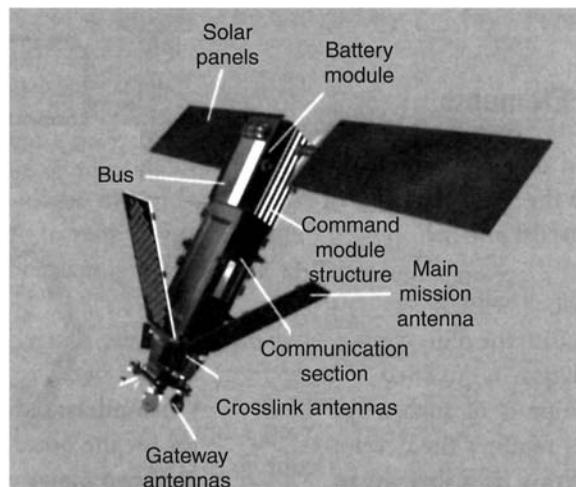
Figure 1.12 The Iridium LEO spacecraft, manufactured by Motorola, Lockheed Martin, and Raytheon. (*Source:* Lockheed Martin.)

1.2   System Elements

Implementation of a communication satellite system is a major undertaking. Fortunately for the great majority of current and future satellite users, several systems already exist and are open for business in every part of the world. There are more than 250 operating geostationary satellites offering satellite capacity globally. Users can lease capacity on domestic satellites and from regional and international operators as well.

In some cases, the requirements may be so extensive that a dedicated satellite or constellation is justified. The purpose of this book is to define and describe all the parts of such a system because an understanding of satellite communication requires the development of a feel for the breadth of the technology. This section describes the system in terms of two major parts, the space segment and the ground segment.

1.2.1   Space Segment

Figure 1.13 is an overview of the main elements of the space segment. As can be seen from the figure, placing a satellite into orbit and operating it for 12 or more years involves a great deal. Placement in orbit is accomplished by contracting with a spacecraft manufacturer and a launch agency and allowing them about three years to design, construct, and launch the satellite. Chapters 6, 7, and 8 provide
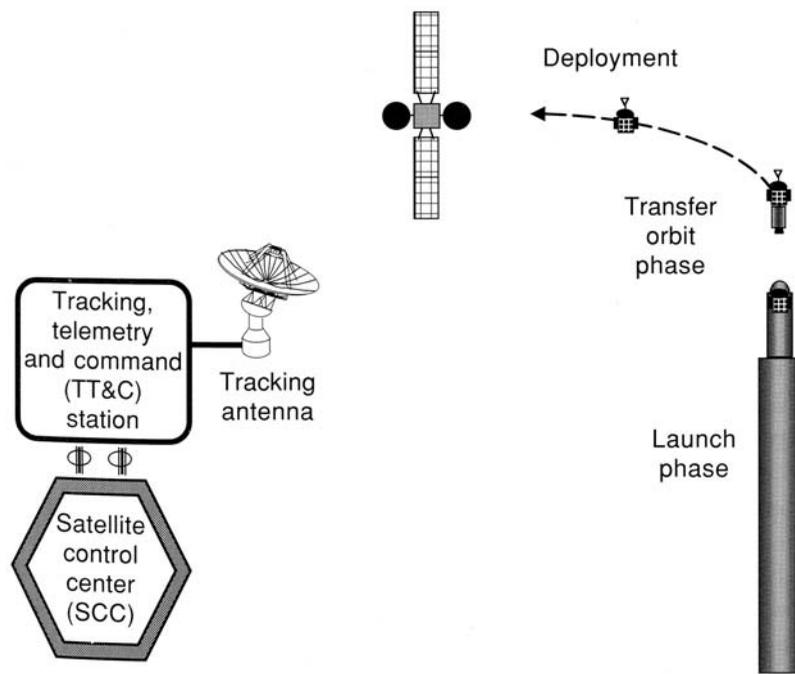
Figure 1.13   The elements needed to implement the space segment of a communication satellite system.

considerable detail on the design of a communication satellite, while Chapter 10 reviews the launch aspects.

After the spacecraft is placed in the proper orbit, it becomes the responsibility of a satellite operator to control the satellite for the duration of its mission (its lifetime in orbit). As discussed in Chapter 11, that is a fairly complex task and involves both sophisticated ground-based facilities as well as highly trained technical personnel. The complexity of the problem increases for a non-GEO satellite and a constellation because there are many more satellites that are in constant motion relative to the surface of the Earth. The tracking, telemetry, and command (TT&C) station (or stations) establishes a control and monitoring link with the satellite. Precise tracking data are collected periodically via the ground antenna to allow the pinpointing of the satellite's position and the planning of on-orbit position corrections. That is because any orbit tends to distort and shift with respect to a fixed point in space due to varying gravitational forces from the nonspherical Earth and the pull of the sun and moon.

Eventually, the onboard liquid or molecular propellant used for orbit correction is exhausted and the satellite must be removed from service. Retirement of a GEO satellite usually involves using a small reserve of propellant to increase the velocity sufficiently to raise the orbit a few hundred kilometers. The retired satellite will remain in orbit for eternity; with its repeater turned off, it cannot interfere or conflict with the operation of usable satellites in GEO. Satellites in LEO are subject to atmospheric drag, which, if not corrected, causes the orbit to decay and the satellite to reenter the atmosphere and burn up. This is, in fact, how LEO satellites may be removed from service.

The second facility shown in Figure 1.13 is the satellite control center (SCC), which houses the operator consoles and data processing equipment for use by satellite control personnel and supporting engineering staff. Other common names for an SCC include operations control center (OCC) and satellite control facility (SCF). The SCC could be at the site of the TT&C station, but more commonly it is located some distance away, usually at the headquarters of the satellite operator. The actual satellite-related data can be passed between the sites over data and voice lines (either terrestrial or satellite). Adding a second TT&C site improves system reliability and would be required for non-GEO satellite operations. As mentioned previously, the level of training of the personnel at the SCC is quite high, involving such fields as orbital mechanics, aerospace and electronic engineering, and computer system operation.

In the launch phase, the satellite must be located and tracked from the ground as soon as it has been released from the last rocket stage. That service normally is obtained from the spacecraft manufacturer. The satellite operator could employ its own TT&C station to participate in transfer orbit maneuvers (actions prior to placement in GEO), and the station can be the point from which commands are transmitted to the satellite to deploy its antennas and solar panels. Some non-GEO systems can be launched directly into the operational orbit, avoiding the transfer orbit phase entirely. Touch-up maneuvers to fine-tune the orbit would be anticipated for GEO and non-GEO satellites alike.

Routine operations at the SCC and the TT&C station are intended to produce continuous and nearly uniform performance from each satellite. Actual communica-

tions services via the microwave repeater aboard the satellite do not need to pass through the satellite operator's ground facilities, although any suitably located TT&C station could serve as a communications hub for a large city. Often, the TT&C station includes separate communications equipment to access the satellite repeater for the purposes of testing and monitoring its performance.

One particularly nice feature of a GEO satellite is that the communications monitoring function can be performed from anywhere within the footprint (see Figure 1.3); for example, the SCC can have its own independent monitoring antenna not connected with the TT&C station. GEO and non-GEO satellites can be monitored by several ground antennas strategically positioned around the coverage region. Such antennas often are vital for measuring satellite repeater output and for troubleshooting complaints and problems (users may blame the satellite for poor or no reception when the difficulty lies with their Earth stations).

Another problem area for which monitoring is essential is dealing with harmful interference to communications services. Also called double illumination, it occurs when an errant Earth station operator activates a transmitter on the wrong frequency or even on the wrong satellite. In the congested orbital arcs serving North America, Europe, and eastern Asia, double illumination must be detected quickly and the source identified so disruption of valid communications services is minimized. While the vast majority of interference problems are unintentional, there have been a few instances of intentional jamming and possible piracy (e.g., unauthorized usage of satellite capacity). Satellite operators and users are motivated to act quickly on those rare instances. Such problems, while serious when they occur, are relatively infrequent; satellite transmission is perhaps the most reliable means of long-haul communications.

1.2.2   Ground Segment

The ground segment provides access to the satellite repeater from Earth stations to meet communications needs of users (television viewers, information network providers, enterprises, disaster workers, and Web surfers). A typical ground segment is illustrated in Figure 1.14; a single satellite is shown to indicate that the links are established through its repeater rather than directly from Earth station to Earth station. Incidentally, *Earth station* is an internationally accepted term that includes satellite communication stations located on the ground, in the air (on airplanes), or on the sea (on ships). Many commercial applications are through Earth stations at fixed locations on the ground; thus, the international designation for such an arrangement is fixed satellite service (FSS). Related to FSS is broadcasting satellite service (BSS), targeted to individual home reception of a variety of broadcast information (e.g., TV, radio, and data). Mobile satellite service (MSS) has been in operation for some time, offering interactive voice and data services for ships, aircraft, and individuals. Fixed Earth stations have experienced a tremendous reduction in size at the same time that satellites have grown in size and power capability. Earth stations have evolved from the first international FSS behemoths with 30m antennas (Figure 1.15) to the inexpensive BSS systems, with their half-meter fixed dishes (Figure 1.16). For the new generation of MSS Earth stations, the requirements are being met with handheld devices about the size of a PDA.
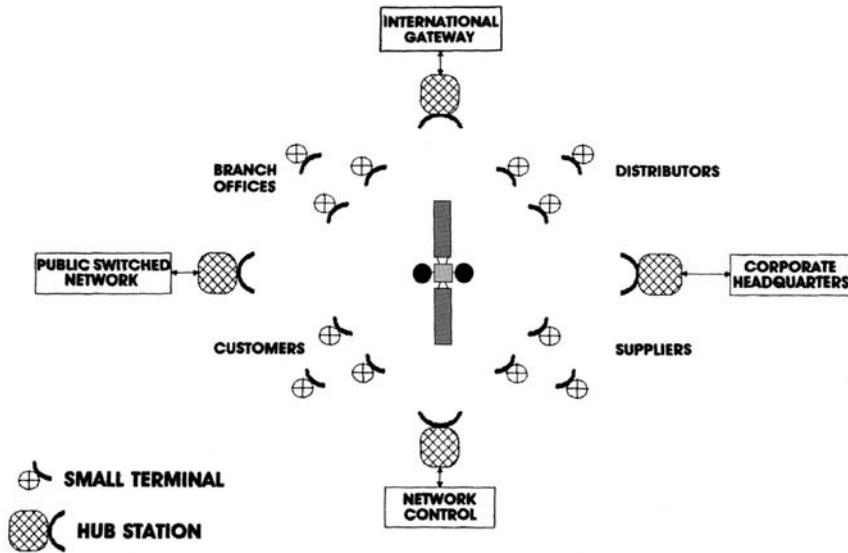
Figure 1.14   The ground segment of a satellite network providing two-way interactive services to a variety of locations.

Figure 1.15   This INTELSAT Earth station antenna in Hong Kong was built at 30m to be able to operate with low-power C-band satellites such as INTELSAT III.

The majority of user fixed Earth stations (see Figure 1.17) are customer-premise equipment that the user can employ directly with a personal computer (PC), a telephone, or other terminal device. They are inexpensive enough to justify their placement at points of low traffic demand such as small towns, branch offices, or suppliers. Figure 1.17 shows an example of a very small aperture terminal (VSAT), which is a compact and inexpensive Earth station intended for that purpose.

The ground segment, therefore, is not a single, homogeneous entity but rather a diverse collection of facilities, users, and applications. It is constantly changing and evolving, providing service when and where needed. Some networks, such as

Figure 1.16   The DSS equipment offered by Sony Electronics Corporation consists of a 45-cm satellite antenna, a decoder box, and a remote control. (*Source:* Photograph courtesy of DIRECTV, Inc.)

DW7700 Satellite Router

Figure 1.17   An example of a small customer-premise VSAT used in two-way data communications. (*Source:* Hughes Communications, Inc.)

those used for BSS or MSS applications, may be homogeneous in the sense that a common air interface (CAI) is used for all links over the satellite. Satellites used in an FSS, on the other hand, can support a variety of applications within the common bandwidth of several megahertz. The CAI concept still might apply to a subnetwork of VSATs used by a large bank in China or for telephone services to rural villages in Brazil.

One of the four large stations shown in Figure 1.14 is a control point for a portion of the network or perhaps even the entire network. Implemented as an automated network managment center, the control Earth station is critical in

networks in which traffic flows are dynamic (i.e., changing with time) and in which remote sites are unmanned. Using modern network management workstations and remote microprocessor technology, the network management center can measure and maintain service quality by remote control. Remote control is a particular feature of a satellite network because the satellite provides a common relay point for control messages and other management data. Realizing that type of management integration is a much more difficult problem in large terrestrial telecommunication networks.

1.2.3   Overall System

Providing a telecommunication service via satellite involves more than launching a satellite and installing Earth stations. The process of converting the idea of a particular service, like TV broadcasting or Internet service to airplanes, is one of engineering a complete system. In systems engineering (detailed in Chapter 12), the planners begin with the objectives in mind, identify alternative approaches to meeting this object, and then optimize the design of the system that ultimately meets those objectives.

Some of the original satellite communication applications were engineered using this process. For example, FSS satellites used to connect major cities across oceans needed to meet critical constraints that resulted from what could be placed in orbit with current rockets and provide an adequate relay between currently available Earth station antennas. The process of systems engineering produced the Intelsat 4 satellite with 12 transponders at 36-MHz bandwidth each. In a more detailed manner, the first BSS satellite system was optimized through systems engineering by trading off the cost of the user terminal (which had to be affordable and still attractive to a large consumer base) against the cost of the broadcast center where hundreds of TV channels would be processed and delivered through the satellites.

A number of satellite communications systems are discussed in Chapter 12 as a means to illustrate how systems engineering is applied to the creation of optimized space and ground segments. The important point is that the system must be evaluated as one overall entity rather than focusing attention on either making the best satellite or the smallest possible user device.

1.3   Satellite Orbit Configurations

Having already introduced the concept of GEO and non-GEO satellites, we can now examine some of the technical particulars. Behavior of a satellite in Earth orbit follows Kepler's laws of planetary motion (Figure 1.18), which can be restated for artificial satellites as follows:

1. The orbit of each satellite is an ellipse with the Earth centered at one focus.
2. The line joining the satellite and the center of the Earth sweeps out equal areas in equal times. The area is bounded by the arc segment of the orbit and two lines that extend from the center of the Earth.

$$P = 1.659 \cdot 10^{-4}(6378+h)^{3/2}$$

(P in minutes)
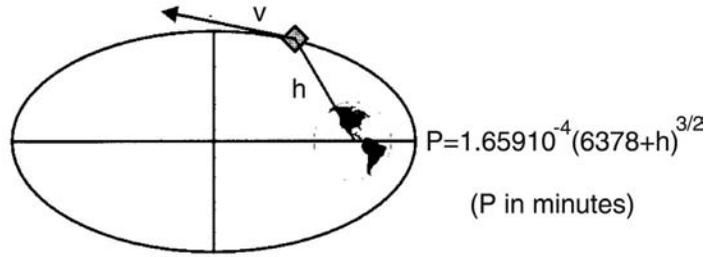
Figure 1.18   Illustration of Kepler's laws of planetary motion.

3. The square of the period of a satellite is proportional to the cube of its mean distance from the center of the Earth. That relationship is plotted in Figure 1.19 to give an idea of the relationship between altitude and orbital period.

Figure 1.20 provides three classical Earth-orbit schemes and their basic properties. LEO systems employ satellites at altitudes ranging from 500 to 1,000 km. Over that range, the orbit period is between 1.6 and 1.8 hours, the higher orbit resulting in a slightly longer period of revolution. The reason for that small change in period is that it is the distance from the center of the Earth that determines the period, not the elevation above sea level. The altitude of a medium Earth orbit (MEO) is around 10,000 km (a period of about 6 hours). Between 2,000 and 8,000 km, there is an inhospitable environment for electronic components produced by the Van Allen radiation belt.

The principal advantage of LEO satellites is the shorter range that the radio signal has to traverse, requiring less power and minimizing propagation delay. The range of delay for a single up-down hop is shown at the top of the right-hand graph in Figure 1.19. Their short orbital period produces relatively brief durations when a given satellite can serve a particular user. For altitudes in the range of 8,000 to 10,000 km, a MEO satellite has a much longer period and thus tends to
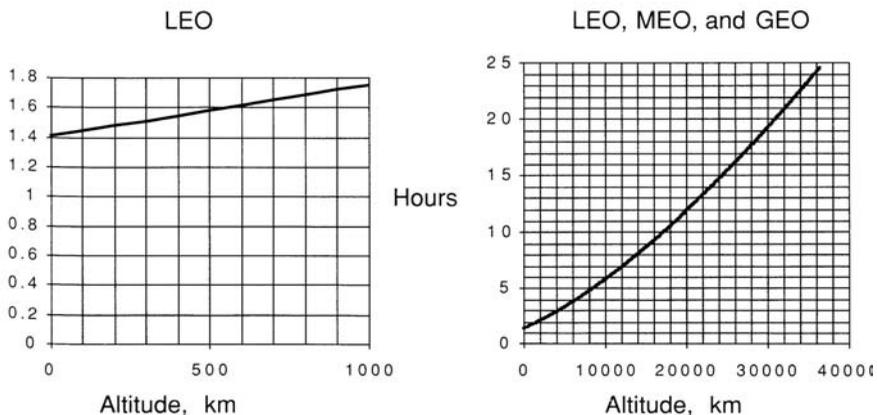
Figure 1.19   Orbit period (in hours) versus altitude, based on Kepler's third law.
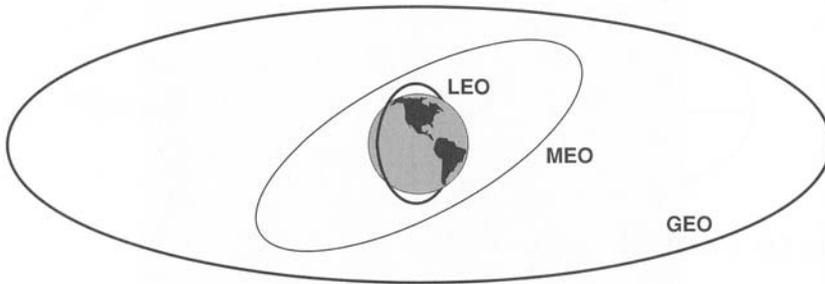
Figure 1.20 The three most popular orbits for communication satellites are LEO, MEO, and GEO. The respective altitude ranges are 500 km to 1,000 km for LEO; 5,000 km to 12,000 km for MEO; and 36,000 for GEO. Only one orbit per altitude is illustrated, even though there is a requirement for multiple orbits for LEO and MEO satellites to provide continuous service.

"hang" over a given region on the Earth for a few hours. Transmission distance and propagation delay are greater than for LEO but still significantly less than for GEO. In the case of the latter, there are really two classes of orbits that have a 24-hour period. A geosynchronous orbit could be elliptical or inclined with respect to the equator (or both). The special case of an equatorial 24-hour circular orbit, in which the satellite appears to remain over a point on the ground (which is on the equator at the same longitude where the satellite is maintained), is called GEO. A 24-hour circular geosynchronous orbit that is inclined with respect to the equator is not GEO because the satellite appears to move relative to the fixed point on the Earth. A GEO satellite would not require ground antennas that track the satellite, while an inclined geosynchronous orbit satellite might. Other non-GEO orbits have been used at various times, such as the highly elliptical Earth orbit (HEO) to allow coverage of northern latitudes.

1.4 Frequency Spectrum Allocations

Satellite communications employ electromagnetic waves to carry information between ground and space. The frequency of an electromagnetic wave is the rate of reversal of its polarity (not polarization) in cycles per second, defined to be units of hertz (Hz). Alternating current in a copper wire also has that frequency property; if the frequency is sufficiently high, the wire becomes an antenna, radiating electromagnetic energy at the same frequency. Recall that wavelength is inversely proportional to frequency, with the proportionality constant being the speed of light (approximately 300 million meters per second in a vacuum).

A particular range of frequencies is called a frequency band, while the full extent of all frequencies for zero to infinity is called the electromagnetic spectrum. In particular, the radio frequency (RF) part of the electromagnetic spectrum permits the efficient generation of signal power, its radiation into free space, and reception at a distant point. The most useful RF frequencies lie in the 300-MHz and 300,000-MHz range, although lower frequencies (longer wavelengths) are attractive for certain applications. While frequencies are all expressed either in hertz or meters

of wavelength, they are not created equal. Some segments of the band can be transmitted with greater ease (e.g., can propagate through the atmosphere and physical media with less loss and temporary fading) than others. Also, differing levels of natural and man-made noise interfere with the transfer. Figure 1.21 indicates the absorbtion by air at various frequencies. We see that above about 30 GHz, propagation grows in difficulty due to higher levels of atmospheric attenuation effects. Not shown in the figure is that manmade noise becomes a significant detriment below about 1 GHz. The absorption is measured in decibels and can be adjusted for Earth elevation angle with the formula indicated in Figure 1.21.

The optimum piece of the spectrum for space-to-Earth applications lies between about 1 and 4 GHz, which we call the *noise window* for microwave transmission. Above about 12 GHz, there is a significant amount of signal absorption by the atmosphere. There is a point of near total absorption at 66 GHz, due to resonance by oxygen. Those frequencies cannot be used for Earth-to-space paths but can be applied for intersatellite links.

The most detrimental atmospheric effect above 4 GHz is rain attenuation, which is a decrease of signal level due to absorption of microwave energy by water droplets in a rainstorm. Due to the relationship between the size of droplets relative to the wavelength of the radio signal, microwave energy at higher frequencies is absorbed more heavily than that at lower frequencies. (Sections 1.4.4 and 1.4.5 provide more background.) Rain attenuation is a serious problem in tropical regions of the world with heavy thunderstorm activity, as those storms contain intense rain cells.

An RF signal on one frequency is called a carrier, and the actual information that it carries (voice, video, or data) is called modulation. The types of modulation available are either analog or digital in nature. Our focus is on digital modula-
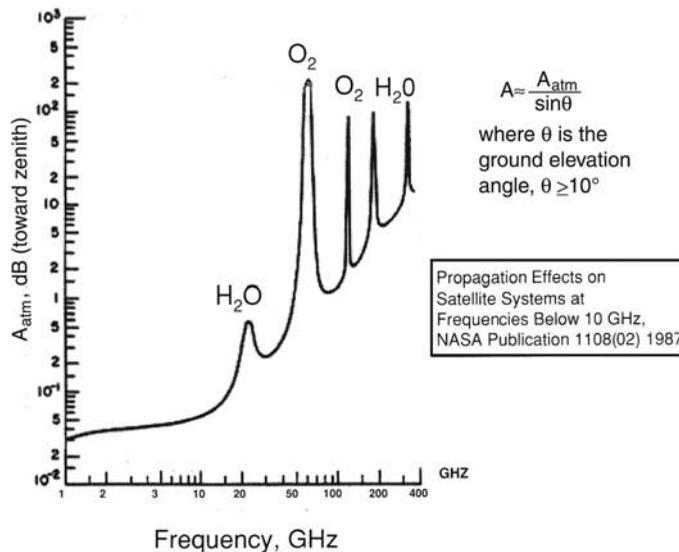
Figure 1.21   Atmospheric attenuation effects for space-to-Earth paths as a function of frequency (clear air conditions); A_{atm} applies for an Earth station located on the equator at the same longitude as the GEO satellite.

tion since this is how most forms of data are transmitted (an exception being some legacy frequency modulated cable TV channels). Once the carrier is modulated, it develops sidebands that contain a representation of the original information. The modulated carrier therefore occupies an amount of RF bandwidth within the frequency band of interest. For a given modulation method, the bandwidth is proportional to the information bit rate. If two carriers are either on the same frequency or have overlapping bandwidths, then radio frequency interference (RFI) may occur. To the user, RFI can look or sound like background noise (which is neither intelligible nor particularly distressful), or it can produce an annoying effect like herringbone patterns on a TV monitor or a buzzing sound or tone in the audio. In digital modulation, interference and noise introduce bit errors that either degrade quality or produce information dropouts. When the interfering carrier is comparable in power level to the desired (wanted) carrier, the interference effect is classed as harmful, a condition similar to the radio jamming encountered in the shortwave broadcast band.

1.4.1 ITU Spectrum Allocations and Regions

Frequency bands are allocated for various purposes by the International Telecommunication Union (ITU), a specialized agency of the United Nations that has its headquarters in Geneva. Members of the ITU include essentially every government on the planet, which in turn are responsible for assigning radio frequencies in allocated bands to domestic users. Because the RF spectrum is a limited resource, the ITU has allocated the same parts of the spectrum to many countries and for many purposes around the globe. The consequence is that users of radio communications must allow for specified amounts of RFI and must be prepared to deal with harmful interference if and when it occurs. When there are disputes between countries over RFI or frequency assignments, the ITU often plays the role of mediator or judge. It cannot, under present treaties, directly enforce its own rules. The rules and the frequency allocations are subject to review and modification at the biannual World Radiocommunication Conference (WRC), conducted by the ITU and all its members. The spelling of the word *radiocommunication* was contrived by the ITU itself to reflect the vital connection between the *radio* spectrum (a limited resource that the members must learn to share) and *communication* (perhaps the most valuable application of that spectrum).

The spectrum of RF frequencies is depicted in Figure 1.22, which indicates on a logarithmic scale the abbreviations that are in common usage. The bottom end of the spectrum from 0.1 to 100 MHz has been applied to the various radio broadcasting services and is not used for space communication. In fact, the electrically charged layer of the atmosphere referred to as the ionosphere is opaque at these frequencies. The frequency bands of interest for satellites lie above 100 MHz, where we find the VHF, UHF, and super high frequency (SHF) bands. The SHF range has been broken down further by common usage into subbands with letter designations, the familiar L-, S-, C-, X-, Ku-, and Ka-bands being included. Generally, Ku-band and those below it are the most popular because of the relative low cost of available equipment and the more favorable propagation characteristics. The Ka-, V-, and Q-bands employ millimeter wavelengths and are potentially useful
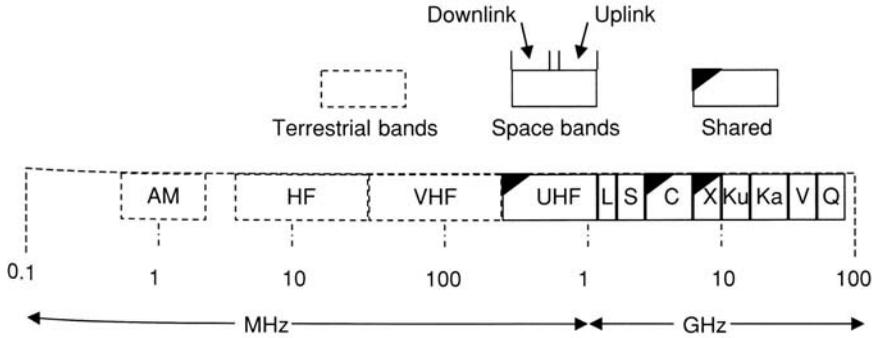
Figure 1.22   The radio frequency spectrum identifying commonly used frequency bands and their designations; high frequency (HF), very high frequency (VHF), and ultra high frequency (UHF).

for very high bandwidth transmissions into small receiving antennas. There are difficulties in applying the millimeter-wave bands because of increased rain attenuation and hardware expense. The letter designations, in general, first were used during World War II as classified code words for the microwave bands applied for radar and other military or government purposes. Today, they are simply shorthand names for the more popular satellite bands, all lying in the range of 1 GHz (1,000 MHz) to 60 GHz.

A typical satellite band is divided into separate halves, one for ground-to-space links (the uplink) and one for space-to-ground links (the downlink). That separation is reflected in the design of the satellite microwave repeater to minimize the chance of downlink signals being re-received and thereby jamming the operation of the satellite. The Iridium satellite system (discussed in [1]) transmits and receives on the same frequency but not at the same time.

Uplink frequency bands allocated by the ITU are typically slightly above the corresponding downlink frequency band, to take advantage of the fact that it is easier to generate RF power within an Earth station than it is onboard a satellite. It is a natural characteristic of the types of RF power amplifiers used in both locations that the efficiency of conversion from alternating current (ac) power into RF power tends to decrease as frequency increases. Along with that, the output from the Earth station power amplifier usually is greater than that of the satellite by a factor of between 10 and 100. More recently, satellite systems that make extensive use of VSATs and mobile terminals allow less uplink power, so the cost of the Earth station can be minimized.

An important consideration in the use of microwave frequencies for satellite communication is the matter of coexistence. Figure 1.22 indicates that many of the satellite bands are shared, which means that the same frequencies are used by terrestrial microwave links. Parts of the Ku- and Ka-bands, on the other hand, are not shared with terrestrial links, so only satellite links are permitted. When sharing is invoked by the ITU Radio Regulations (the so-called Red Books), the users of different systems must take standard precautions to reduce the possibility for RFI. That involves keeping the radiated power below a prescribed threshold and locating

transmitting and receiving antennas so as to minimize the amount of interference caused. A band that is not shared, therefore, is particularly valuable to satellite communications, since terrestrial microwave systems can be totally ignored. More recently, L- and S-band spectrum has been authorized for satellite and terrestrial use by the same operator, a concept called ancillary terrestrial component (ATC), or alternatively, complementary ground segment (CGS).

Even if the band is not shared, we still have to allow for multiple satellites and Earth station operators, who could interfere with one another. We employ the process of frequency coordination in cases where a new satellite or Earth station operator could cause unacceptable or harmful interference to an existing operator. Frequency coordination is necessary to control interference among satellite systems that use the same frequency band and operate in adjacent (or even nearly adjacent) orbit positions. The procedures for conducting frequency coordination are developed and administered through the ITU.

The ITU has divided the world up into three regions: Region 1 is Europe and Africa (including all of Russia); Region 2 is North America and South America; and Region 3 is Asia and the Pacific. The regionalization, illustrated in Figure 1.23, was somewhat arbitrary, and radio signals do not obey such boundaries. The regions reflect a kind of political/economic subdivision of the world for the purposes of allocating spectrum to ITU members. At WRC meetings, all regions are combined into one, while there can still be Regional Radiocommunication Conferences (RRCs) that are attended by members of a single region. Such RRCs are rare, however.

The rest of this section reviews the major segments of the satellite allocations and identifies the particular ranges and some of the key features and applications.

1.4.2 VHF and UHF Frequency Ranges

The VHF and UHF bands are defined to be in the ranges 30 to 300 MHz and 300 to 3,000 MHz, respectively. For the purposes of satellite communication, we are
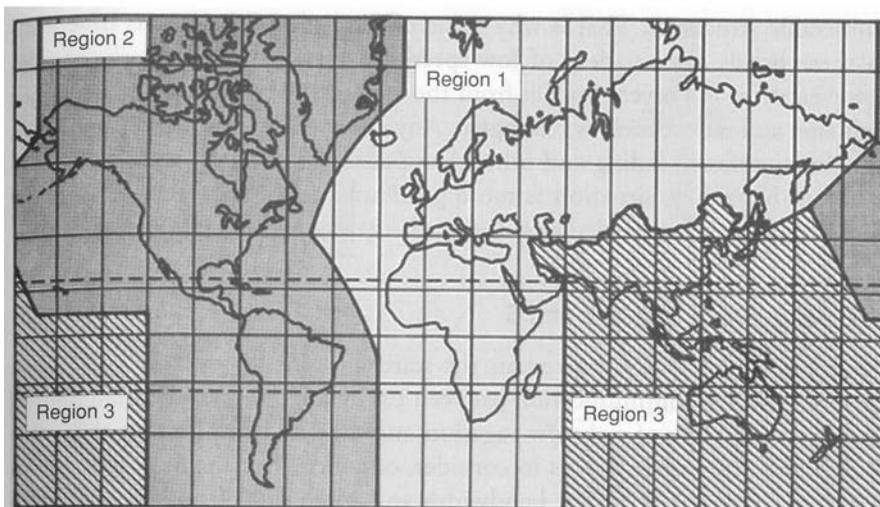
Figure 1.23   The ITU Regions (ITU radio regulations).

concerned with application of the band between about 400 MHz and 1,000 MHz. That range already is highly contested for terrestrial wireless applications, notably cellular; mobile data; and various radio systems used for government and emergency services. In the past, this range was used by the early scientific satellites and for telemetry and command purposes. More recently, some of the small non-GEO systems (referred to as "little LEOs") adopted this range because of the case of generating power and the convenience of using simple antennas on the satellites and user terminals. The amount of bandwidth available is limited simply because of the low range of frequencies. The Western military have exploited UHF on a global basis for maritime and tactical services.

While the satellite hardware and user terminal design are relatively straightforward, the propagation at these frequencies certainly is not. There are advantages to using longer wavelengths to bend around obstacles and to penetrate nonmetallic structures. That is why some of the little LEO systems have chosen these bands for a variety of low-speed data services, particularly paging. However, there is a severe impact from the ionosphere, which can twist, bend, attenuate, and reflect these wavelengths. Any system that uses VHF/UHF must allow for significant fading and other disruption of the transmission, often on a random basis. The situation is more pronounced in the tropics around the geomagnetic equator, particularly during the spring and fall equinox seasons.

1.4.3 Microwave Bands: L and S

The frequency of 1 GHz represents the starting point for microwave applications that involve communication between Earth stations and satellites. Signals travel nearly by line-of-sight propagation and are less hampered by the ionosphere. There are other factors to consider, of course, but the basic fact is that these frequencies have greater bandwidth and more stable propagation under most conditions than do frequencies below 1 GHz. We first explore the range up to about 3 GHz, which includes L- and S-bands.

With available spectrum much greater than VHF/UHF but less than C- or Ku-bands, the L- and S-bands are particularly effective for providing rapid communications by way of mobile and transportable Earth stations. Two-way voice and data service can be provided through a ground antenna as simple as a small Yagi (TV-type antenna), a wire helix, or even a quarter-wavelength rod (which would be about 4 cm long). The use of relatively high satellite transmit power for each individual channel of communication (voice or data) also helps to reduce the size and cost of the receiving terminal. The trade-off is in the number of satellites per hemisphere and the quantity of voice channels per satellite: instead of being measured in the hundreds of thousands for one C- and Ku-band satellite, capacity of such lower frequency satellites ranges from a thousand to tens of thousands of channels. That can be adequate for high-value mobile telephone or data communications or one-way mobile broadcasting. A number of S-band satellites are being used to broadcast audio to vehicles, with as many as 100 different channels of music, talk radio, and driver information. ST Telecom, a Korean wireless operator, offers mobile TV to handsets like that shown in Figure 1.24. Both the mobile radio and handheld TV versions rely on a combination of satellites and terrestrial

Figure 1.24   MobileTV from Samsung and TU Media of Korea. (*Source:* Samsung.)

rebroadcast transmitters. In Korea, these transmitters are even located on subway trains so commuters can watch their favorite TV dramas while riding to and from work.

The use of such simple antennas on the ground, taking advantage of high power per channel in the satellite, restricts the number of satellites that can operate at the same time within the same footprint of coverage. That applies to GEO and non-GEO satellite systems alike. Figure 1.25 illustrates how an Earth station's antenna beam influences satellite orbit spacing. Satellite A is at the center (peak) of the oval-shaped antenna beam emanating from an Earth station. Operation at a relatively low frequency causes that beam to be broad enough to interfere with the operation of a closely spaced satellite at position B. This figure illustrates a transmitting earth station as the principle applies to reception as well. The chords drawn within the beam show that the strength of the antenna beam (its gain) in the direction of A is greater than in the direction of B, indicating that RFI probably would be harmful (although not the maximum possible). A common way to control RFI under that condition would be to segment the frequency band by using nonoverlapping bandwidths for the RF carriers of the respective satellite systems. Alternatively, the same bandwidth can be employed using code division multiple access (CDMA), discussed in Chapter 5.

For the satellite at position C, the main beam of the ground antenna provides no measurable performance, and proper isolation is achieved. That means satellite C can operate on the same frequencies as satellite A without causing or receiving harmful RFI. Compare that to the situation illustrated in Figure 1.26 for a higher frequency like C- or Ku-band. Automatically, the ground antenna beamwidth is
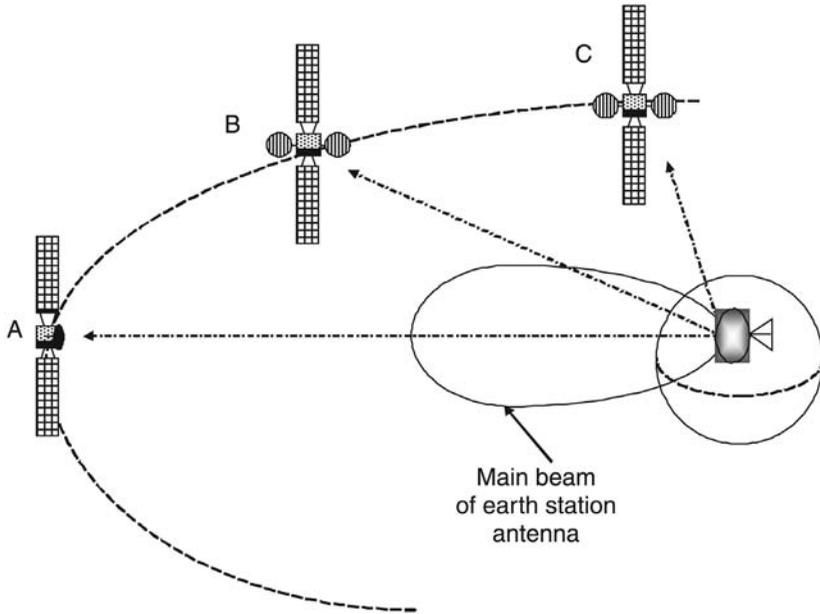
Figure 1.25 Effect of wide-beam Earth station antennas on GEO arc utilization.
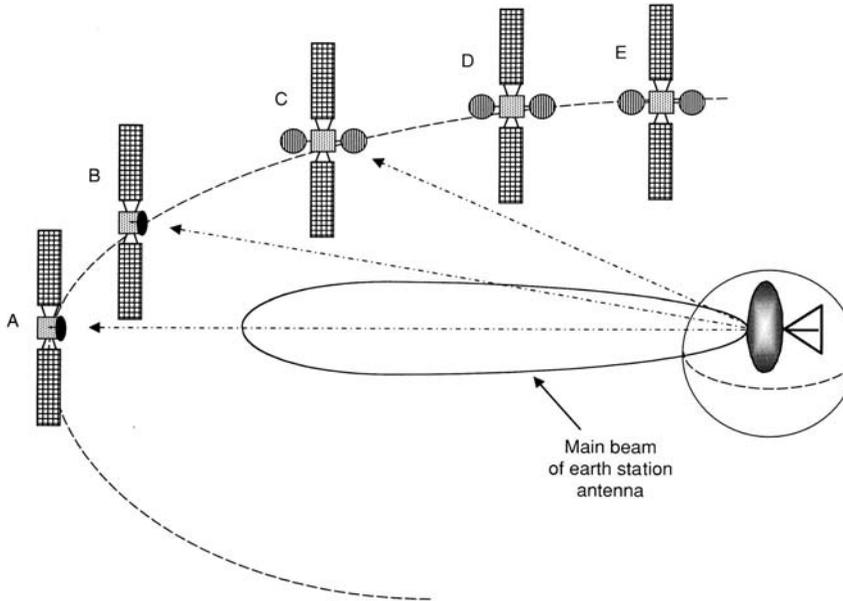
Figure 1.26 Improved GEO arc utilization from narrow Earth station antenna beams.

narrowed to the point that three satellites can be moved closer together while maintaining the desired isolation. The same result is obtained at L- or S-band if the diameter of the Earth station antenna is increased. Whenever we use a narrower beam to allow more satellites to operate, we introduce a need to point the ground

antenna more precisely. Typically, we would use a reflector-type antenna, familiar to home DTH installations. The advantage to the more broad beams for simple L- and S-band antennas is that they do not have to be pointed toward the satellite, thus facilitating mobile and personal communication applications.

The Global Positioning Satellite (GPS) system operates in a segment of L-band and employs non-GEO satellites at approximately 26,500-km altitude. Each satellite transmits in the same bandwidth using a different CDMA spreading code. Thus, a receiver such as a car navigation system with its single omnidirectional antenna can recover data from multiple satellites and from this data compute a precise location [2].

An important clarification to be made at this point is that while L- and S-bands may be forgiving as to propagation around natural and man-made obstacles, it is still desirable and often necessary that there be a clear line-of-sight path between the Earth station and the satellite. This means that nothing other than the Earth's atmosphere should be present along the path; otherwise any physical media such as wood, stone, concrete, metal, or earth will reduce signal strength and potentially introduce forms of interference referred to as multipath. This is discussed further in Chapter 4.

Larger parabolic reflector antennas are practical for L-band ship-to-shore satellite communications. These satellites are operated by Inmarsat, the global provider of international maritime communications. Until 1990, ships at sea depended primarily on HF radio communication using Morse code; that was replaced by Inmarsat terminal usage as the service achieved 100% penetration on large commercial vessels. An Inmarsat terminal can pass one or more digital telephone channels plus two-way data. The ship's directional antenna is protected from the elements by an umbrella-shaped "radome." To compensate for the rolling and pitching of the ship, the antenna is attached to a controlled mount that centers the beam on the satellite. The satellite repeater translates the link to C-band for transmission to a fixed Earth station. Telephone and data traffic then can be routed to distant points over the public telephone network and the Internet. The maritime mobile satellite system is being expanded, but the number of satellites that can operate simultaneously around the world is limited to fewer than 25. Such mobile services provide reliable communications with ships, airplanes, vehicles, and individuals as they move at will.

S-band, nominally centered at 2.5 GHz (just below C-band), was used for the downlink on the first experimental geosynchronous satellite, SYNCOM. It is even closer than C-band to the optimum frequency for space communications and is one of the bands preferred by the U.S. National Aeronautics and Space Administration (NASA) for communication with scientific deep-space probes. However, the amount of bandwidth is much less than that afforded by C- and Ku-bands.

The Globalstar non-GEO system employs S-band for the downlink to mobile users. That choice was driven by the need for adequate bandwidth at a frequency that can adapt to the wide variety of situations that users will experience in their day-to-day lives. Satellite digital audio radio service (S-DARS) and mobile TV likewise employ the S-band.

1.4.4   Microwave Bands: C, X, and Ku

The most popular microwave bands exploited for satellite communication are the C-, X-, and Ku-bands. These bands, which lie between 3 and 15 GHz, offer substantially more bandwidth than L- and S-bands. C-band maintains its popularity because it is the mainstay of such industries as cable TV in North America and satellite DTH in Asia. On a worldwide basis, C-band is very popular for domestic and international telephone services for thin routes to small cities (and countries) and to reach rural areas of such large countries as China, Russia, Brazil, and Mexico. X-band, on the other hand, is the established medium for fixed services for government and military users. The ground terminals that have been developed for these applications are designed to operate under all types of weather by soldiers and security forces in a variety of countries. Ku-band was first exploited for high-capacity trunking links between North America and Europe but was later adopted in western Europe for domestic services, particularly TV distribution. It continues as the band of choice for DTH and VSAT networks in developed countries.

   C-band was the first part of the microwave spectrum to be used extensively for commercial satellite communication. Another common designation we use is 6/4 GHz, which identifies the nominal center of the uplink frequency band (5.925 to 6.425 GHz) followed after the slash (/) by the nominal center of the downlink frequency band (3.700 to 4.200 GHz). (The uplink/downlink convention is not an industry standard, and others have reversed the order to express the numbers in ascending order.) Additional C-band spectrum was allocated by the ITU in the 1990s. Those bands, termed the expansion bands, add about another 300 MHz in each direction.

   The principal benefit of C-band is the low level of natural and man-made noise coupled with a very small degree of attenuation in heavy rain. The simplest means of overcoming noise is to increase the level of received signal power from the radio link, either by transmitting more power toward the receiving antenna or by increasing the size of the receiving antenna. Considerably more detail on the subject of radio link engineering is given in Chapter 4. Under line-of-sight conditions, C-band requires less signal level to provide reliable communications.

   Equipment technology and availability were factors in the favor of C-band's initial selection for satellite services. In the early years (1965 to 1970), C-band microwave hardware was obtainable from other applications such as terrestrial microwave, tropospheric scatter communications systems (which use high-power microwave beams to achieve over-the-horizon links), and radar. No break-through in the state of the art was necessary to take advantage of the technical features of C-band technology. Today, the equipment has been made very inexpensive (relatively speaking) because of competition and high-volume production for the global market.

   With all its benefits, C-band still is constrained in some ways because of the international requirement that it be shared with terrestrial radio services. Traditionally, C-band Earth stations were located in remote places, where terrestrial microwave signals on the same frequencies would be weak. The potential problem runs in both directions: the terrestrial microwave transmitter can interfere with satellite

reception at the Earth station (i.e., the downlink), and RF energy from an Earth station uplink can leak through antenna side lobes toward a terrestrial microwave receiver and disturb its operation. That incompatibility is addressed through the ITU's processes of: (1) requiring both types of stations to control emissions through technical sharing criteria, and (2) conducting terrestrial frequency coordination when signals could traverse national borders.

Shielding is the technique by which sharing can be made to work (Figure 1.27). Let's say that an Earth station must be located about 30 km from a terrestrial microwave station that either transmits or receives on the same C-band frequency. The signals could be blocked or at least greatly attenuated by a natural or man-made obstacle located very near the Earth station antenna but between it and the terrestrial microwave station. As stated at the beginning of this chapter, microwave signals travel in a straight line, and one would expect that an obstacle would block them entirely. However, microwave energy can bend over the top of such an obstacle through electromagnetic diffraction. Visible light, being electromagnetic in nature, also will bend over a knife-edge obstacle. What diffraction does is cause the wave to propagate over the top of the obstacle and thereby potentially interfere with reception on the other side. The amount of bending can be predicted and is a function of the distances between the source, the obstacle, and the receiver, as well as of the height differences (indicated in Figure 1.27 by H_1 and H_2). If the height differences are large, causing the antenna or antennas to lie well below the top of the obstacle, little signal will reach the receiver and good shielding is therefore achieved. Note that shielding is equal for both directions of propagation (i.e., from Earth station to terrestrial microwave tower and vice versa).

A distance of greater than 50 km usually provides adequate natural shielding from the curvature of the Earth augmented by foliage and man-made structures. Obviously, if the microwave station is on top of a high building or a mountain, the Earth station siting engineer will have to look long and hard for adequate natural shielding. Man-made shielding in the form of a 10–20-m metal or concrete wall has proved effective in such difficult situations.

C-band must live with yet another aspect of sharing: the ITU regulation that attempts to protect terrestrial radio receivers from direct satellite radiated signals. The level of such signals is low relative to that emanating from an Earth station;
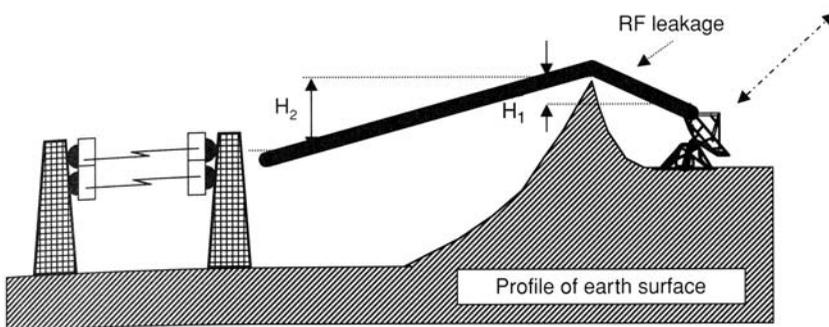
Figure 1.27   The use of terrain shielding to block RFI between a terrestrial microwave link and an Earth station.

however, if satellite power is allowed to increase without bound, it is theoretically possible to produce measurable interference into a terrestrial receiver. Figure 1.28 depicts the situation that has the greatest potential for such interference: when the antenna of the terrestrial station is pointing at another terrestrial station along the horizon but on a direct path to a satellite in orbit. To deal with that potential problem, the regulations limit the power flux density (power per square meter on the surface of the Earth produced by the satellite) to an amount that would not cause significant interference to any terrestrial microwave receiver no matter where it might be located. In addition, terrestrial microwave antennas are not to point in the direction of GEO.

Government and military satellite communications systems in the United States and some other countries employ X-band and, on a limited basis, Ka-band. With an uplink range of 7.90 to 8.40 GHz and a downlink range of 7.25 to 7.75 GHz, X-band is used extensively for military long-haul communications links, much as C-band is used on a commercial basis. In highly specialized cases, Ka-band is being applied because very narrow spot beams can be transmitted to and from the satellite.

The frequency band that has done more to interest new users of satellite communications is Ku-band, a part of the spectrum lying just above 10 GHz. Portions of Ku-band are not shared with terrestrial radio, which has some advantages over C-band particularly for thin-route and receive-only services using inexpensive terminals with small-diameter antennas. The precise uplink and downlink frequency ranges allocated by the ITU vary to some degree from region to region.

Effectively, three sections of Ku-band have been allocated to different services on an international or a domestic basis. The most prevalent is FSS, which is the service intended for one- or two-way communications between fixed points on the ground. All of C-band and much of Ku-band are allocated to FSS for wide application in international and domestic communications. Part of Ku-band is subject to
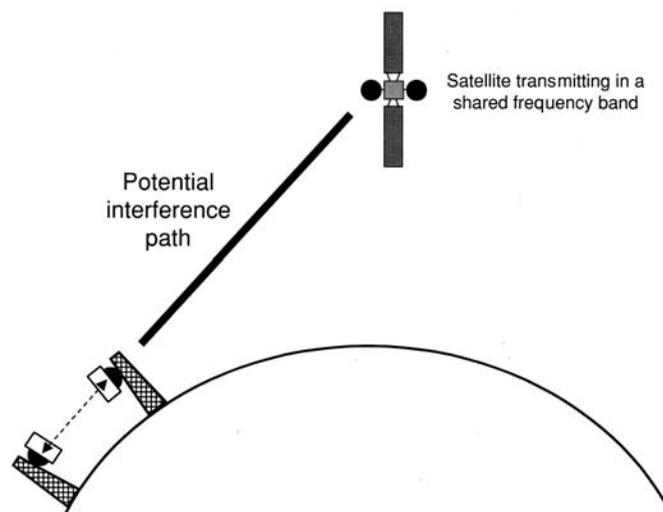
Figure 1.28   The potential for direct RFI from a satellite into a terrestrial microwave receiver where the frequency band is shared.

the same coordination and siting difficulties as C-band. This particular part of Ku-band is referred to as 14/11 GHz, where the uplink range is 14.00 to 14.500 GHz, and the downlink range is 10.95 to 11.7 GHz (minus a gap of 0.25 GHz in the center). Only the downlink part of the allocation is actually subject to sharing.

A portion of the Ku allocation for FSS that is not shared with terrestrial services is referred to as 14/12 GHz (the uplink range again is 14.00 to 14.50 GHz, and the downlink range is 11.70 to 12.20 GHz). The availability of 14/12 GHz is limited to Region 2, which comprises North America and South America, and can be used only for domestic communications services. North America, in particular, has seen wide use of 14/12 GHz since about 1982; services in this band were first introduced in South America in 1992. Power levels from these satellites are not subject to the same restrictions as at C-band, although there is an upper limit, to minimize interference between satellites. Ku-band satellite operations in the rest of the world (i.e., Regions 1 and 3) are restricted to the 14/11 GHz shared allocation. In some instances, a Region 1 (Europe and Africa) or Region 3 (Asia) country can make 14/11 GHz appear like 14/12 GHz simply by precluding domestic terrestrial services from that band. However, terrestrial radio services in adjacent countries are not under their control, so international coordination still must be dealt with in border areas.

A third segment of Ku-band, referred to as 18/12 GHz, is allocated strictly to BSS. As with the 14/12 FSS band in Region 2, the BSS band is not shared with terrestrial services. Its intended purpose is to allow television broadcasting and other DTH transmissions from the satellite. There are two regulatory features of this band that make direct broadcasting to small antennas feasible. Because the band is not shared with terrestrial radio, the satellite power level can be set at the highest possible level. Adjacent satellite interference could be a problem but is precluded by the second feature: BSS satellites are to be spaced a comfortable 6 or 9 degrees apart, depending on the region. In comparison, while there is no mandated separation between FSS satellites, a 2-degree spacing has become the standard in crowded orbital arc segments such as North America, Europe, and East Asia.

The operational advantages of 14/12 and 18/12 GHz lie with the simplicity of locating Earth station sites (without regard to terrestrial radio stations) and the higher satellite downlink power levels permitted. The latter results in smaller ground antenna diameters than at C-band, all other things being equal. However, as discussed earlier in this section, Ku-band is subject to higher rain attentuation, which can increase the incidents and duration of loss of an acceptable signal. Figure 1.29 indicates the relative amounts of extra downlink power, measured in decibels, needed to reduce the outage time to four hours a month (on average). Ka-band (30/20 GHz) is included for completeness and is discussed in Section 1.4.5. The decibel (dB) is a logarithmic scale (like the Richter scale used to measure earthquake intensity) and is discussed in Chapter 4.

As shown in Figure 1.29, the amount of decibels to overcome a fade is also a strong function of the elevation angle from the Earth station to the satellite in orbit. Rain cells, which are large atmospheric volumes (wider than they are high), produce the greatest attenuation. Because of the flattened shape, radio waves that take off at a low elevation angle relative to the ground must pass through a greater
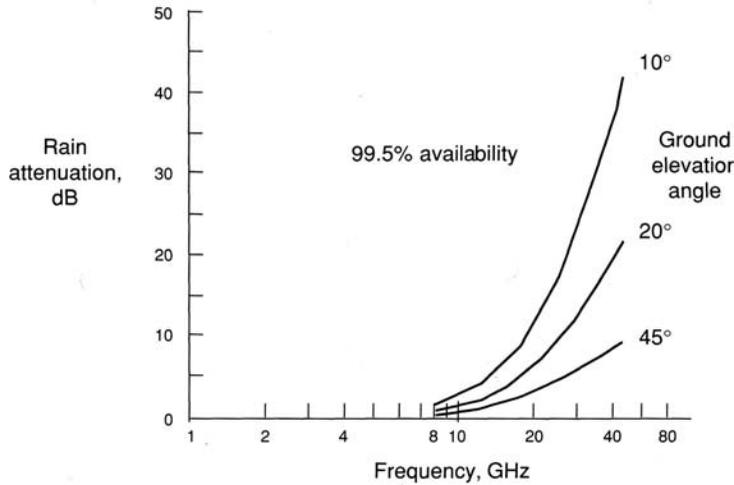
Figure 1.29   Rain attenuation as a function of frequency for different ground elevation angles.

thickness of rainfall. Consequently, elevation angles of 30 degrees or greater are preferred for Ku-band frequencies and higher. Another important variable is the local climate: deserts are less affected than tropical regions. In general, the need for greater power margin at Ku-band tends to reduce some of the benefits obtainable by virtue of the higher satellite powers permitted by international regulations. A proper design at Ku-band still can deliver a quality TV, voice, and data signal that is maintained for a satisfactory period of time. Only for those regions with the highest rainfall, such as a tropical rain forest, and where there is a high incidence of thunderstorms, would Ku-band tend to be less practical than C-band.

1.4.5   Millimeter Wave and Higher: Ka-, Q-, and V-Bands

Beginning at around 20 GHz and extending to suboptical frequencies, we find the millimeter-wave spectrum range. These frequencies were viewed as the domain of the researcher until around 1970, when satellite planners began to look ahead to the time when the C- and Ku-bands would be heavily used. Also, they saw some unique applications that might employ the greater bandwidth and smaller wavelengths that came with these allocations. First to be considered was Ka-band, with its 30/20-GHz allocations and about 2.5 GHz of available spectrum. The frequencies above about 40 GHz, designated as Q- and V-band, are now being evaluated for space and ground applications because of the wide bandwidths. During heavy rainstorms, the atmosphere will appear to be nearly opaque at these frequencies, so applications must allow for some periods of total outage.

There is a potential for the use of Ka-band (30/20 GHz) for commercial applications in cases where C- and Ku-bands are fully occupied. During the 1970s, the Japanese space agency, NASDA, in conjunction with the government-owned telecommunications company, NTT, launched some experimental communications satellites that operate in the Ka-band frequency range. The experimental links generally showed that the ample bandwidth that was allocated could in fact be

used under certain circumstances. Severe rain attenuation at these high frequencies may require that a given Earth station include two receiving sites separated sufficiently so that only one is heavily affected at a time. This principle of diversity reception allows the downlink to be switched to the site with the acceptable signal. That works because the most intense part of the storm is smaller than the spacing between receiving sites. In the United States, NASA experimented at 30/20 GHz with the Advanced Communications Technology Satellite (ACTS). The ACTS demonstrated a series of interesting and possibly useful applications in the area of high-speed data transfer and mobile communications, all at Ka-band. A kind of frequency "land rush" developed in the mid-1990s in the United States, when several companies filed applications for spectrum and GEO orbit positions with the Federal Communications Commission (FCC). At the time of this writing, only two of these systems actually reached orbit. The U.S. government also purchased Ka-band satellites for military use. Prospects for advanced Ka-band systems are reviewed in Chapter 2.

1.4.6   Guided and Unguided Optical Properties

Optical communication is based on two essentially different schemes—guided waves through transparent medium and unguided waves in free space. The relationship between optical wavelengths and those of microwave frequencies is illustrated in Figure 1.30. The electromagnetic spectrum (above) shows frequency in Hertz and wavelength in meters. At the right end, the wavelength is only one millionth of a meter, also referred to as either one micron or 1,000 nanometers (nm). The corresponding frequency is approximately 300 trillion Hz, or 300 terahertz (THz).

The optical spectrum, shown at the bottom of Figure 1.30, is reversed from that of the electromagnetic spectrum. Thus, 800 nm at the low end actually corresponds to a frequency that is higher than the dotted line at 1,000 nm (corresponding to 300 THz). The curves represent absorption characteristics of glass; there are three useful "windows" for optical transmission purposes. Most optical communication systems operate within the first window in the infrared, which is from 800 to 900 nm and not visible to the human eye. Interestingly enough, this amounts to 30 THz of potential bandwidth.

The now-familiar fiber communication is based on guided optical transmission through glass fiber. A typical fiber link includes an optical transmitter, the fiber itself, and an optical receiver. Glass fibers that are used fall into two general categories: multimode fiber that is generally limited to short distances and single-mode fiber that is applied to links of significant length. Both are based on the principle that light waves are reflected within the fiber by the transition of refractive index produced by a change in glass density.

Unlike fiber optics, unguided optical systems transmit light through the atmosphere and space, much like their microwave counterparts. This technology has been the subject of significant experimentation over the last three decades, resulting in several space communications trials. Such work in the United States has been done for the U.S. government and is highly classified.

The European Space Agency (ESA) developed the first commercial optical laser communication system called Semiconductor Laser Intersatellite Link Experiment
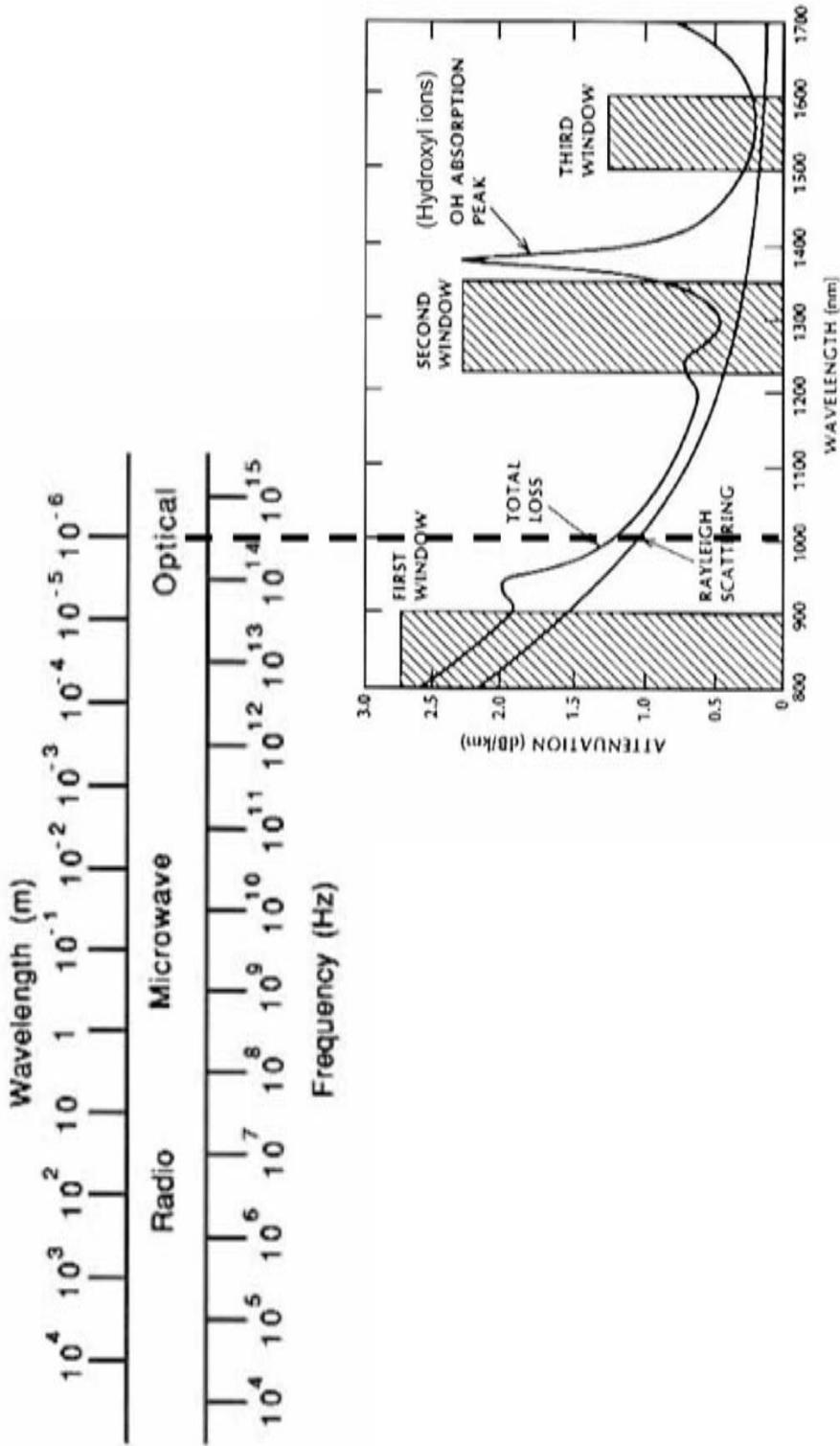
Figure 1.30 Spectral location and wavelengths of optical transmission.
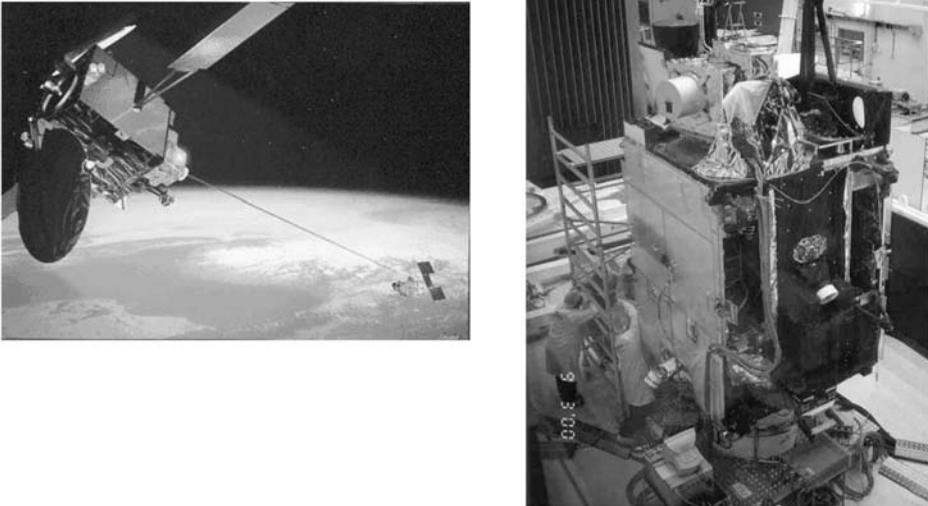
Figure 1.31 SILEX optical transmission system on the ARTEMIS satellite, linked with a LEO Earth
observation satellite. (*Source:* European Space Agency.)

(SILEX). The mission, shown in Figure 1.31, is a free-space optical communication
system that consists of two optical communication payloads to orbit on the ESA
Advanced Relay and Technology Mission Satellite (Artemis) spacecraft and on the
French Earth-observation spacecraft SPOT-4. It allowed data transmission at
50 Mbps from low Earth orbit (LEO) to geostationary orbit (GEO) using GaAlAs
laser diodes and direct detection. ESA has continued with the development of
Nd-YAG laser-based coherent laser communication systems and related hardware
technologies.

Efforts are underway to employ space-based laser communications for national
defense. The technology for this purpose will take some years to refine to the point
that an operational system can be deployed.

References

[1]   Elbert, B. R., *The Satellite Communication Applications Handbook,* 2nd ed., Norwood,
      MA: Artech House, 2004.
[2]   Kaplan, E. D., and C. J. Hegarty, (eds.), *Understanding GPS: Principles and Applications,*
      2nd ed., Norwood, MA: Artech House, 2006.

Evolution of Satellite Communication

The satellite communication industry has grown tremendously since the first edition of this text, and with it the number of professionals and the range of activities have grown as well. With many of the originators of the basic technical and operating principles having reached retirement, it is likely that a significant fraction of industry participants lack a historical perspective. What follows is an overview of key developments and events; it is not meant to be exhaustive. It can only be said that satellite communication is an extremely rich field of endeavor, encompassing space technology, telecommunications systems, information technology, entertainment television and audio, and the Internet. Add to that the introduction of new markets in satellite-based mobile communication, and you begin to comprehend the possibilities for new historical chapters.

Figure 2.1 is a basic timeline of many of the critical events. A familiarity with the past satellite programs and applications will aid the reader in understanding why today's satellite systems and services are the way they are. This chapter can be read in sequence or deferred until the basic concepts covered in Chapters 3 through 8 have been absorbed. Additional historical background can be found in [1].

2.1 Source of the Original Idea

While various individuals in the early part of this century recognized the existence of the 24-hour geostationary orbit, it is the late science fiction author Arthur C. Clarke who is credited with conceiving of the application to communication satellites. His article on extraterrestrial relays in *Wireless World* magazine back in 1945 specifically describes a three-satellite worldwide network, including a drawing similar to Figure 1.1. As a result, the GEO is also called the Clarke Orbit, in recognition of Mr. Clarke's forethought. He knew at that time that the requisite electronic and spacecraft technology would take a few years to develop; in fact, it was not until the early 1960s that the first workable spacecraft was built and launched into geosynchronous orbit.

Following the launch of the Soviet Sputnik satellite in 1957, the United States accelerated its space program and many proposals were developed for space relay satellites. The geosynchronous satellite was thought by many to be too speculative, and efforts were centered instead on LEO passive reflector and active repeater technology. Programs like Echo, Advent, Telstar, Courier, and Relay were pursued
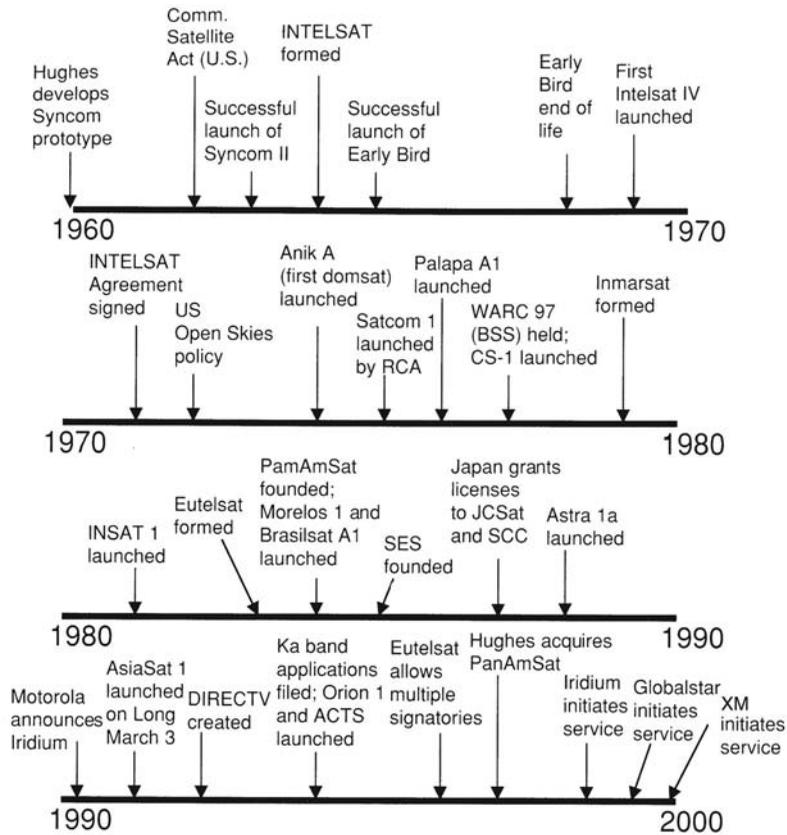
Figure 2.1   Timeline of critical events in the satellite communications industry.

by NASA (and AT&T in the case of Telstar) in hopes of coming up with a viable system that would be competitive with terrestrial radio, buried cable, and transoceanic cable. The feeling in that community was that LEO was the appropriate orbit because of the reduced time delay and the difficulty of constructing and operating a workable GEO spacecraft. It is interesting that a few non-GEO systems have been deployed although their long-term future is still being debated.

2.1.1   SYNCOM

In 1959, Dr. Harold A. Rosen and two of his colleagues at Hughes Aircraft Company, Tom Hudspeth and the late Don Williams, identified the key technologies for a simple and lightweight active repeater communication satellite for launch into GEO. Dr. Rosen was the product champion and systems engineer; Tom Hudspeth (a former radio mechanic on the Spruce Goose) provided the microwave know-how that produced a compact but effective repeater for space application; and Don Williams, a brilliant Harvard-trained mathematician, figured out how the satellite could be placed into orbit and oriented for satisfactory operation.

The basic concept used a drum-shaped spinning body for stability with tiny gas jets to alter the attitude (orientation) in space. In 1960 at Hughes Laboratories

in Los Angeles, the team built a working prototype that demonstrated the feasibility of spin stabilization and microwave communication. Supported by Dr. Allen Puckett, then vice president (and later chairman of the board of Hughes Aircraft), they convinced NASA and the Department of Defense to go ahead with the launch of SYNCOM, an acronym for ''synchronous orbit communications satellite.'' On the second attempted launch of the Thor Delta rocket in July 1963, SYNCOM II became the first operational geosynchronous satellite providing intercontinental communications. Little did the team know that SYNCOM, shown in Figure 2.2, would be the precursor for a powerful telecommunication medium and the foundation of whole new industries.

2.1.2 COMSAT

The idea of developing a commercial communications business based on the use of satellites can be attributed to the administration of President John F. Kennedy, which issued ground rules in 1961 for the U.S. operation of an international
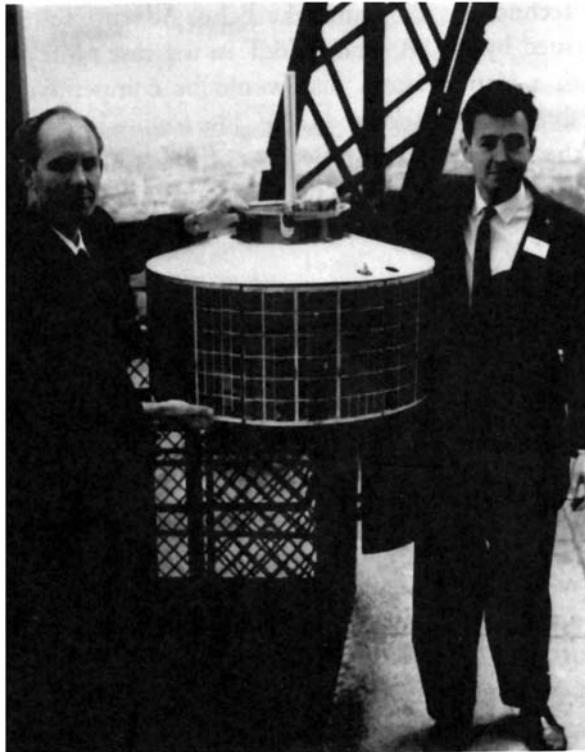
Figure 2.2 Historic photo taken from high in the Eiffel Tower shows the world's original synchronous communications satellite, which Hughes Aircraft Company demonstrated at the Paris Air Show in 1961 in an effort to find a sponsor to launch it. ''This is as high as it will ever get,'' observed one skeptic. Scientists pooh-poohed the satellite as unworkable until NASA backed Hughes in the SYNCOM program and paved the way for the current age of space communications. The original satellite embodied design features now used in every commercial satellite in service. Thomas Hudspeth (left) designed the spacecraft's electronics, and Dr. Harold A. Rosen (right) was the leader of the group that conceived and hand-built the satellite.

communications satellite system. President Kennedy favored private ownership and operation of the U.S. portion of an international system that would benefit all the nations of the world, large and small. The Communications Satellite Act of 1962 resulted from that initiative, establishing the charter for the Communications Satellite Corporation (COMSAT). Under the leadership of its first chief executive, Dr. Joseph Charyk, and technical vice president, Siegfried Reiger, the company developed a workable satellite system and helped to encourage worldwide expansion. Over the years, COMSAT gradually moved away from that core and into a variety of related and unrelated businesses, including ownership of the Denver Nuggets basketball team. Nevertheless, many of the innovations in satellite communication networks and operation were COMSAT's creations.

COMSAT raised money through the sale of common stock both to the public and to the U.S. common carriers, particularly AT&T. At the time, it was assumed that subsynchronous satellites would be used basically for two reasons. First, the U.S. communications companies felt that the time delay associated with the propagation of radio waves, which travel at the speed of light (approximately one-quarter of a second for the GEO uplink and downlink combined), would prove to be unacceptable to a significant percentage of telephone subscribers. Second, there existed a lingering doubt about the viability of the technology necessary to deliver and operate the spacecraft. Those fears were substantially dissolved when SYNCOM demonstrated the feasibility and quality of synchronous orbit satellite communications. That success allowed COMSAT to create INTELSAT and, at the same time, to save more than half its original $200 million of capital (a lot of money at the time) that otherwise would have been needed to launch the required LEO constellation.

2.2   Evolving Satellite Coverage

Once Hughes, NASA, and COMSAT proved that GEO satellites were the wave of the future, it remained a difficult task to bring the first systems into practical existence. That was the role of the global satellite operators, along with regional and domestic systems. All three demonstrated that GEO satellites could fill a wide variety of telecommunications needs, in line with the advantages of satellite communications outlined in Chapter 1.

2.2.1   Global Service: INTELSAT, PanAmSat, and Orion

Clearly, the first and most demonstrable need for the commercial satellite was to provide international communications links across the oceans. At the beginning of the international GEO satellite era, most telecommunications companies around the world were government owned or controlled (with the exception of North America). In time, a trend toward privatization and even competition appeared, which eventually affected the structure of international satellite operation. The following paragraphs outline the evolution of the now well-established international satellite system.

2.2.1.1 Early Bird

In March 1964, COMSAT contracted with Hughes Aircraft Company for the construction of two spin-stabilized satellites using C-band (SYNCOM used S-band, which was authorized only for NASA experimentation). Early Bird, the well-recognized name for this first commercial satellite, was launched on a McDonnell Douglas Delta rocket in the spring of 1965. The satellite worked perfectly for six years, well past its design life of two years. The second spacecraft was never launched and now resides in the Air and Space Museum of the Smithsonian Institution in Washington, D.C. Early Bird was positioned over the Atlantic Ocean and was first used to link stations in Andover, Maine, and Goonhilly Downs, England, providing voice, telex, and television service. Having gained such renown, *bird* continues to be used as a nickname for all GEO satellites.

2.2.1.2 INTELSAT

The intended international nature of the system led the U.S. government and COMSAT to establish a new organization in 1964 called the International Telecommunications Satellite Consortium (INTELSAT). Later, *Consortium* was changed to *Organization*, and INTELSAT grew to become the preeminent satellite operator in the world. The INTELSAT Agreement, a treaty among member countries, was signed in 1971. COMSAT continued as the U.S. signatory, while the Government Post Office (now British Telecom) fulfilled that role for the United Kingdom. In time, INTELSAT evolved into more of a business, purchasing spacecraft based on the global demand for telephone and TV services and other noninternational applications. INTELSAT was completely privatized and is currently owned by private investors.

   This quasi-governmental approach served INTELSAT and the world community well for more than three decades. The signatories and their respective users employed the capacity to implement international telephone and TV links to augment terrestrial transmission means like fiber optic cable and microwave links. One important application is the use of INTELSAT links for fiber optic cable restoration. This was used again when an earthquake off of Taiwan severed several cables. In 1984, INTELSAT introduced the INTELSAT Business Service (IBS), for business communication private lines. In some countries, corporate users could install and operate their own Earth stations for direct access to the INTELSAT space segment. The service became very popular, because it was based on digital transmission and the pricing was attractive relative to public leased lines. INTELSAT then proceeded to digitize all its international telephone links, resulting in the extensive use of intermediate digital rate (IDR) carriers. The widespread use of IDR carriers at capacities of 2 Mbps and higher improved the efficiency and quality of telephone services over INTELSAT satellites. INTELSAT followed a deliberate process to establish the requirements for its satellites and to procure them from manufacturers. Each generation of spacecraft, indicated in Figure 2.3, was marked by a significant upgrade in capability.

   The big step up in satellite capability happened with the introduction of the Hughes Aircraft-built Intelsat IV (Figure 2.4), a spacecraft many times larger than its commercial forerunners. A major innovation came in the way the repeater and
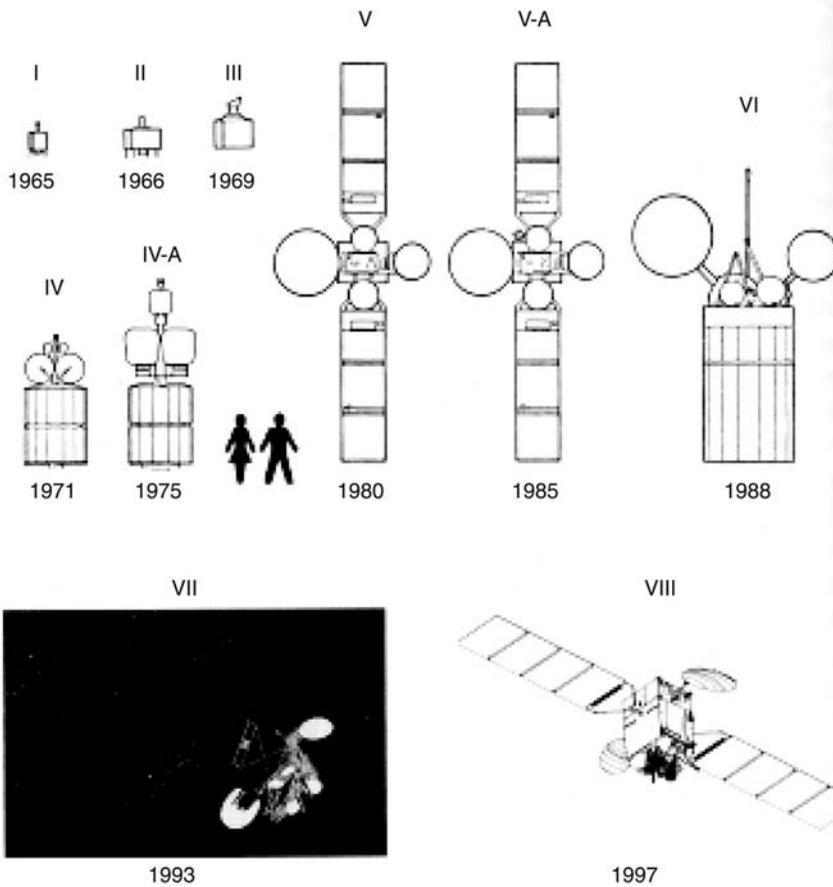
Figure 2.3 Eight generations of INTELSAT satellites.

the antenna were attached. Instead of spinning the repeater with the spacecraft and despinning the antenna, the repeater with the antenna were despun as a package. The configuration resulted from the work of Dr. Anthony Iorillo and is called the GYROSTAT. With considerably more power and communications equipment, the C-band repeater on Intelsat IV contained 12 individual channels of approximately 36 MHz each. The selection of the now standard 36-MHz RF bandwidth resulted from the technical trade-off involving the bandwidth required for a single TV channel (using FM) and the maximum number of amplifiers that could be carried and powered. The term *transponder* has come into common usage to refer to the individual RF channel, typically of 36-MHz bandwidth. The first Intelsat IV was launched in 1971, and a total of seven have each provided 10 or more years of service. The Intelsat IV series became the backbone of international communications, and the last of the series was retired at the end of 1985. The transponder concept has been used in every subsequent commercial satellite operating at C- and Ku-band.

Figure 2.4   The Intelsat IV spacecraft. (*Source:* Photograph courtesy of Hughes Aircraft Company.)

INTELSAT saw the need to go beyond the design of Intelsat IV to increase the capacity in the Atlantic Ocean region (AOR). Dr. Rosen of Hughes Aircraft developed a practical approach to increase satellite capacity while retaining the basic configuration of the Intelsat IV spacecraft. The concept, called frequency reuse, doubles the 500 MHz of allocated bandwidth at the satellite by directing two independent beams toward the visible face of the Earth (the east and west hemi beams in Figure 2.5). Shaping of each beam causes the footprint to cover the land mass of the eastern or western hemisphere, leaving the open sea and some land extremities outside the coverage area (that is necessary to prevent the signals in one beam from interfering with signals on the same frequency in the other beam).

Because of continued expansion of the INTELSAT system, in terms of both the number of member/users and the traffic demands, INTELSAT found that the problems of traffic overload in the AOR and also the Pacific Ocean region (POR) required another doubling of capacity per satellite. The 14/11-GHz portion of Ku-band was selected as the means of carrying heavy links in the regions, freeing up C-band for general connectivity among countries. A satellite with both C- and Ku-bands generally is referred to as a hybrid satellite. The Ku-band transponders are connected to spot beams that point toward heavy traffic regions, such as Western Europe and the northeastern United States (ES and WS in Figure 2.6).
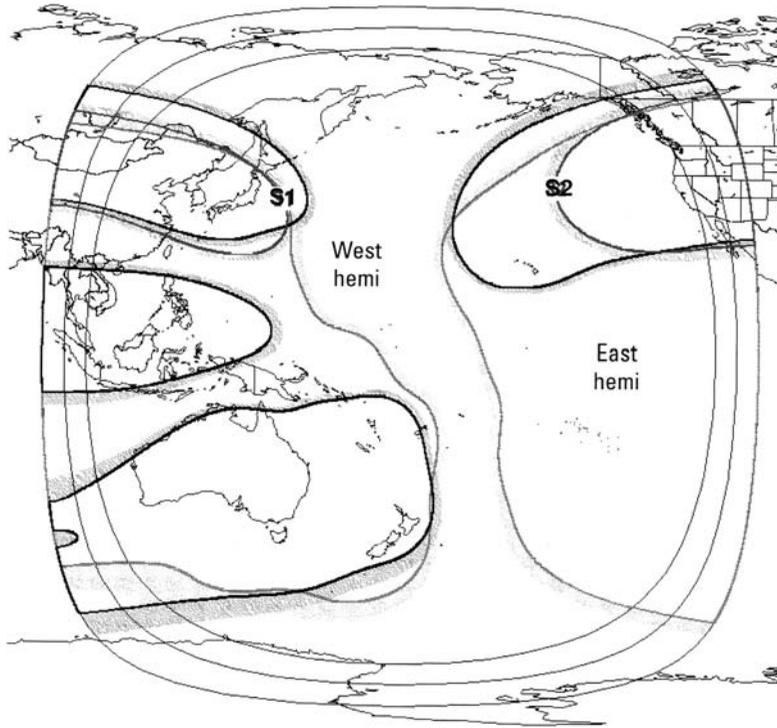
Figure 2.5 Intelsat VII Pacific Ocean region frequency reuse coverage; hemi and zone beam cover-
age at C-band and spot beam (S1 and S2) coverage at Ku-band. (*Source:* Intelsat.)

Having conceived the hybrid approach, COMSAT proceeded to conduct an interna-
tional tender for the supply of Intelsat V spacecraft.

The Intelsat V program was begun, with Ford Aerospace and Communications
Corporation, in Palo Alto, California, winning the development and production
contract. (The company was later purchased by Loral and continues as Space
Systems/Loral.) Ford chose the three-axis approach (also called body-stabilized),
whereby an internal spinning wheel and small jets stabilize the satellite to keep the
footprint over the proper portion of the Earth (more on this in Chapter 8). With
Intelsat IV and IV-A providing excellent service throughout the 1970s, the first
Intelsat V was launched in 1980.

The C-band repeater on Intelsat V is very similar to that of Intelsat IV-A, with
the exception that frequency reuse was taken to another dimension with the use
of polarization discrimination. In that approach, two signals on the same frequency
and in the same footprint are isolated from each other by cross-polarizing their
electromagnetic waves. A more complete explanation can be found in Chapter 4.
This allows a doubling of bandwidth in the Intelsat V global coverage beam. Having
a Ku-band repeater onboard provides the extra capacity desired and even allows
cross-connection between stations using C-band and those using Ku-band.

Conceived in the late 1970s to provide for the widest possible expansion of
international satellite services, the massive Intelsat VI spacecraft was defined by
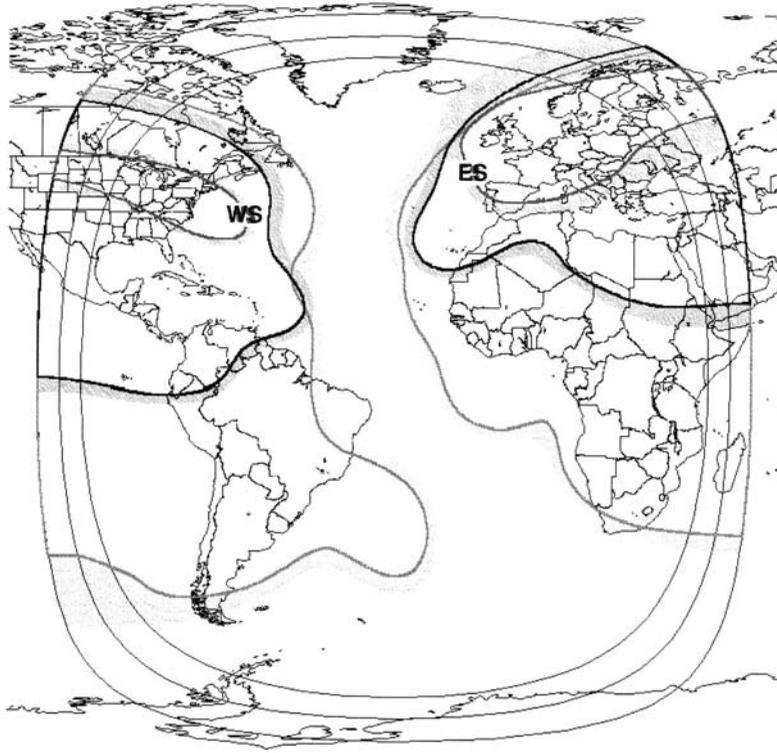
Figure 2.6 Intelsat VI AOR frequency reuse coverage; hemi and zone beam coverage at C-band and spot beam (west and east) at Ku-band. (*Source:* Intelsat.)

INTELSAT to continue the decline in cost per circuit. Hughes Aircraft won the contract in 1982 for the design and production of this very capable satellite. Again, it was a hybrid design (C- and Ku-bands), but frequency reuse was carried to an even greater stage to triple the capacity of the satellite compared to Intelsat V/V-A. A new feature on Intelsat VI was satellite-switched time division multiple access (SS/TDMA), which increases the efficiency of spot beams for high-capacity links and allows their reconfiguration with minimal impact on the Earth stations.

The 1990s saw INTELSAT expand its global system through the construction and launch of the Intelsat VII and VIII series of hybrid spacecraft. A total of nine Intelsat VII satellites were constructed by Loral, and eight reached operational status, beginning with the 701 satellite in October 1993. The next production contract was with GE Astro (now Lockheed Martin) for Intelsat VIII, which concentrates on delivering greater C-band capacity for overall system expansion. Each of these spacecraft contains 38 C-band transponders and 6 Ku-band transponders. The next spacecraft to enter the production cycle was Intelsat IX, constructed by Loral and first launched in 2001 by Arianespace.

In addition to funding technology development and investment in new satellites, INTELSAT created many innovative ground-based systems of communication. The initial system used analog transmission of telephone and video signals. However, it was clear to the engineers at COMSAT who provided most of the technical

foundation for the international system that digital technologies would provide substantial benefits. Through the efforts of COMSAT Laboratories, the internal research organization, key transmission technologies like time division multiple access (TDMA) and demand assignment multiple access (DAMA) were invented and perfected.

In a step toward privatization, INTELSAT spun off Newskies Satellites to operate selected satellites for commercial services. Newskies set up shop in the Netherlands. The conversion of INTELSAT from a quasi-governmental institution to an aggressive global service company was completed in 2005. As the new Intelsat, it is the largest satellite operator in the world with over 50 GEO satellites. A good chunk of this expansion resulted from the acquisition of PanAmSat, described next.

2.2.1.3   PanAmSat

Founded in 1984 by legendary satellite entrepreneur Rene Anselmo, PanAmSat became a powerhouse in the global marketplace for space segment services. Anselmo started the company by using his own personal funds to purchase a spacecraft from GE Astro (now part of Lockheed Martin), coupling it with an inexpensive launch on the first Ariane rocket. The real challenge that the company overcame was the prohibition on any U.S. company competing with INTELSAT in the international satellite communication market. PAS-1 was successfully placed in orbit in 1988 to provide services in the AOR. PanAmSat's first customer was CNN, which selected PAS-1 to transmit programming to Latin America.

Anselmo broke through the political barrier and built his company from the ground up. That involved considerable work in Washington to obtain support from the U.S. government to force COMSAT and INTELSAT to allow competitive international satellite operator to enter the business. Prior to that time, INTELSAT used the powerful provisions of the INTELSAT Agreement to preclude signatories for encroaching on its territory. After PanAmSat reached profitable operation, Anselmo became famous for his steadfast courage against difficult odds. PanAmSat's advertisements in various satellite industry publications featured a cartoon dog named Spot, which indicated Anselmo's disregard for the entrenched monopolies then prevalent.

PAS 1, the first satellite launched by Mr. Anselmo's startup company, contained 18 C-band transponders, split between 8.5W SSPA and 16.2W TWTA amplifiers, and 6 Ku-band transponders, each with 16.2W amplifiers. The satellite served the AOR, in direct competition with INTELSAT's primary region of business. The initial business was to serve Latin American TV customers, but PanAmSat was also successful at penetrating the U.S. to U.K. market. That new market entry required that PanAmSat obtain landing rights, which is the permission of the local government to deliver or originate transmissions within its border. Landing rights acquired cheaply (the way PanAmSat did it) eventually can be worth a lot of money. A second and a third satellite were added to the lineup in the Atlantic. Pacific Ocean service was initiated in July 1994 with the launch of the Hughes-built PAS 2 satellite. That bird carried 16 C-band and 12 Ku-band transponders and was very popular for trans-Pacific TV transmission, Asia program distribution, and

VSAT networks. Service was extended to the Indian Ocean region in January 1996 with the launch of PAS 3.

After Anselmo's death in 1995, the company entered into a merger with Hughes Communications, to form a new PanAmSat Corporation, which was the largest satellite operator in the world at the time. On May 16, 1997, after the merger, PanAmSat owned and operated a fleet of 14 satellites; four PanAmSats, designated PAS 1, 2, 3, and 4; seven Galaxy satellites, designated as Galaxy 1R, 3R, 4, 5, 6, 7, and 9; and three Satellite Business Systems (SBS) satellites, designated SBS 4, 5, and 6. The SBS satellites were assets acquired by Hughes Communications from MCI Corporation in 1991 and all have been retired by the time of this writing. The Galaxy satellites generally serve the U.S. market (see Figure 2.7) and were contributed by Hughes to the merged company. Galaxy 1 was launched in 1983 for use by the leading cable networks in the United States, including HBO, CNN, the Disney Channel, and ESPN. Subsequently, Hughes Communications built the Galaxy System into the most financially successful satellite operation in history. To attest to its market power, the name of Galaxy is preserved by Intelsat for all of its satellites serving North America.

PanAmSat expanded its system by launching additional satellites for international services. Those satellites were built by Loral and Hughes and offer higher power and more focused service to specific markets. The combination of PAS and Galaxy satellites gave PanAmSat a profitable base and allowed the company to expand in the late 1990s.

2.2.1.4 Orion

The official name of the second private international satellite startup company is Orion Network Systems. Formed in October 1982 and initially the brainchild of
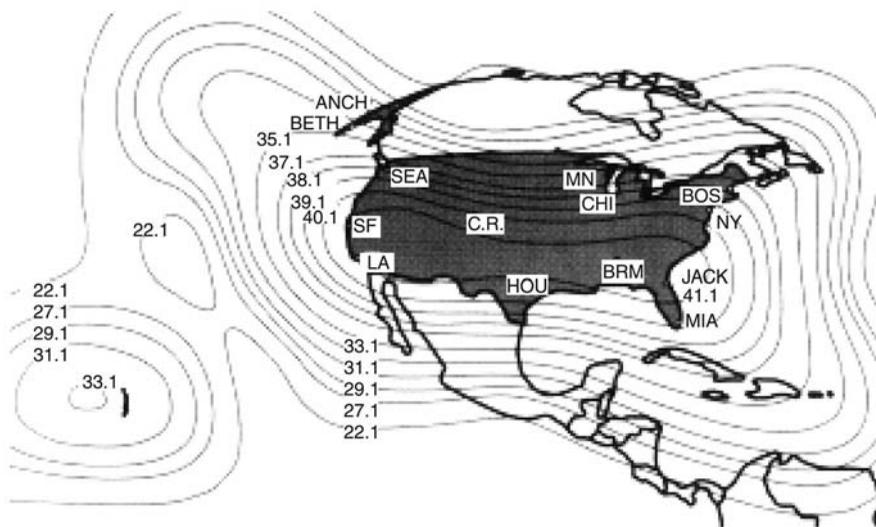
Figure 2.7 U.S. domestic satellite footprint coverage in C-band (99° WL).

John Puente, another famous satellite entrepreneur, Orion had a slow beginning. Like PanAmSat, Orion selected the AOR market and the strategy of gaining market access in the United States and the United Kingdom. Orion chose to build the company as a joint venture with British Aerospace (BAe) and thereby obtained access to spacecraft technology and funding. Orion contracted with BAe for two satellites but completed only the purchase of Orion 1, launched in November 1994.

Orion contracted with Space Systems/Loral for a second satellite, and the company was subsequently acquired by Loral. The Orion assets were combined with those of the Telstar system, discussed later in this chapter, to form Loral Skynet. The second Orion satellite was launched in October 1999 and later renamed Telstar 12.

2.2.2   Regional Coverage: EUTELSAT and SES

International systems, notably INTELSAT, were created to meet long-haul requirements for telecommunications transmission. Because the satellites must provide broad coverage, they typically are not optimized to the needs of particular countries or affiliated groups of countries. The first to do this on significant scale were the Western European countries, in the formation of EUTELSAT. The parallels with INTELSAT are clear, including the basic strategy of a space segment cooperative whose members provide their own Earth stations.

The EUTELSAT project appeared in 1977 as a regional provider along European lines, including the focus for development of the indigenous manufacturing base. The first EUTELSAT satellite was launched in 1983 by the European Space Agency (ESA). The organization was formally established as an operating company in 1985 with an initial membership of 17 Western European countries. Its first director general was Andrea Caruso, a well-recognized figure in INTELSAT circles. With the opening of Eastern Europe, EUTELSAT grew remarkably and the membership included Bulgaria, Estonia, Kazakhstan, and Slovenia. In time, EUTELSAT has become the leading provider of satellite capacity over the now-extended Europe. This role is extremely vital to a wide range of public and private users. Like INTELSAT, EUTELSAT was a treaty organization that operates like an intergovernmental body. The official government members, called Parties, met once a year in an Assembly of Parties to make decisions on long-term objectives.

And like Intelsat, EUTELSAT has made a rather smooth transition to a purely commercial company with private investor and public ownership. Under the leadership of Giuliano Berretta, Chairman of the Board–CEO, EUTELSAT was transformed into a competitive and commercial organization that ranks as number three in terms of global satellite operators. Through the 2005 IPO, EUTELSAT Communications stock is listed on the Paris Stock Exchange (ticker symbol ETL).

A completely commercial satellite company called Société Européene des Satéllites (SES) was established in 1985 in Luxembourg. Modeled after the Hughes Galaxy system (described later in this chapter), SES sought to create a focal point for the distribution of satellite TV throughout Europe. It began in 1988 with the launch of Astra 1A into the 19.2°E Luxembourg orbit slot to become the distribution point for the Sky TV services in the United Kingdom.

Subsequently, SES added six more Ku-band GEO satellites to complete the plan shown in Figure 1.8, effectively providing enough TV transponder capacity for the entire region. Two major acquisitions boosted SES into one of the world's largest operators: Americom, the U.S. company started by RCA and sold to SES by GE; and New Skies, the spinoff from INTELSAT that SES purchased through a stock transaction.

2.2.3   Domestic Systems: Telesat, Westar, and Palapa

The technology and economics of commercial satellite communication were established during the expansion of INTELSAT. However, the real opportunities for business development appeared as domestic satellite systems came into being during the 1970s, multiplied in the 1980s and defined the marketplace by 2000. New operators that based their businesses on commercial services and customers grew rapidly, extending from developed regions to the rapidly developing parts of the world. The history and worldwide activity of domestic satellite systems are reviewed in the following sections.

2.2.3.1   Canada

The first country to proceed with its own domestic satellite system was Canada, which contracted with Hughes Aircraft for a 12-transponder HS-333 satellite similar in capability to Intelsat IV but weighing half as much. The spacecraft launched in 1974 and shown in Figure 2.8, had a beam that produced a single Canadian
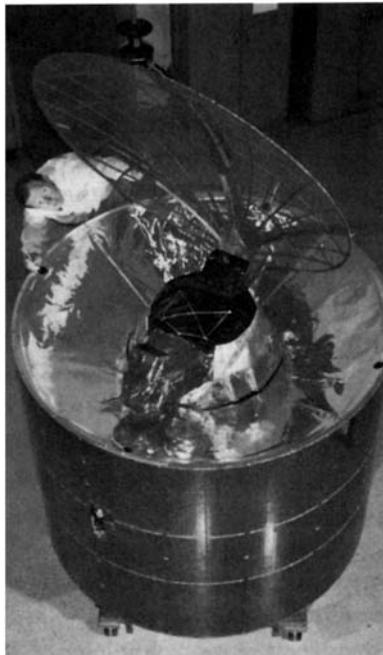
Figure 2.8   The Hughes HS-333 spacecraft. (*Source:* Courtesy of Hughes Aircraft Company.)

footprint. A company called Télésat Canada was established in much the same way as COMSAT; however, its purpose was to offer satellite communication services within Canadian boundaries. The system has proven useful in extending reliable communications into the remote and northern reaches of Canada as well as demonstrating the attractiveness of using a domestic satellite to distribute television programming. The name *Anik*, meaning *little brother* in the native Inuit language, was adopted for each of the satellites. Télésat went on to introduce Ku-band into domestic service with Anik B, built by RCA Astro Electronics, and Anik C, built by Hughes Aircraft. The original Anik A satellites were replaced in 1983 with C-band 24 transponder satellites named Anik D, built by the team of Spar Aerospace Limited of Canada and Hughes Aircraft. Two Anik E hybrid C- and Ku-band spacecraft were launched in the 1990s.

The Anik F series was developed in the early 2000s to expand the capacity, coverage and frequency spectrum usage of the Telesat fleet. Anik F1 and F2 are Boing 702 series satellites and provide C-, Ku-, and Ka-band capabilities. An artist's rendering of the F2 satellite in orbit is shown in Figure 2.9. Anik F2R and F3 are comparable, but built by EADS Astrium. Telesat also engages in the consulting business and has assisted many newcomers to the industry in acquiring and operating their first satellites. The ownership of Telesat changed as leading investor, Bell Canada Enterprises, agreed in 2006 to sell its stake to a Canadian government pension group and Loral Space and Communications.

2.2.3.2 U.S. Domestic System

The policy issues in the United States took longer to settle, and it was the administration of President Richard M. Nixon that allowed the industry to start through the "Open Skies" policy of 1972. That meant simply that any U.S. company with the financial resources and demonstrable need for telecommunications could be granted

Figure 2.9   Anik F2 (Telesat Canada).

the requisite authorization to construct a system and operate its satellites in orbital slots assigned by the FCC. The coverage footprint of most U.S. domestic satellites is similar to that shown in Figure 2.7.

Six companies applied to the FCC under the Open Skies policy and one of them, the Western Union Telegraph Company, was the first to proceed. Their first two Westar satellites, of the same Hughes Aircraft design as Anik A, went into operation in 1974, augmenting Western Union's existing nationwide terrestrial microwave system. Only slightly delayed were COMSAT and AT&T with their Comstar satellites, which provided 24 transponders using polarization frequency reuse and a spacecraft based on the Hughes Aircraft Intelsat IV-A. This system, also used by GTE, provided long-distance telephone circuits to expand the nationwide network and offered both diversity and flexibility in meeting a variety of future requirements. Replacement of three Comstars was accomplished in the early 1980s with satellites of the Telstar 3 series, built by Hughes Aircraft but owned and operated by AT&T alone. GTE went ahead with its own successor system; two other companies, American Satellite Corp. and Spacenet, entered the fray as well.

The polarization frequency reuse technology of Intelsat V and Comstar was adopted by RCA American Communications (Americom), now a subsidiary of SES. RCA had pioneered three-axis stabilization for LEO weather satellites and adapted the technology to a GEO communications spacecraft. Designed and built by RCA Astro Electronics, of Princeton, New Jersey, the spacecraft was particularly compact and efficient; it achieved the capability of Comstar with significantly less weight and volume. That allowed the spacecraft to fit on the Delta instead of the much larger and more costly Atlas Centaur. The Satcom series of satellites, first launched in 1976, proved to be particularly popular for video programming distribution to cable television systems, a market pioneered by HBO, a subsidiary of Time Warner. Americom continued to launch satellites, including in 1985 two of their Satcom K and in 1996 their GE series (now designated AMC).

The U.S. market underwent further expansion in terms of the number of operators and satellites. With Hughes, AT&T, and GTE establishing themselves as powerful forces, the market became nearly glutted with C- and Ku-band transponders by 1987. That proved to be too much for Western Union, which decided to exit business by selling its in-orbit satellites to Hughes. Another satellite venture, SBS, the Ku-band pioneer in the United States, tried to gain a foothold in the rapidly growing field of private corporate telecommunications networks. However, being ahead of its time did not prove adequate for financial success, and SBS, too, folded its satellites into the Hughes Communications constellation. The AT&T Telstar system was acquired by Loral and became the core of its entry into the business of operating satellites as well as construcing them.

As discussed previously, Intelsat acquired PanAmSat in 2006 and as a result became the largest provider of GEO satellite capacity in the United States. Americom likewise continues in a strong position as a part of SES. Loral is maintaining directed satellite service capabilities through its Skynet subsidiary. In addition, Echostar is offering up a portion of its FSS Ku-band capability to users in the United States, while both Telesat and SatMax are serving U.S. customers as well. This has made the United States the most dynamic and efficient satellite service market in the world.

2.2.3.3 Indonesia

The year 1976 witnessed the introduction of a major domestic satellite system in the Republic of Indonesia. Implemented in only 18 months under the direction of the late Mr. Soehardjono, director general of Post and Telecommunications, it used two Hughes Aircraft satellites, which were of the same design as Westar and Anik A. A ground segment of 40 Earth stations was installed and turned up in parallel, allowing President Suharto to demonstrate its operation to his people on the thirty-first anniversary of the independence of the Republic. The Palapa system was named for a mythical fruit that the ancient king, Gadjamada, pledged not to eat until all of Indonesia was united.

The project was completed on time and within budget by a coordinated team consisting of Perumtel of Indonesia (now known as PT Telkom), Hughes (the spacecraft supplier, system architect, and integrator of 10 of the initial 40 Earth stations), ITT (the telephone switch supplier and integrator of 15 of the initial Earth stations), Ford Aerospace (the integrator of the remaining 15 stations), and TRT (supplier of all terrestrial tranmission links). This integrated system, illustrated in Figure 2.10, provided telephone and television service throughout this archipel-ago nation spanning approximately 5,000 km. The cost of implementing and operating the system proved to be a small fraction of what it would have taken to do an even lesser job with terrestrial approaches.

The system moved into its second generation in 1984 with larger Palapa B satellites, each with twice the capacity of the Palapa A. The second Palapa B failed to reach proper orbit, was recovered by the Space Shuttle, and subsequently was replaced in 1987 by a similar spacecraft. PT Telkom was privatized and still provides most of the domestic telephone service in Indonesia. Another satellite operator,
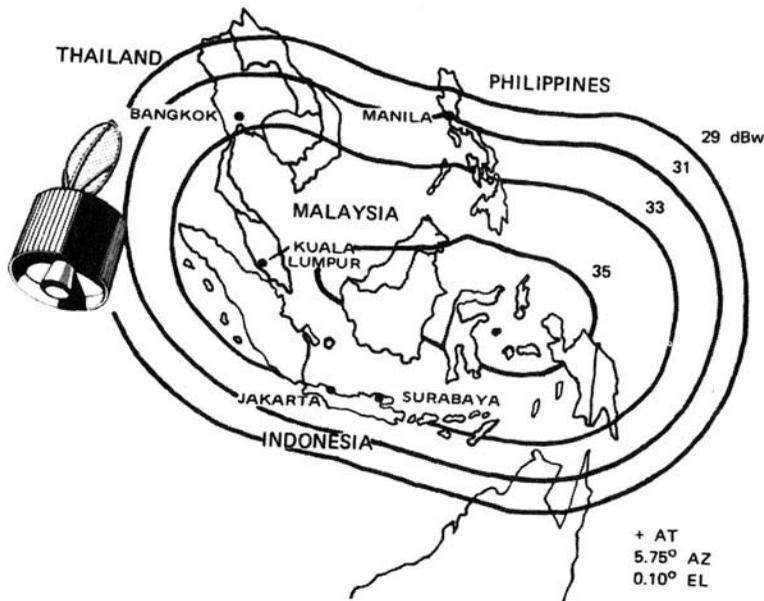
Figure 2.10   The Palapa A system.

Satelindo, continued with the Palapa C series, which is composed of hybrid C- and Ku-band satellites. For financial and other reasons, PT Indosat acquired Satelindo and continues as the Palapa operator. PT Telkom, on the other hand, has acquired some of its own satellites, using the names Telkom 1 and 2 rather than Palapa.

2.2.3.4 Applications Technology Satellite

As the U.S. government agency responsible for space system research, NASA has been involved in several geostationary communication satellite projects since SYNCOM. The Advanced Technology Satellite (ATS) series consisted of six satellites that were used for communications demonstrations throughout the United States and the world. Hughes Aircraft built ATS-1 through 5 to demonstrate a variety of technologies and applications. ATS-1 had a VHF repeater that was used for basic communications on a push-to-talk bais (like a taxi radio). However, the satellite provided some essential services in the Pacific for medical consultation, experimental courses, and emergency communications. Services were maintained through 1983 with ATS-3.

ATS-6 (also known as ATS-F), the most advanced and the largest of the six, was built by Fairchild Industries of Germantown, Maryland, and was launched in May 1974 directly into synchronous orbit by a Titan 3C rocket. A rendering of ATS-6 is shown in Figure 2.11. With its 9-m unfurlable mesh antenna, ATS-6 was used in the first direct broadcasting experiments, with the emphasis placed on educational programming. After providing service in the United States to isolated regions, the satellite was moved into position to be used in India.

Another NASA experimental satellite program called Communications Technology Satellite (CTS), also named Hermes, was jointly undertaken with the Canadian government. Designed and built in Canada but launched in 1976 from the United States on a Delta, CTS operated at Ku-band (14/12 GHz) and demonstrated the capability of high-power satellite broadcasting in this frequency band. One-way and two-way services were provided with small-aperture ground antennas. Of particular note is that CTS was used in the restoration of communications during the Johnstown, Ohio, flood disaster in 1977, proving the effectiveness of compact, transportable ground stations operating at Ku-band.

NASA's more recent experimental satellite was ACTS, which was built by a team of GE Astro, Motorola, COMSAT, and TRW. As with previous NASA communication satellite experiments, ACTS set out to explore new applications in an underdeveloped part of the radio spectrum. In this instance, it was Ka-band. The unique feature of ACTS is that its communication payload incorporates onboard processing and steerable, spot-beam antennas that allow routing of signals within the spacecraft repeater. The Ka-band frequencies used on ACTS are new capabilities for U.S. communications satellites (the Japanese, however, were first to launch a 30/20 GHz satellite). The footprint of various spot beams produced by the ACTS satellite are shown in Figure 2.12. The ACTS system was made available for experiments by industry, universities, and other government agencies as well as for demonstrations that evaluate new service applications.
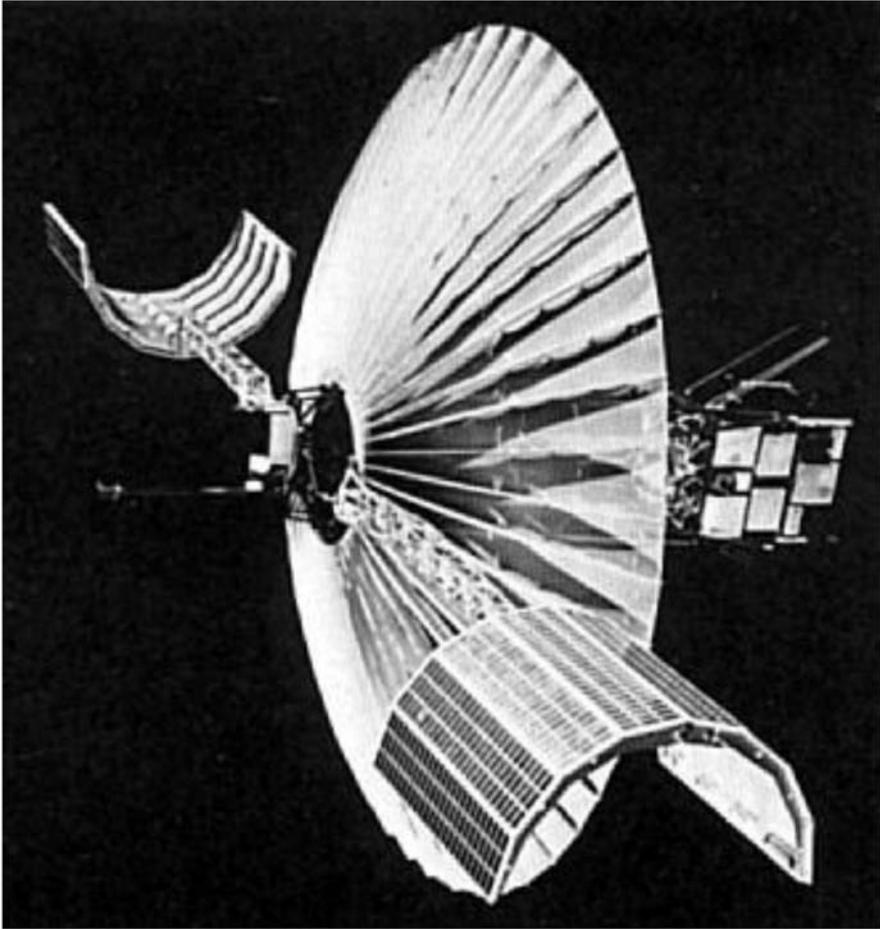
Figure 2.11   The ATS-6 satellite in orbit. (*Source:* NASA.)

2.2.3.5   Japan

As the leading industrialized country in Asia, Japan has aggressively sought to be a power in the field of satellites and satellite communication systems. The Japanese government funded several satellite projects for experimental and industrial development purposes.

The first true Ku-band direct broadcast satellite (DBS) satellite was developed by GE for the national broadcast network, NHK. There were technical problems with high-power traveling wave tube amplifiers (TWTA) and their associated power supplies, so the spacecraft failed to meet expectations. A follow-on DBS program in Japan included satellites designed and built by RCA Astro Electronics and Hughes. Interestingly, the Japanese government became one of the most successful DBS operators in the world (in terms of the number of viewers) by offering a package of U.S. programming material. Regular cable channels like ESPN and Financial News Network (now merged into CNBC) and TV networks like ABC and CBS were combined into a continuous feed of non-Japanese material that was available only on DBS. Another unique feature of the system is that it was used
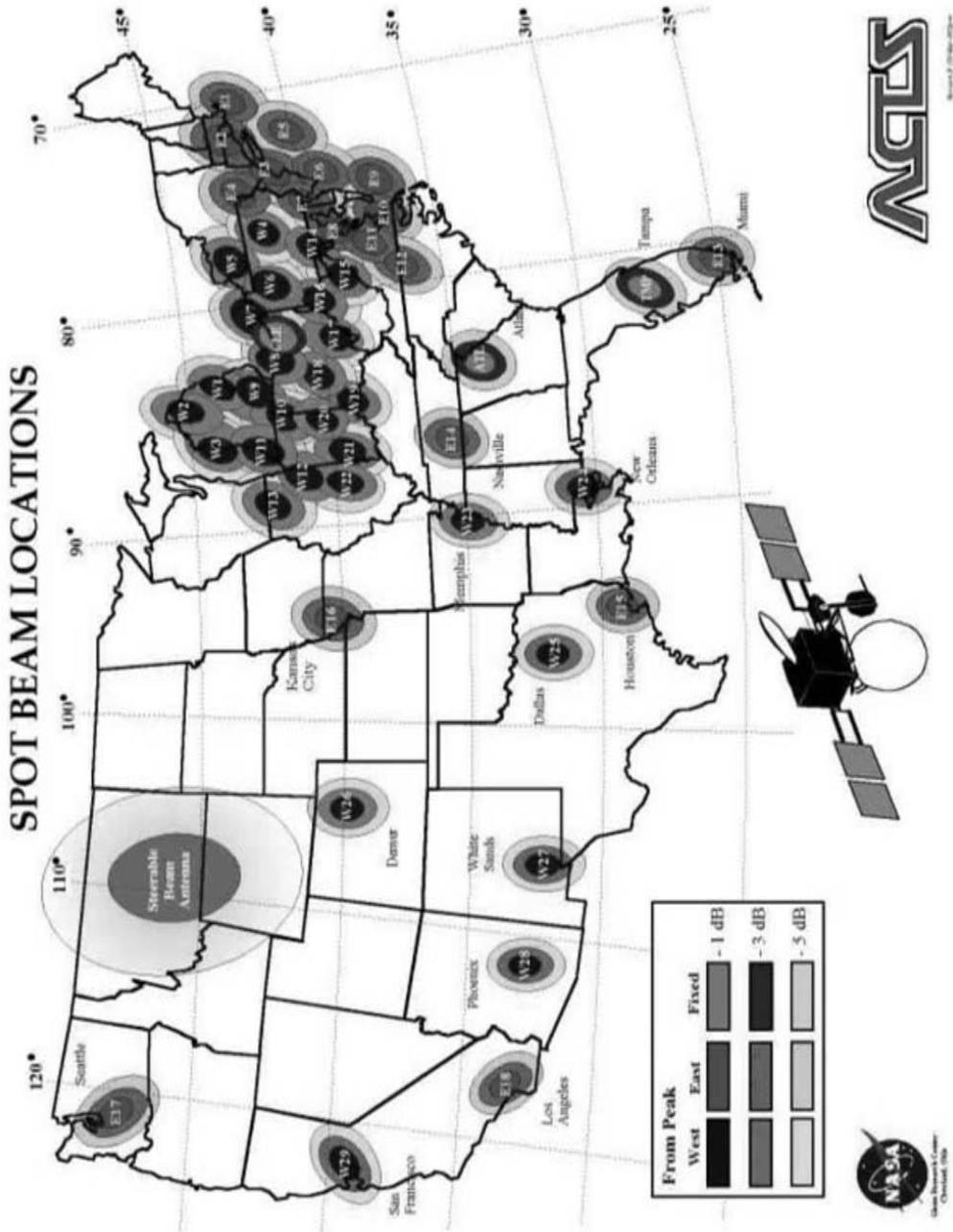
Figure 2.12 An illustration of the spot beams produced by the ACTS satellite, using fixed and switched Ka-band channels. (*Source:* NASA.)

to transmit high-definition TV (HDTV) on a trial basis. The NHK channels were joined by private service providers who used the same space segment to deliver subscription services.

Satellite communications began to be privatized in 1987, when licenses were granted to Japan Communications Satellite Corporation (JCSAT) and Space Communications Corporation (SCC). JCSAT was an extremely interesting company because of its ownership: 40% Itochu, 30% Mitsui and Company (both highly respected Japanese trading companies), and 30% Hughes. The combined company successfully launched two Ku-band satellites and largely cornered the domestic market. These Ku-band satellites became the delivery system for Sky PerfecTV, a successful Japanese DTH operator. SCC, heavily backed by Mitsubishi, the largest trading group in Japan, was plagued by problems with its first satellites supplied by Ford Aerospace (later Space Systems/Loral) but eventually went into operation using satellites built by Loral and Hughes. A third satellite operator, Satellite Japan, never went into business, but the partners ended up buying out Hughes' interest in JCSAT (renamed JSAT). Subsequently, JSAT has grown into the largest satellite operator in Asia and was renamed Sky Perfect JSAT.

2.2.3.6 India

India has conducted an extensive scientific and technological research program in space, under the stewardship of the Indian Space Research Organization (ISRO). Headquartered in Bangalore, India's fifth-largest city and the capital of the southern state of Karnataka, ISRO has created a domestic industrial base capable of designing and building spacecraft and placing them in orbit with an indigenous rocket. The first domestic satellite for India, called INSAT 1, was launched in 1981 and performed the dual function of providing communication services at C- and S-band (for TV broadcast) and meteorological imaging. The first-launched of these complicated three-axis spacecraft, designed and built by Ford Aerospace, reached orbit but experienced technical problems with deployment of antennas and operation of propulsion. A modified second spacecraft was launched in 1983 and met its service objectives.

ISRO successfully transferred technology to develop INSAT 2, the second-generation spacecraft that continued to offer domestic transponders and weather imaging. A trend toward limited privatization of telecommunications started in 1995, and other satellites were introduced into the domestic market.

2.2.3.7 Middle East

After many years of planning and discussion, the Arab League countries embarked in 1976 on a regional system called the Arab Satellite Communications Organization. It was structured along the lines of EUTELSAT, with membership including Algeria, Bahrain, Egypt, Iraq, Djibouti, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia (the largest and most powerful member), Somalia, Sudan, Syria, Tunisia, the United Arab Emirates (UAE), and Yemen. The satellite and system were to become known as Arabsat.

Design and construction of the first series of satellites were contracted to Aerospatiale, with a significant portion of the three-axis technology provided by Ford Aerospace. The satellites, first launched in 1985, had communication capability similarly to Insat in that C- and S-band services are possible. What ensued was a number of years of technical and political difficulties, which negatively affected the initial business prospects. Eventually, Arabsat developed a more functional and businesslike structure and became profitable. Arabsat 2B, for example, provides TV and telecommunication services at 30.5° East. All services are delivered at Ku-band, including DTH television.

A number of other companies in the Middle East have embarked on their own domestic and regional satellite programs. Nilesat 901 was built by EADS Astrium for the Egyptian Satellite Company. Turksat, from Turk Telekomunikasyon AS of Turkey, was launched in August 1994, providing Ku-band capacity to the region. Another program of interest is Thuraya, a GEO mobile system supplied by Boeing. Thuraya was developed by Etisalat, the public telecommunications operator in the UAE, and a consortium of telecommunication operators in the Middle East. Capacity is also provided to the Middle East by such established operators as INTELSAT, EUTELSAT, and AsiaSat. A cooperative spacecraft program called Amos 1 involved the Israel Aircraft Industries along with Daimler Aerospace and Alcatel. This Ku-band spacecraft was successfully launched on May 15, 1996, on an Ariane 4 rocket. In addition to a Middle East beam, Amos 1 has a Central Europe beam that delivers HBO programming to Hungary, Poland, and the Czech Republic. Amos 2 and 3 have since been launched and colocated with Amos 1 at 4°W.

2.2.3.8 Latin America

Latin America embraced satellite communication almost immediately after the launch of Early Bird. With vast expanses of underdeveloped territory and difficult natural obstacles between its countries, Latin America proved to be a loyal supporter of the INTELSAT system. Mexico (technically in North America but close to South America by virtue of a common heritage) and Brazil (the largest and most industrialized economy in the region) were among the first in the world to construct Standard A Earth stations near their capital cities. Later on, Mexico and Brazil initiated their satellite systems in 1982 using Hughes Aircraft designs. In the case of Brazil, the first two satellites were assembled in Canada by Spar (now MacDonald, Dettwiler and Associates, Ltd.) and provided 24 C-band transponders each. Each satellite incorporated a hybrid C- and Ku-band payload, intended to provide a variety of communications services to large- and small-diameter Earth stations. One important Ku-band application was the extension of tele-education to remote villages in Mexico, something that was previously started with terrestrial microwave.

Both Mexico and Brazil (Brasilsat) were successful in implementing their space segments in 1985. Other Latin American nations have taken action to develop their own orbit resources. Argentina initiated its own satellite program called Nahuel in the 1990s, contracting with DASA for the system. The unique aspect of Nahuel is that the contract grants DASA the right to operate the satellite to provide services in Argentina and around the region. Colombia has applied for

orbit positions beginning in 1980 but had not initiated construction of its own satellite at the time of this writing. Another regional project named Simon Bolivar (later changed to AndeanSat) has been under development by the nations of the Andean Pact, which includes Colombia, Venezuela, Ecuador, Peru, and Bolivia. Venezuela may undertake this project by employing satellite and launch technology from China.

In the 1990s, Mexico and Brazil were in a position to maintain and expand their respective satellite systems. Telecomunicaciones de Mexico (Telecomm), the government agency responsible for Morelos, contracted with Hughes Space and Communications International for two medium-power hybrid spacecraft, Solidaridad 1 and 2. An Ariane 4 rocket launched Solidaridad 1 on November 20, 1993, resulting in the spacecraft being placed at 109.2 degrees west longitude. The satellite was the first to carry L-band in addition to C- and Ku-bands. The Mexican government spun off the satellite business as SatMex, which was acquired by Telefonica Autry and Loral Space and Communications. Embratel, the Brazilian long distance telephone company and operator of the Brasilsat system, was sold to MCI of the United States in 1998. Brasilsat, on the other hand, was purchased by SES and renamed StarOne.

2.2.3.9 Africa

The African continent, with the exception of the Middle East, still remains the area with the lowest level of economic and telecommunications development in the world. However, it cannot be ignored when considering markets for satellite capacity and, ultimately, domestic spacecraft. Telephone density in the region is the lowest in the world at around 5 per 100, although an acceleration is under way through installation of new lines and cellular coverage.

African nations have long been members and users of the INTELSAT system. That contributed to the telecommunications development of many countries, allowing the more politically stable foreign investment in natural resource and transportation industries. The pooling of demand for those transponders led to the creation of Regional African Satellite Communications (RASCOM), a 42-member cooperative of public operators. RASCOM-QAF1 was contracted to Alcatel, but the launch in 2008 was unsuccessful. Meanwhile, Nigeria lofted its first domestic satellite in 2007.

2.3 Specialized Systems: DTH and Mobile

Satellites that employ the FSS spectrum allocations are intended for a variety of telecommunication services that can satisfy basic infrastructure needs of a country or a region. Such systems are employed on a global basis and form the core of the industry. In time, existing and new satellite operators saw an opportunity to create satellite-based business in the areas of DTH television and mobile communications. As discussed in detail in our other work [2], these systems reached operational status in the 1990s and are today the fastest growing segments. We review the systems that took the risk of being first to market.

2.3.1   DTH Development

The concept behind DTH is to deliver TV and other broadcast services directly to the individual user's home or portable device. Doing that required a step improvement in satellite performance to reduce the size and the cost of the user installation. In addition, the service must be controlled through scrambling or encryption so that users are encouraged to pay for services. The other key ingredient is the content, which must be attractive enough (and affordable) so that people acquire the equipment and subsequently pay for the service. The ITU initiated the flurry of interest by conducting a landmark World Administrative Radio Conference (WARC), as the WRCs were known prior to 1995, for broadcasting satellites. The conference was held in Geneva in 1977, at a time when only NHK of Japan had initiated a DTH program. WARC-77 made assignments of Ku-band channels and GEO positions to ITU members in Regions 1 and 3 (Region 2 completed action on its assignments at the 1983 Regional Administrative Radio Conference).

Several U.S. companies approached the DTH market in the early 1980s, but unfortunately none survived long enough to establish a permanent service. COMSAT was the most ambitious, forming Satellite Television Corporation (STC) to utilize the new assignments from WARC-77. They contracted with RCA Astro in 1983 for two DBS satellites (small by today's standards but fairly sizable in their time). However, they could not put a viable business together, lacking progamming resources from a TV network partner. Another startup, United States Satellite Communications Inc. (USSCI), obtained FCC approval in 1983 to use the Canadian Anik C satellite to deliver a limited DTH service to a segment of the U.S. market. USSCI ceased operation in 1985, when it ran out of money before achieving adequate revenues. By 1986, the real development of DTH in the United States came in the form of the C-band backyard dish business as an adjunct to normal cable TV program distribution. The large networks, like HBO, CNN, and Disney, sought to increase their revenues overall by gaining backyard dish subscribers. That involved the introduction of scrambling and conditional access control, two key technologies that any DTH system demands. A firm foothold in the U.S. DTH market was eventually established by DIRECTV, Inc., a wholly owned subsidiary of Hughes Electronics Corporation. They started service in 1994, using an array of 150 digitally compressed TV channels.

DTH expanded aggressively through the late 1990s and early 2000s as Echostar built out its DISH Network and News Corporation acquired and further developed DIRECTV. The two established DTH as a substantial segment of the U.S. home TV market, reaching over 25 million homes out of a potential market of approximately 90 million. Both services added local TV channels from over 50 cities and towns through a combination of nationwide broadcast and spot beams to limit coverage to metropolitan areas. The viability of DTH was thereby established, but it is recognized that the service must take on interactive features to compete effectively with cable TV and fiber to the home (FTTH) services from Verizon and AT&T.

News Corp initiated its first DTH venture in the United Kingdom under the SKY brand. This employed the SES Astra satellite, discussed previously, and started out as an analog service. By offering equipment that was affordable and providing

a unique pay TV package, SKY developed a loyal customer base. The officially mandated U.K. DTH operator, BSB, did not get off the ground until 1989. The prestigious backers of BSB included Pearson, owners of the *Financial Times*, and Reed and Granada Television, famous in the United States for the excellent dramatic series shown on the Public Broadcasting System. The U.K. government awarded BSB the WARC-77 assignments, requiring also that they adopt the D-multiplexed analog components (D-MAC) transmission standard that had been mandated by the European Commission. The ensuing delay in getting its DTH platform, generally regarded as technically superior to that of Sky, off the ground proved to be fatal to the business prospects of BSB. On November 2, 1990, BSB and Sky merged to form British Sky Broadcating (B-Sky-B). The combined service was directed by News Corporation, which held a 40% interest in the company. Subsequently, B-Sky-B reached a high degree of acceptance in the United Kingdom and became a very successful business in financial terms.

DTH systems in Asia were centered in two markets: Japan and East Asia. The Japanese approach to DTH is best illustrated by the efforts of NHK to establish early preeminence in the field. Two other Japanese DTH operators, PerfecTV and DIRECTV Japan, entered the market using FSS Ku-band transponders supplied by JSAT and SCC, respectively. These operators subsequently merged to form SkyPerfecTV. The rest of Asia was to be the broadcasting domain of Richard Lee and his company Satellite Television for the Asia Region (STAR). That venture rode along with AsiaSat 1 when it was launched in 1991. STAR assembled a package of Hong Kong–developed programming channels in English and Mandarin Chinese, along with imported programming from the BBC and U.S. movie studios. The target was primarily the C-band DTH market, which could be developed within the AsiaSat 1 footprint. Richard Lee was definitely the starter-upper, but what really made him famous was the sale of the entire operation and licenses to Rupert Murdoch for over $600 million. The finances of the service were based on advertisers. STAR has shown that there is indeed a place for pan-regional programming and for packages targeted to the attractive country markets.

The main thrust of the News Corp purchase of STAR was to obtain a foothold in the expansive Chinese TV market. However, building a TV business inside the People's Republic of China has remained an extremely difficult task. The national TV scene is dominated by state-run China Central Television (CCTV), which also broadcasts on an international basis. Others have offered programming in Mandarin Chinese on a free-to-air basis. The challenge in China, as it is in India, is to build a pay TV business that would grow as the respective economies. A number of Chinese satellite companies have proposed and subsequently launched Ku-band satellites. It is hoped that one or more of these will be able to establish a viable DTH business.

2.3.2 MSS Development

Development of MSS, introduced in Chapter 1, started with Marisat in 1975 and evolved through the various satellites used for maritime applications. Inmarsat, originally an international organization like INTELSAT, was established in 1979 to provide worldwide mobile satellite communications for commercial, emergency,

and safety applications at sea and in the air. Inmarsat services were extended to land-based users with the Inmarsat M and Mini-M standards. The Inmarsat 4 satellites have sufficient power to support portable user terminals, which in the case of B-GAN are the size of a laptop computer (Figure 1.7).

Domestic MSS satellite services began to appear in the mid-1990s, when Optus Communications decided to include an L-band payload on its second-generation Ku-band satellites. The Optus MobileSat team designed and implemented the ground segment to make mobile telephone calls possible from conventional vehicles. In the United States, AMSC was formed from eight applicants for domestic licenses. The consortium floated its own series of stock on the public stock market to raise sufficient capital to fund a satellite and the associated gateway Earth station. Even with a dedicated high-power satellite, the AMSC network could serve only vehicular and portable terminals (similar to the mini-M terminal of Inmarsat). It would take satellites either much higher in power or, alternatively, closer to the Earth to close the link to a hand-held type of MSS telephone instrument.

The real excitement in mobile satellite services was created by Motorola when it announced the Iridium project in 1990. As discussed in Chapter 1, Iridium employs a LEO constellation of 66 satellites (the original proposal was for 77 satellites, hence the selected name Iridium, which is the element in the periodic table with 77 orbiting electrons). Loral followed suit with Globalstar, which was based on CDMA technology from Qualcomm.

The Iridium space segment, which is still in operation at the time of this writing, employs satellites similar in design to GEO satellites, with the exception of its smaller size and configuration adapted to the orbit and operational requirements. Built by the team of Motorola, Lockheed Martin, and Raytheon, the Iridium satellites demonstrate the principle of nearly continuous coverage of the entire globe in the delivery of a packet-switched voice and low-speed data service. The intersatellite links, called cross links, add the benefit that a user with a handheld Iridium subscriber unit can communicate directly with another such user who can be located anywhere else on the Earth's surface. Some of the same capabilities were instilled in the Globalstar system, developed by Loral and Qualcomm. In this case, the satellite is a bent-pipe design without the onboard processing and switching found in Iridium. Using CDMA instead of TDMA, Globalstar has some added features that make the system more flexible in terms of services within its coverage footprint. Another LEO system called Orbcomm also went into operation this time for the purpose of providing low-speed data relay.

The entire MSS business went through an upheaval during the late 1990s and early 2000s, caused to a significant extent by the bankruptcy of Iridium, Inc., in 1999. Globalstar, which completed their system as well, was also the subject of bankruptcy proceedings. Both systems have come out of bankruptcy with new owners who offer services to a much smaller customer base than originally defined. Meanwhile, MSS satellites to serve handheld devices were produced for regional systems in the Middle East (Thuraya) and Asia (ACeS). Inmarsat developed the more powerful and capable Inmarsat 4 series of satellites (Figure 2.13) that provide both Internet access and ISDN services to portable and mobile users.

Another development on the MSS front was the creation of a hybrid concept of satellite and cellular-type wireless communications in the same band, called
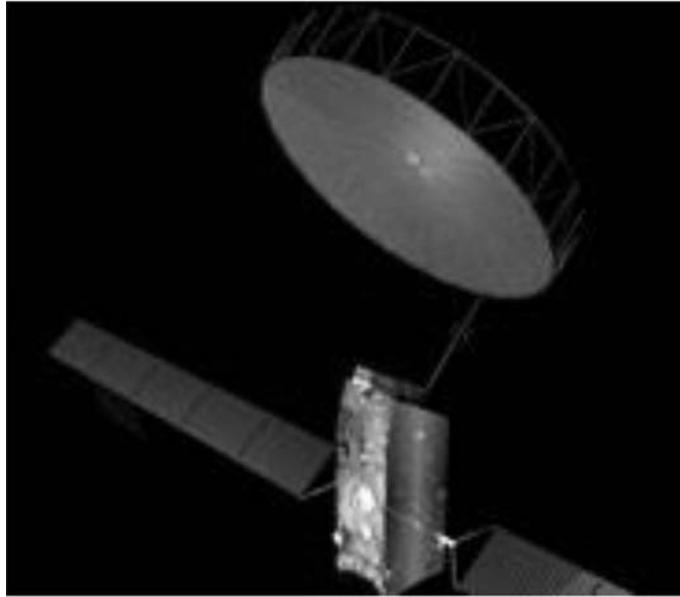
Figure 2.13   Inmarsat 4. (*Source:* Inmarsat.)

ancillary terrestrial component (ATC). The first to obtain approval for such a system was Mobile Satellite Ventures (MSV), an outgrowth of the original AMSC. Their system employs the same L-band spectrum as previously authorized but includes cell sites around the United States that could provide a broadband wireless service in conjunction with the satellites. Other companies to pursue this concept are the new ICO Communications (also a company recovered from bankruptcy by investors) and TerreStar. Both of these systems employ S-band spectrum authorized by the FCC. An illustration of the type of satellite required for ATC is shown in Figure 2.14. In Europe, a similar concept called complementary ground component (CGC) may also be developed using MSS spectrum.

2.3.3   Digital Information Broadcasting

There remains one other application segment for satellite communication, that of providing direct to user transmission of information. Digital audio radio service (DARS) represented an interesting possibility for a significant domestic or global market. Experiments were conducted by NASA using the S-band payload of the tracking and data relay satellite (TDRS). The experiments proved that digitized radio signals can be transmitted to vehicular terminals with a special antenna and receiver. The difficulty was that terrain blockage caused a break in reception, something that probably would not be acceptable to the typical audience. That was overcome either by transmitting the same signal from multiple diversely placed satellites and through terrestrial broadcast towers that fill in gaps. The XM Satellite Radio system was deployed in the United States using this approach; see Figure 2.15 for a satellite photograph. XM, and competitor Sirius Satellite Radio, established DARS as a powerful offering.
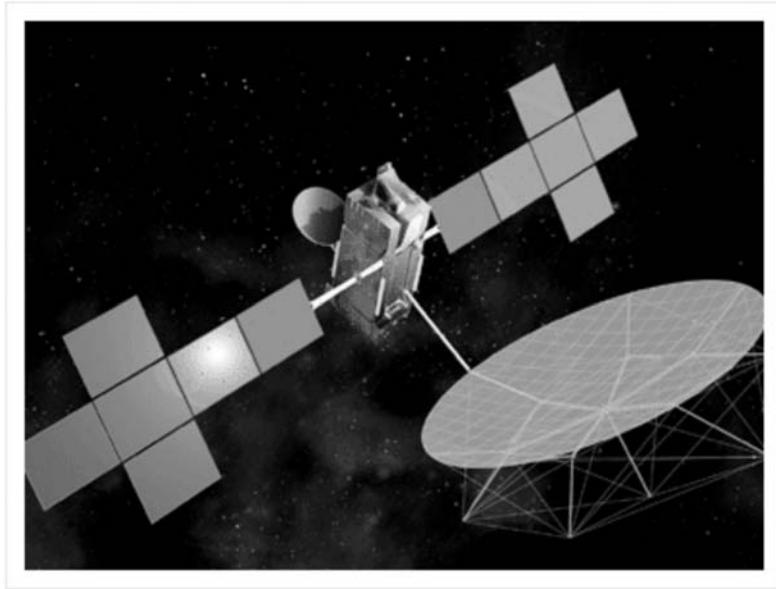
Figure 2.14   ICO ATC satellite. (*Source:* Space Systems/Loral.)

Figure 2.15   XM Satellite Radio spacecraft (Rock and Roll) at Boeing Satellite Systems, and the deployed spacecraft. (*Source:* XM Satellite Radio.)

On the nonaudio side of the picture, several U.S. companies have provided data broadcasting since the late 1980s. Examples include continuous feeds of financial information, like stock market and commodity prices. The typical receiver in that case would be a fixed dish of anywhere from 50 cm to 1.5m and a simple receiver and data demultiplexer. Early participants in the field include Equatorial Corporation, which was first in 1983 to employ spread spectrum for a commercial application, and United Video, which in 1985 used data broadcasting to deliver an online program guide for cable TV before that became a standard facility of

DTH services. The most dramatic development in the field was the DirecPC service from HNS, which allowed a standard PC to receive a 12-Mbps data broadcast through a 60-cm dish.

Data broadcasting has evolved through the facilities of the Internet into a service called digital content distribution. These are private broadcasting systems that typically employ FSS satellites in a point-to-multipoint mode (discussed in Chapter 3). Rather than sending only one kind of information (video, audio, or a particular type of data), a digital content distribution network transfers multiple data and information types. These can be real-time streams of video and audio, text messages, and files of various types. At the remote receiver, the content is displayed and/or stored. A popular use is in consumer retail establishments to provide a medium to entertain and inform customers. Using the Internet Protocol and standard compression and file formats, this type of network can be the medium for updating customer records, software programs, specialized Web sites within a corporate Intranet, and training videos that interact with the student.

2.4   Expansion at Higher Frequency Bands: Ka-Band

Ka-band frequencies long have promised much greater bandwidths than are available at C-, X-, and Ku-bands. Many years of experimentation in Japan, Italy, and the United States have produced a wealth of useful information. NTT of Japan, in particular, was first to employ a commercial Ka-band payload using the CS-1 satellite, also known as Sakura (Cherry Blossom). CS-1, launched in December 1977, was contracted with Ford by the National Space Development Agency (NASDA) of Japan and flew Japanese components made by NEC and Mitsubishi Electric Company. A second generation, comprising CS-2a and CS-2b, was launched in 1983 to maintain a Ka-band capability. Those satellites provided the important feature of connecting Japan's remote islands into an integrated telecommunications network. NTT conducted numerous experiments using fixed and transportable Earth stations in Japan. Later, NTT acquired transponders from SCC and launched its own N-Star satellites to maintain Ka-band capacity to continue its development of applications.

As discussed in Chapter 1, the technical issues at Ka-band are reasonably well understood; however, there remains the challenge of finding applications that can justify making the requisite investments. Almost without regard to the expected challenges, many established and startup satellite operators made application to the FCC and the ITU for Ka-band satellite systems.

Reality set in by 2004, with the result that only three programs actually reached orbit: EchoStar Ka, Spaceway and KaStar (WildBlue). A rendering of WildBlue 1 is presented in Figure 2.16.

The western military organizations have developed specialized satellites with Ka-band as well as X-band payloads. This has the benefit of employing spectrum that was earmarked for government use, particularly in the United States, Europe, and Russia. A series of X-band satellite programs evolved from the Defense Satellite Communications System (DSCS) built by TRW in the 1970s followed by GE (now Lockheed Martin) in the 1980s and 1990s. Boeing produced a new series of large

Figure 2.16   WildBlue 1. (Courtesy of Space Systems/Loral.)

satellites to replace the DSCS series and add government Ka-band capabilities. These wideband global SATCOM (WGS) satellites, illustrated in Figure 2.17 and discussed in Chapter 7, were first launched in 2007 and offer varied capabilities using phased arrays and digital processing technology. Also incorporated is the global broadcast system (GBS) that was first flown on the UHF Follow-On satellites operated by the U.S. Navy. Other members of NATO are likely to employ this series as well as other satellites produced in the United States and Europe.

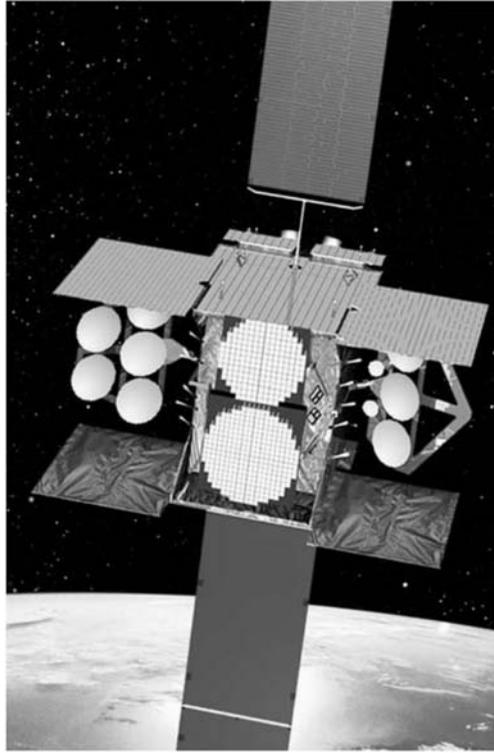**Figure 2.17**   Wideband Global SATCOM (WGS) satellite built by Boeing and first launched in 2007. (*Source:* Boeing.)

References

[1]   Labrador, V. S., and P. I. Galace, *Heavens Fill with Commerce: A Brief History of the Communications Satellite Industry,* West Covina, CA: Satnews Publishers, 2005.

[2]   Elbert, B. R., *The Satellite Communication Applications Handbook*, 2nd ed., Norwood, MA: Artech House, 2004.

Satellite Network Architectures

Any communications satellite performs the function of a microwave relay in a telecommunications network. The most basic type of relay is the bent-pipe style of satellite, which does not alter the nature of the transmissions between Earth stations. The Earth stations in this case must organize their transmissions to efficiently use the satellite's repeater resources and to allow the various points on the ground to transfer or exchange the required information. That then determines the basic network structure for the particular application. In more advanced repeaters with onboard processing, the satellite performs other network functions such as switching the connections between stations or even modifying the format of transmission to further improve the efficiency of transmission. All that is carried out whether there is a single GEO satellite or a constellation of non-GEO satellites serving the network. From the users' perspective, the specific nature of the space segment may be irrelevant. However, the design, operation, and, more important, the required investment, all depend heavily on the way the architecture is put together.

The first part of this chapter reviews the features and generic arrangements of networks independent of specific uses. This provides a cross-reference with regard to the applications that are reviewed at the end of the chapter. More detail on these applications can be found in our other work [1].

The purpose of this chapter is to introduce readers to satellite communications, so we have taken a nontechnical approach in reviewing the range of network architectures and services. Chapter 4, the first technically oriented chapter, explores the fundamentals of satellite microwave links, including such important principles as the decibel (dB), radiowave propagation, and the link budget.

3.1 General Features of Satellite Networks

The property which a satellite network provides via links between users is called *connectivity*. The three generic forms of connectivity are point-to-point, point-to-multipoint, and multipoint interactive, shown in Figures 3.1, 3.2, and 3.3, respectively. The first uses of satellites were for point-to-point links between fixed pairs of Earth stations. As shown in Figure 3.1, communication from one station to the other is on a dedicated path over the same satellite. Therefore, two links are needed to allow simultaneous communication in both directions. That is how the typical telephone conversation or interactive data link functions. The link can

Satellite repeater
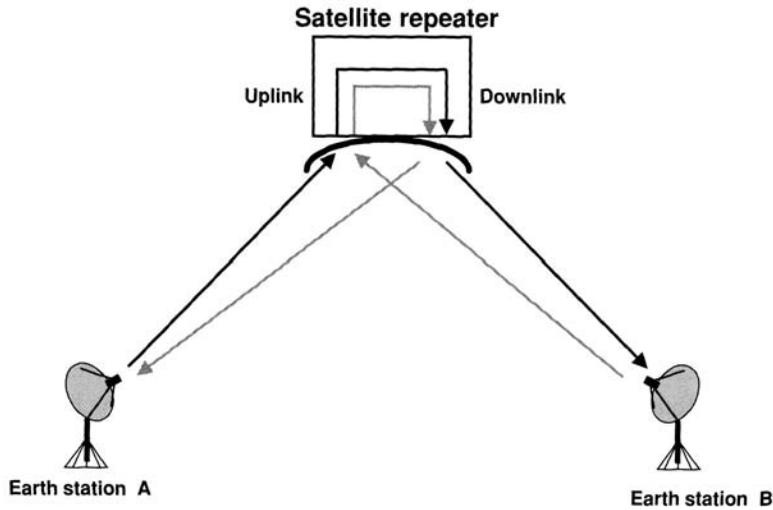
Figure 3.1   The most basic two-way satellite link provides point-to-point connectivity.
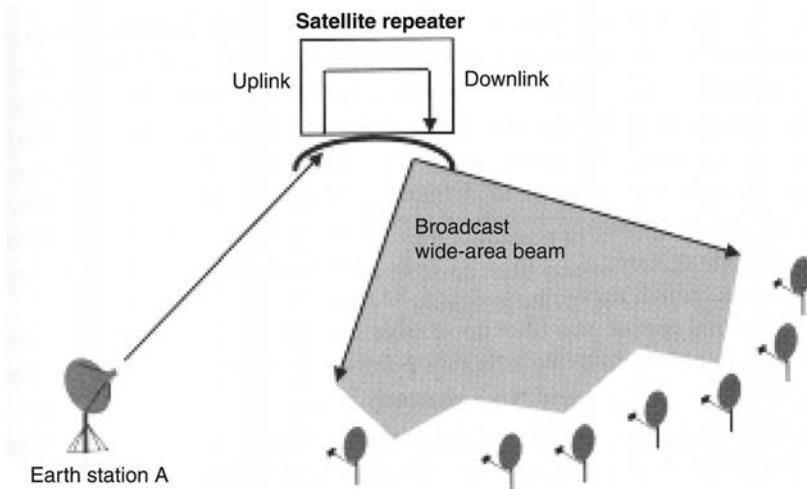
Figure 3.2   Point-to-multipoint (broadcast) connectivity delivers the same information over the satellite coverage footprint.

remain in place for an extended period, called preassigned service, or be put into place for the duration of a conversation, called demand-assigned service.

The point-to-point link allowed satellites to create a worldwide network of telephone circuits, providing a solid foundation for international telecommunications development. That was before the days of high-capacity fiber optic cable, a technology that pushed satellites out of this role. Since the 1980s, satellites are used largely for point-to-multipoint connectivity, also called broadcast. As shown in Figure 3.2, a single uplink station can transmit a continuous stream of information to all receiving points within the coverage area. Satellites like Galaxy 1R in the United States, Astra 1 in Europe, and JCSAT in Japan all use that broadcast
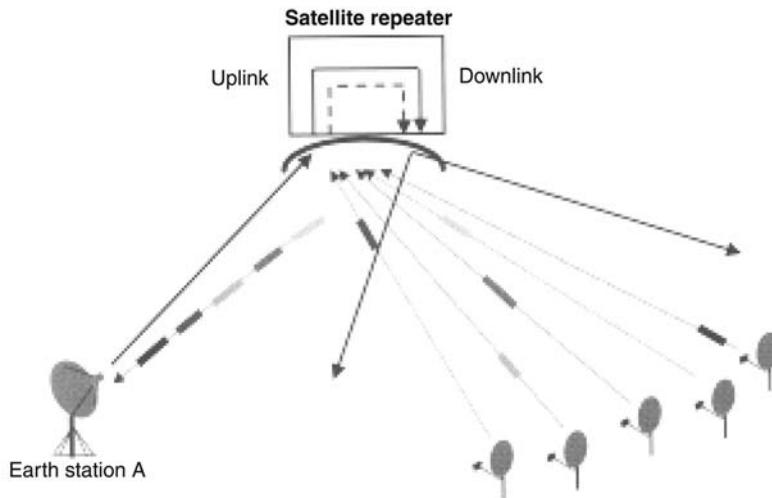
Figure 3.3   Multipoint-to-point (interactive) connectivity supports a star network of low-cost VSATs.

mode to transmit television programming to large user populations, thus using the wide-area coverage of a GEO satellite to its greatest extent.

Adding a transmit capablity to each receive terminal is a technique to convert the broadcast system into a multipoint interactive network, shown in Figure 3.3. Each remote site employs a type of Earth station called a very small aperture terminal (VSAT). The broadcast half of the link transmits bulk information to all receiving points. Those points, in turn, can transmit their individual requests or replies back to the originating point over the same satellite. The value of this approach is at its highest when most of the information transfer is from the large Earth station, shown on the left in the figure. For example, a subscriber might request to download an information file containing a software program or short movie, all in digital form. That request is received by the large Earth station, which then transmits the file over the broadcast link. Files and streaming video can be broadcast to many VSATs using the multicast feature of the Internet Protocol.

Terrestrial copper and fiber optic cable networks can create all those connectivities, but they do so by stringing point-to-point links together. Satellite networks, on the other hand, are inherently multipoint in nature and have a decided advantage over terrestrial fixed networks whenever multipoint connectivity is needed. It depends on many variables to determine if this advantage exists in a particular example. A key variable is the capacity of the link, commonly called the *bandwidth*. In digital communication, bandwidth is measured in bits per second (bps). What controls those variables are the particular aspects of the telecommunications application that the satellite network delivers. The rest of this section reviews many of those variables, relating them to how they are addressed in a particular satellite network design.

3.1.1   Dedicated Bandwidth Services

The commercial history of satellite communication is founded on the use of the GEO, in which satellites and Earth stations are fixed with respect to each other

so that link direction and length do not change. The principal approach to satellite links and services is to dedicate the bandwidth to each transmission service. In this mode, the communication is established and maintained continuously for an extended period of time. The user must make a request to the network operator well in advance of need; that gives the operator enough time to plan ahead for the required satellite bandwidth and associated ground resources. Bandwidth usually is assigned as fixed time slots of a specific duration. They may repeat periodically to provide a near-continuous channel of communications, which is required for voice and video communication.

Figure 3.4 shows a basic arrangement for a dedicated bandwidth service. Each transmission comes from a particular Earth station and is provided on a noninterfering basis with respect to all other Earth stations on the same satellite. As discussed in Chapter 5, techniques of multiple access are used to prevent interference between transmissions. The transmission is formed either as a continuous stream or a noncontinuous burst signal. It may provide a constant bit rate (CBR) type of service, suitable for literally any communication application. Voice and conventional TV give optimal performance under CBR transmission; for intermittent services like electronic mail (e-mail), CBR is much less beneficial. Once established, the CBR link is maintained for an extended duration, sometimes measured in years for a cable TV channel or hours for a temporary wideband data channel to transfer a large file.

Dedicated bandwidth links are either analog or digital in nature. Originally, all satellite tranmissions were analog and the bandwidth was fixed at time of installation. The introduction of digital transmission permitted the bandwidth to be rearranged more conveniently, which was one of the big reasons for making the conversion. A given link could support a single digital data stream of 64 kbps, which is sufficient for a standard telephone channel or a combination of low-speed data channels. More likely, many 64-kbps channels are assembled into a higher speed stream through the technique of time division multiplexing (TDM). TDM is different from TDMA, in which the transmission is made as a burst to allow
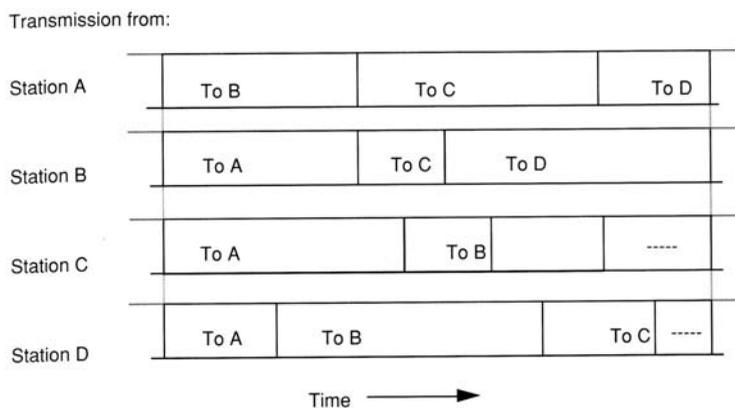

Figure 3.4 Transmission timelines for multidestination TDM carriers that connect four stations, using point-to-point links. This arrangement provides full interconnectivity among the four stations.

several stations to share the same RF channel. All these aspects are covered in Chapter 5.

TDM channels are specified in terms of a digital hierarchy that follows standard bandwidth group sizes. Table 3.1 lists the familiar North American T and European E hierarchies, respectively. One T1 provides 1.544 Mbps of data transmission, 24 individual 64-kbps user channels, or an appropriate mix of both. Likewise, the European E1 channel is 2.048 Mbps in capacity, suitable to transmit 30 user channels at 64 kbps each. While the T and E hierarchies are somewhat different, they both employ the 64-kbps basic channel of communication. It is up to the operator or ultimate user to determine how the channels are applied. For example, if the user is a telephone company, the T1 link can carry 24 simultaneous voice conversations. In a data communication application, the E1 channel is useful in connecting many remote clients to a mainframe computer used as a large server.

Greater bandwidth and service flexibility is offered by the synchronous digital hierarchy (SDH) approach, reviewed in Table 3.2. Although the North American scheme, called synchronous optical network (SONET) is structured slightly differently, the combined system of global digital transport provides a seamless way to connect fiberoptic and advanced broadband satellite networks at the Gbps level.

There are many other ways that a dedicated bandwidth channel is employed. In DTH systems, the capacity of a single link might be 28 Mbps, which is broken down through multiplexing into 5 to 10 digital TV channels with sound and control information. A given DTH satellite with 16 high-power transponders will carry 16 of those data streams, which is equivalent to between 80 and 160 TV channels. With analog TV, the satellite would support only 16 TV channels (e.g., one per

Table 3.1   The Plesiochronous Digital Hierarchy (PDH) as Employed in North America and Japan (1.5 Mbps) and Europe, South America, Asia, and Africa (2 Mbps) (Entries Expressed in kbps)

Level	North America (1.5 Mbps)	Japan (1.5 Mbps)	Europe, South America, Asia, and Africa (2 Mbps)
1	1,544	1,544	2,028
2	6,312	6,312	8,448
3	44,736	32,064	34,368
4	97,728	139,264	

Table 3.2   SONET (North America) and SDH (Worldwide) (Entries Are Expressed in Mbps)

SONET Level	SONET Designation	Bit Rate (Mbps)	SDH Level	SDH Designation
STS-1	OC-1	51.840		
STS-3	OC-3	155.52	1	STM-1
STS-9	OC-9	466.560		
STS-12	OC-12	622.080	4	STM-4
STS-18	OC-18	933.120		
STS-24	OC-24	1,244.160		
STS-36	OC-36	1,866.240		
STS-48	OC-48	2,488.320	16	STM-16

transponder). The dedicated digital bandwidth approach, therefore, is popular for providing many TV channels and the associated sound programming.

One potential difficulty with dedicated bandwidth is that it may require a great deal of planning effort to ensure that the capacity is allocated as efficiently as possible. The network operator must assess: (1) the quantity of individual channels of communication (64 kbps or digital video), and (2) where they are to be provided. The second point means that the Earth stations must be equipped with the proper equipment to land their respective carriers and demultiplex the individual information streams. A way around part of that problem is to use dynamic bandwidth allocation, also referred to as statistical multiplexing, to adjust individual usage of channels to near-instantaneous need. As illustrated in Figure 3.5, statistical multiplex reduces the total peak capacity by using bandwidth that otherwise would have been wasted. Channels 1, 2, and 3 on the left have maximum bandwidth requirements over time with A, B, and C, respectively. Using dedicated bandwidth, the combined channel on the right would require the maximum value of A, B, and C. With statistical multiplexing, the maximum instantaneous bandwidth is Σ, less than what would have been required by adding the three maxima.

There are two ways to apply dynamic bandwidth allocation: time slot reassignment (TSR) and packet switching. In TSR, the fixed total channel capacity of a given link is divided into fixed time slots each corresponding, for example, to an 8-bit byte of capacity. A standard 64-kbps telephone channel requires 1 byte to be transmitted in each $125\text{-}\mu$s time frame. Other types of applications require a different quantity of bytes to be transmitted within that interval. The key behind TSR is that the time slots are assigned to individual users as they need them and not on a dedicated basis. One might say that is not dedicated bandwidth because of the dynamic nature of allocation. However, the reason it remains in that category is that the satellite link has a fixed total bandwidth.

Packet switching, covered in more detail in Section 3.1.3, also can be used to add dynamic bandwidth allocation to an otherwise dedicated bandwidth link. The Internet Protocol is a commonly applied packet switching scheme. Instead of fixed bytes of capacity, the packet approach transfers the user data to the link as it arrives at the point of access. Temporary storage called buffering is required to adjust for differing arrival rates of different user information. Variable-length packets are used by standard protocols like TCP/IP and Frame Relay. A fixed-length packet, called a cell, can offer dedicated bandwidth or on-demand connection, as
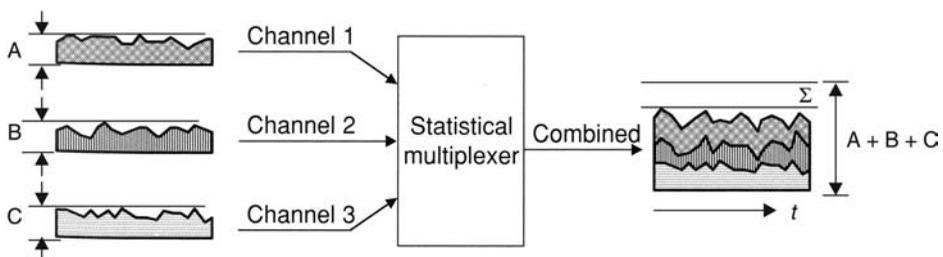

Figure 3.5 An illustration of the principle of statistical multiplexing, demonstrating how total channel bandwidth is reduced.

discussed in Section 3.1.2. Asynchronous transfers mode (ATM), an international standard for cell switching, is used in applications where dynamic bandwidth allocation provides significant improvement over static approaches.

Non-GEO systems are not easily applied to dedicated bandwidth services because the link is interrupted each time a new satellite must be acquired by the Earth station. Communication must be handed off to another beam or satellite without loss of connection. That forces the non-GEO operator to employ one of the other transmission modes to be discussed in the following paragraphs.

3.1.2   Circuit-Switched Services

Many communication applications are active only when a user has actual need for service. A common example of that is the cellular telephone network, which gives a CBR type of service but only for the limited duration of a call. Other CBR applications, like video teleconferencing and image transfer, can employ circuit-switched service. Users like this type of service because they pay for only what they need. While network capacity is only provided on demand, each end of the circuit needs a dedicated termination facility so calls can be made and accepted on demand. The common telephone and the local loop to the local telephone exchange are examples of such a termination facility. For video teleconferencing, the termination facility might consist of television cameras, video compression equipment, and a special type of Earth station that can access the satellite directly.

A circuit-switched connection can be established using dedicated bandwidth with a constant bit rate; alternatively, the logical equivalent of a dedicated connection can be produced using a variable bandwidth service like packet switching. The latter is referred to as a *virtual circuit* because the end points and their applications are defined for the connection and that relationship is not physical (as if by hard electrical connections) but rather remembered by the network for the duration of the "call." The virtual circuit medium is very common in packet switching for persistent on-demand communications needed for Voice over IP (VoIP) and downloading files and Web pages. While the end points are defined in a virtual circuit, the path that the data travels need not be specified or even used consistently for all of the packets that comprise the information.

Figure 3.6 is called a *ladder diagram,* used to illustrate the basic approach to establish and release a circuit. In a telephone network, the user makes a request for service by lifting the telephone receiver (going "off hook"); the network detects that state and responds by providing a dial tone. After the user has entered the destination by dialing digits, the network takes a number of steps to make a connection between the user and the ultimate destination (i.e., another user): the distant user is notified of the incoming call by a ringing signal; the connection is established when the party on the distant end goes off hook, allowing both ends to communicate normally. When the need for service ends, one of the users signals that situation by hanging up (going "on hook"), which then deactivates the connection and deallocates the capacity. In a typical commercial telephone network, many other details must be addressed properly. For that to happen, the network requires a distributed computer intelligence to be able to recognize the request, to track and control available capacity, to signal the initiating and accepting ends of the
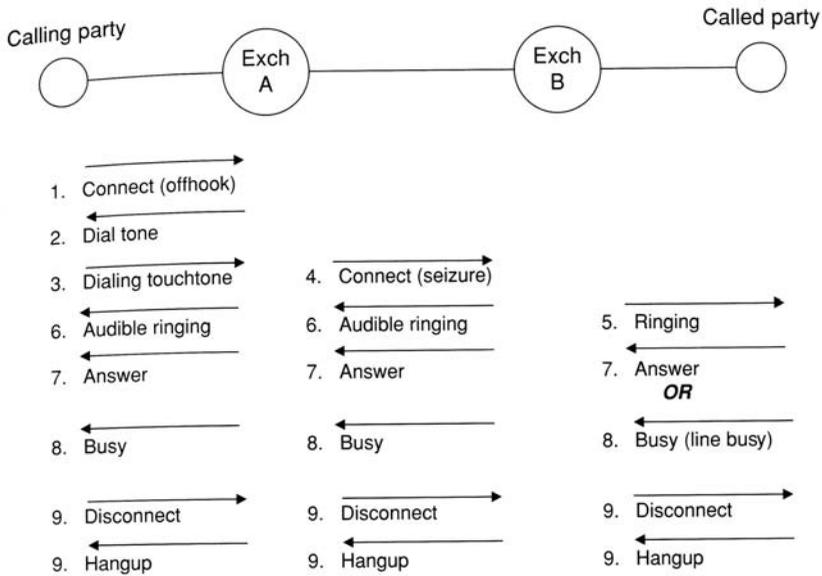
Figure 3.6   A ladder diagram illustrates the basic steps used to establish an ordinary telephone call.

connection, and to provide service management facilities like billing and fault detection and correction. Few, if any, of those capabilities are needed in the case of dedicated bandwidth communications.

In satellite communication, the Earth station network responds to call requests that are given to it either by directly connected telephone users or the public telephone network itself. This type of application is referred to as *demand assignment* because capacity for the circuit is assigned to the particular connection on demand. The capacity that this comes from is a pool of narrowband frequency channels or of time slots within a common timeframe. Control of such a network can be a complex problem. For a traditional GEO satellite with a single footprint, a single control station can manage an entire demand assignment network. That is possible because the control station can receive requests directly from all user stations and transmit control signals to them.

Figure 3.7 illustrates the principle for a telephony network using a variety of single-user and shared multiple-user Earth stations. The Earth station at the upper left responds to a locally connected telephone (a) that goes off hook by sending a service request to the master station. If the master is able to provide satellite capacity, it signals the requesting station; that results in a dial tone being transmitted back to the originating telephone instrument. The calling party can then enter a dialing sequence that informs the network of the destination city, local exchange, and called subscriber. Depending on the particular implementation of the control network, the dialing digits are transferred to the control station, which determines the destination Earth station (indicated at the upper right) and signals it that a connection is to be established. Satellite capacity is allocated to the connection, and the two Earth stations perform a quick check that there is a workable signal path. The called station then signals the local telephone network to complete the connection to the final called party (b), with the final step being the ringing of that
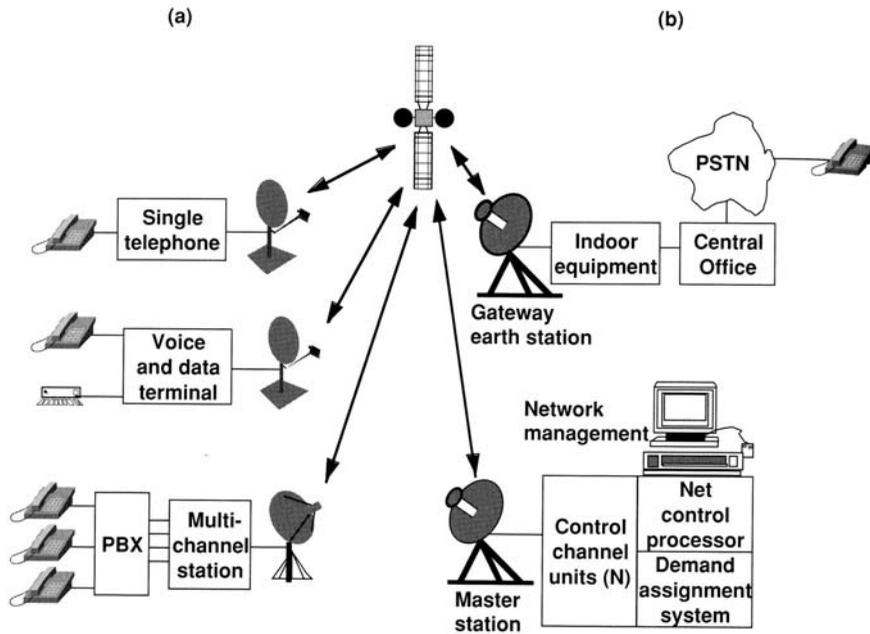
Figure 3.7   (a, b) Typical arrangement of a demand assignment satellite telephone network with central network control and management.

telephone instrument. Once answered, conversation can commence. Taking down of the connection works in the opposite direction.

The process becomes more complicated for a multiple spot beam satellite, either with or without an onboard processor. The control station may not be able to monitor all satellite transmissions, so additional steps are added. Also, a processor-type repeater may need to be configured for the connection. In a non-GEO system, the call may involve connection through multiple satellites and cross-links. Also, handover from satellite to satellite adds complexity. Disturbances and dropouts along the satellite path during either call establishment or actual use can produce a failed or dropped call, respectively. Whether GEO or non-GEO, call setup time can be an issue if it takes more than about 5 seconds. Apart from the possible sources of complexity, a satellite circuit switched network can provide exceptional service in terms of flexible connection for users who could be located anywhere in the coverage area.

3.1.3   Packet-Switched Services

Data communication networks have adopted packet switching as the predominant mode of transferring information, either on a point-to-point or point-to-multipoint basis. In the 1990s, TCP/IP reached critical mass in private and public network applications. In addition, the ATM protocol is a form of packet switching. Both technologies are based on the concept of the data protocol, in which a stream of digital information is broken down (or assembled, as the case may be) into packets

that satisfy certain constraints as to length. To each packet is added a prescribed number of bits called the *header*, used in the process of reliably transferring the packet from source to destination. Once the packet has reached the destination, the header is removed and the original information stream reformed. The protocol often is used to transmit acknowledgment back to the source or, if the data transfer was corrupted, to transmit a request for retransmission.

There are two fundamental schemes for packet switching: the virtual circuit (described previously) and datagram (also called connectionless). Datagram packets are addressed for a particular destination as opposed to a particular circuit. Therefore, no preparation is necessary to set up a virtual circuit, and packets are handled independently. The network routes each packet as if it were a dedicated message from a particular source to a particular recipient. The most classic datagram protocol is the Internet Protocol (IP) network layer of TCP/IP. Interestingly, TCP uses the IP datagram service to perform the information transfer function. It goes through a circuit connection routine before actually offering packetized data from a particular stream of data onto the link.

The connection-oriented and connectionless paradigms represent two distinct ways of implementing a packet switched network. As with TCP/IP, they can be combined to further optimize the performance of the network. Many applications employ a connection-oriented service because it provides: (1) reliable information transfer, (2) no loss of any portion of the input, and (3) acceptable delay (also called latency). This all must occur as the links and network are loaded by user information, a process that often is dynamic. That is because it usually is not possible to predict the precise demand for data transfer, in terms of the number of simultaneous users and their individual data communication needs. A datagram/ connectionless service offers minimum delay in exchange for reduced reliability of information transfer.

3.1.4 Flexibility Features

Satellite communication networks offer a variety of benefits to the system engineer and the user alike. The following paragraphs summarize several of those benefits and put them in context of the overall advantages of the medium introduced in Chapter 1.

3.1.4.1 Implementation of Satellite Networks

Installation of a fiber optic cable system requires first that the right-of-way be secured from organizations such as governments, utility companies, and railroads. Hundreds or even thousands of sites must be provided with shelter and power (and even access roads in the case of terrestrial microwave). After the entire system has been installed and tested, all the equipment must be maintained to ensure continuous service. Still, one outage along the route probably will put the entire chain out of service until a crew and equipment can arrive on the scene to effect repair. When the primary high-capacity fiber optic cable is cut, it can be devastating to the communications of a major city or industrial site. The way around that is

to add redundant links, which can be costly in terms of capital and maintenance expenses.

In contrast, the time to install an Earth station network is relatively short, particularly if the sites are close to where service is provided. In the past, implementation times from Earth stations were lengthened not because of site construction but rather because electronic equipment had to be special-ordered and then manufactured. The low production volumes (because satellite communications require less equipment in general than terrestrial) discouraged manufacturers from mass producing standardized equipment and holding inventory for future sales. In today's larger and more competitive Earth station equipment market, higher manufacturing volumes along with the greater use of standardized digital systems have allowed equipment suppliers to reduce cost and maintain on-the-shelf inventory. The time to implement satellite networks and add stations has been reduced from 1–2 years down to a few months or even weeks.

3.1.4.2   Expansion of the Network

With a proper network architecture, new Earth stations can be added without affecting the existing stations. Increasing the number of receive-only Earth stations (e.g., DTH receivers) is particularly easy and economical because operation of existing stations is not affected.

Dealing with two-way communication Earth stations is a bit more complicated because some setup and coordination are needed to expand an operating network. To add an Earth station to an old analog point-to-point network would require dismantling the equipment at each old station to be linked with the new station. That major drawback of the older systems has been eliminated with programmable digital technology. These more flexible digital approaches, described later in this chapter, can be assumed in virtually every application involving two-way communications. VSATs and multimedia terminals are designed for easy introduction into a working network, which is vital when one considers how large such networks have become.

3.1.4.3   Simplification of Network Routing

Many corporations and government agencies operate private backbone networks for use in enterprisewide information processing and dissemination. Client/server networks and corporate intranets are popular, representing a large commitment in financial and human terms. Such networks are extended to temporary sites, including drilling platforms, ships, and aircraft. Once a major public network has been modernized, it can be expanded or modified by running new local cable loops and by making appropriate wiring changes in the telephone offices. However, time delays of many weeks to months are still involved, beginning from the moment when a user submits a significant service order. The public telecommunication service provider (common carrier) often must perform some network engineering, install equipment if necessary, make the required wiring changes, and then test the resulting circuits for proper operation. If the circuit or circuits cross the boundaries between the terrestrial networks of different common carriers, the process must

be run simultaneously between the various organizations and coordination between them handled in some manner. (The reliability aspects of this particular problem are discussed in Section 3.1.5.)

After AT&T divested its local telephone company operations in 1984, it began to concentrate on long-distance service. A nationwide network of a corporation like General Motors was forced to traverse the network of all of independent regional companies, along with the long-distance networks of AT&T, MCI and US Sprint. That made life both complicated and difficult, as there no longer was one service provider to deal with. The consolidations of AT&T and Verizon have not substantially altered this aspect of corporate network architecture. A similar situation exists in developed regions of Europe and East Asia.

In a modern satellite network, only the end connections are involved because the satellite itself provides all the intermediate transfer or routing. Unlike terrestrial networks that must deliver multipoint connectivity by extending terrestrial links to each and every point to be served, the satellite network reaches all points within a common footprint. There are terrestrial radio techniques that mimic a point-to-multipoint satellite network by placing broadcast transmitters on towers or mountaintops. AM and FM radio and TV work on a point-to-multipoint basis, and cellular mobile telephone is an excellent multipoint-to-point system. However, all such terrestrial techniques are severely restricted as to range because of line-of-sight radio propagation. To extend well beyond that geographical limitation, point-to-point links must be established between the radio towers to chain the broadcast or cellular stations together. Until that is accomplished on a uniform basis, the satellite network remains the most practical means of having a unified approach.

3.1.4.4 Introduction of New Services

Expansion of a satellite network can add new services ahead of terrestrial means. Perhaps the clearest example is the long-distance transmission of full motion color television, which could not be carried over the first generation of analog trans-oceanic telephone cables. It was not until the advent of terrestrial microwave radio in North America that coast-to-coast TV transmission was possible. Satellite repeaters in the FSS have sufficient bandwidth to carry several TV channels, along with an array of voice and data traffic. The telephone local loops, which bring voice and data services into the office and home, are still limited in their capacity. A family of multiplexing technologies called digital subscriber line (DSL) addresses this shortcoming. Home cable television is enhanced with two-way interactive services. WiFi and WiMAX standards likewise offer a local wireless broadband capability in suburban and rural towns. Any and all of those services can be included in or added to networks of small Earth stations, particularly the VSAT operating at Ku- and Ka-bands.

A given satellite network can achieve all three connectivities individually or simultaneously, while a terrestrial network usually is restricted to a point-to-point capability. It is not uncommon for a user to implement a point-to-point satellite network (see Figure 3.1) involving 10 to 50 Earth stations and then add a broadcast capability to extend the network to hundreds or even thousands of receiving points. Any one of the point-to-point stations could then be used as an uplink site to

broadcast digital information or video programming (see Figure 3.2) on a full-time or occasional basis. The multipoint-to-point capability (see Figure 3.3) can be installed in the future by adding a transmit "retrofit" package to many of the smaller receive-only stations.

3.1.5   Reliability of Satellites and Links

The remaining features are more difficult to explain and quantify; however, they ultimately can be the factors that decide in favor of satellite transmission over terrestrial transmission. The mere fact that the typical satellite link requires only one repeater hop tends to make the satellite connection extremely reliable. Complexity is increased for dual-hop intercontinental links over GEO statellites, as well as for multiple hops and intersatellite links for non-GEO systems. Nevertheless, the number of such transfers is substantially less than for terrestrial alternatives that must extend the same distance.

　　The reliability of a communication link typically is specified by the percentage of time it is capable of providing service, called the *availability*. Figure 3.8 indicates the number of outage hours per year that results from a given availability value. The engineering of the link, described in Chapter 4, must properly take into account the frequency bands, fade margin requirements, and orbit geometry. When that is done and an established satellite system is employed, the link will be up and usable well in excess of 99% of the time. In fact, satellite engineers normally talk of link reliabilities of 99.9% for FSS applications, which equates to an outage or downtime of 9 hours a year. Normally, outage is segmented into durations of a few minutes distributed mainly through the rainiest months. In the case of MSS systems, availability is reduced by the impact of mobile fading due to obstructions and, in the case of non-GEO constellations, satellite motion and handover.

　　Cable systems are susceptible to accidental breakage or deterioration of the cable itself, and outages of several hours or even days at a time do occur. Therefore, proper terrestrial system design requires that diverse routes be provided. Buried cable with diverse routing and automatic switching can satisfy a 99.99%-level
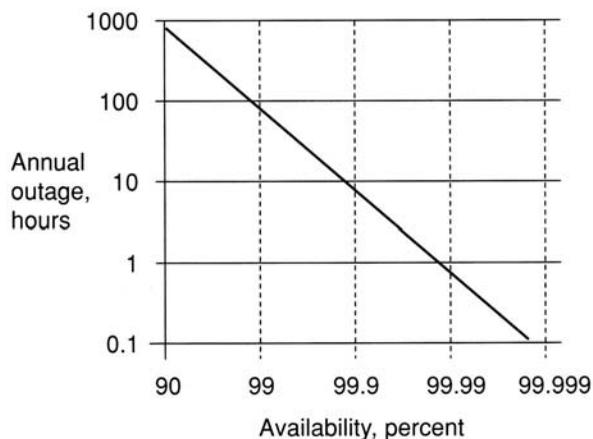

Figure 3.8   Service availability in percentage versus equivalent outages per year in hours.

requirement, although the cost of implementation would be within the reach only of relatively wealthy organizations (national telephone operators, governments, and major industrial corporations).

Equipment failures on satellite links do occur, and for that reason backup systems may be appropriate. A communications satellite contains essentially 100% backup for all its critical subsystems to prevent a catastrophic failure. Double failures, for example of primary and backup equipment, have occurred, so it is vital that procedures be in place to routinely check operation. The individual transponders or transmitters within the repeater section usually will not be backed up 100%, so a fractional loss of capacity is possible at some time in the useful orbital life. Experience with modern commercial satellites has been excellent, and users have come to expect near perfection in the reliability of the spacecraft. However, because total satellite failures (including launch and on orbit) can and do occur, satellite operators need contingency plans of the type discussed in Chapter 10.

A common cause of communications disruption is the RFI problem described in Chapter 1. Harmful RFI is a fairly routine but infrequent occurrence, and satellite operators are reasonably well equipped to respond to and identify the source of the problem (which almost always is accidental and of short duration). A severe interference usually is isolated to one frequency, allowing the satellite operator to reassign users to different frequencies or transponders until the RFI is removed.

The reliability of satellite communication is enhanced by the fact that virtually all the ground facilities can be under the direct control of one using organization. If a problem occurs with equipment or its interface with other facilities such as telephone switches or computers, the user's technical support personnel can easily identify and eventually reach the trouble spot. Restoration of service thus can be accomplished conveniently and quickly. In contrast, terrestrial linkups can involve many organizations that provide segments of the overall service, complicating the necessary troubleshooting and follow-up.

3.1.6   Quality Features and Issues

The quality of satellite communications or any communications service must be defined in user terms. The following sections identify different approaches to measuring quality of transmission as perceived by humans. If the information is analog in nature, quality is measured objectively by using a signal-to-noise ratio or subjectively by surveying people (subjects) who have been exposed to the service. Quality in data communications boils down to the quantity of valid data that reaches the distant end (the bit or packet error rate), along with the time it takes to get there (the latency).

3.1.6.1   Signal Reproduction

For a single point of transmission, a satellite is nearly ideal for delivering a signal of the highest quality. Modern GEO FSS satellites radiate sufficient power into the geographical footprint to be received by ground antennas of diameters in the range of 45 cm to 10m (18 inches to 32 feet). The GEO and non-GEO MSS systems

introduced in the late 1990s reach antennas the size of your thumb. Because satellites use line-of-sight transmission in directions perpendicular to the atmosphere, the frequency and duration of link fades are reduced. Satellite links aimed near the horizon, at a low elevation angle, do experience fading similar to terrestrial microwave.

The predominant mode of transmission over satellites is digital, in which information is transformed from its analog form into a data stream. Such a process alters the ideal quality of the signal, introducing a certain level of distortion due to quantization (assigning the analog signal to discrete voltage levels that can be represented by numerical values) and compression (altering the actual information to reduce the required number of bits). As long as the bits are transferred through the link with little or no error, the quality remains exactly as it was after the initial conversion process. Whether a satellite or a terrestrial link is used, signal quality may be maintained at all times.

Many terrestrial networks suffer from man-made noise and various kinds of short interruptions (''glitches'') while satellite links experience primarily receiver noise, which is constant and easily compensated for with power or improved forward error correction. Satellite-based mobile communication is affected by various forms of blockage, which makes these services somewhat more glitchprone than terrestrial cellular networks. That can cause data dropouts that will affect the consistency and continuity of the delivered information.

When these factors are properly taken into account, satellite communication can be engineered for high circuit quality and reliability. Designers can then select appropriate Earth station equipment that will provide that quality confidently. The application in which these aspects of quality play the greatest role is video. From a viewer perspective, the quality of the delivered video signal is the same as the signal created at the studio or played from the originating video tape machine.

3.1.6.2   Voice Quality and Echo

The issue of quality of voice transmission over satellites has received a lot of attention because of heavy competition among satellite and terrestrial communication companies. As discussed in Chapter 2, the use of the GEO for communications relays was a controversial topic prior to the first use of SYNCOM in 1962 because of the quarter-second delay introduced by the long transmission path. Voice communications over a GEO satellite can be made acceptable to over 90% of telephone subscribers, as has been proved by numerous quality surveys. Terrestrial and LEO systems that span continents and oceans do not suffer from noticeable delay and hence are potentially more desirable from that standpoint alone. People who routinely make phone calls over the Internet often experience delays comparable to the GEO path.

The delay itself is not as detrimental as the presence of a returned echo to the speaking party. Both terrestrial and satellite links are capable of producing echo, as illustrated in Figure 3.9. At the left of the figure, speech from a female talker is converted into electrical energy in the voice frequency range (300 to 3,400 Hz) by the handset and passes over the single pair of wires to the telephone equipment
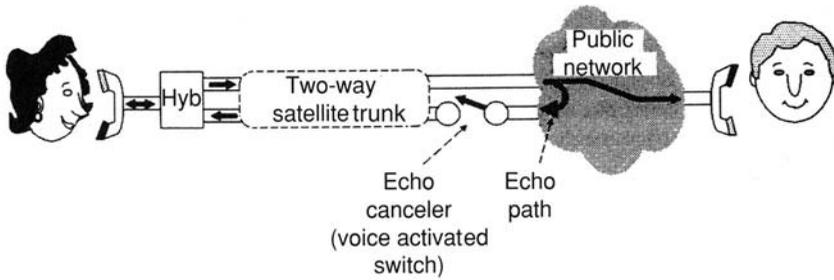
Figure 3.9 Telephone echo over a satellite is caused by electrical reflection at the distant end, where it can be eliminated by an echo canceler.

used to connect to a long-distance circuit. The same pair of wires allows the speech from the distant end (where the male talker is listening) to reach the female talker.

The connection point onto the long-distance circuit breaks the two directions in half, segregating the sending and receiving wire pairs. The device that routes the energy properly between the two-wire local loop and the four-wire trunk line is called a *hybrid*. The typical configuration has a hybrid on each end; however, Figure 3.9 shows that the male talker is connected through an undefined terrestrial network within which several hybrids could exit. The echo path is produced within one or more of the unseen hybrids, allowing some of the female talker's speech energy to make a U-turn and head back over the link toward the female talker, arriving in a delayed (and potentially annoying) fashion. Because the echo is the result of uncontrollable factors in the terrestrial network, it must be blocked or removed in some manner. Obviously, a satellite circuit with its quarter-second delay is subject to an echo that is heard approximately one-half second after the talker utters his or her first word. Undersea cable circuits might introduce an echo return delay of only 50 ms or less, but that still would be unacceptable for commercial telephone service. The same issue applies to digital cellular and Internet phone calls.

The simplest and most effective way to block the echo is to use a voice-activated switch at the opposite end of the circuit, as shown in Figure 3.9. Whenever the female talker is speaking, the control circuitry detects the presence of the incoming speech on the upper wire pair, and the switch on the lower wire pair is opened. When she stops talking, however, the switch closes automatically and the male talker is free to speak and be heard by the female talker. A similar switch would have to be placed on the female talker's end to protect the male talker from his own echo. Other features are necessary to make the switch respond to characteristics of human conversation, such as when one party needs to interrupt the other (double talking).

All long-distance telephone networks can achieve excellent echo control with the digital *echo canceler*, which is included with a digital telephone exchange or satellite network. An echo canceler works with the digitized version of the speech and mathematically eliminates the echo from it. It does that initially by measuring and characterizing the echo path through the hybrid and terrestrial network. From that information, the canceler determines how to subtract a sample of the incoming

speech from the return path to the distant talker. After an initial training interval, the echo is suppressed by a factor of about 10,000 (e.g., 40 dB). Additional echo reduction is accomplished with a voice-activated switch that can totally block the return path when only one person is speaking. To allow double talking, only the echo cancelation function is operating to permit both sides of the circuit to work simultaneously. While the details of how this technology works are beyond the scope of this book, the important point is that there is strong evidence that modern echo cancelation along with high-quality digital transmission produces satisfactory service for even single-hop GEO satellite networks with a total delay of 500 ms or less. For delay greater than 500 ms, even the echo canceler cannot overcome the effect of latency on the interactive quality of a telephone circuit.

3.1.6.3   Data Communications and Protocols

The key to successful computer-to-computer communication over satellite is to employ the right protocols and coding schemes. Computers that communicate over terrestrial links transmit data in blocks of data, or *words,* as shown in Figure 3.10. (The specific structure of protocols is contained in standards, particularly TCP/IP). A protocol provides a set of rules that the two ends follow in assembling the words for transmission and determining if the received words are valid. In a highly-reliable protocol, provision is also made to automatically request retransmission if an error is detected at the distant end. The "conversation" illustrated in Figure 3.10 shows what happens when the computer at the receiving end detects an error and requests retransmission from the sending end (on the left). The sequences starts at the upper left and moves left to right and right to left, following the direction of the arrows
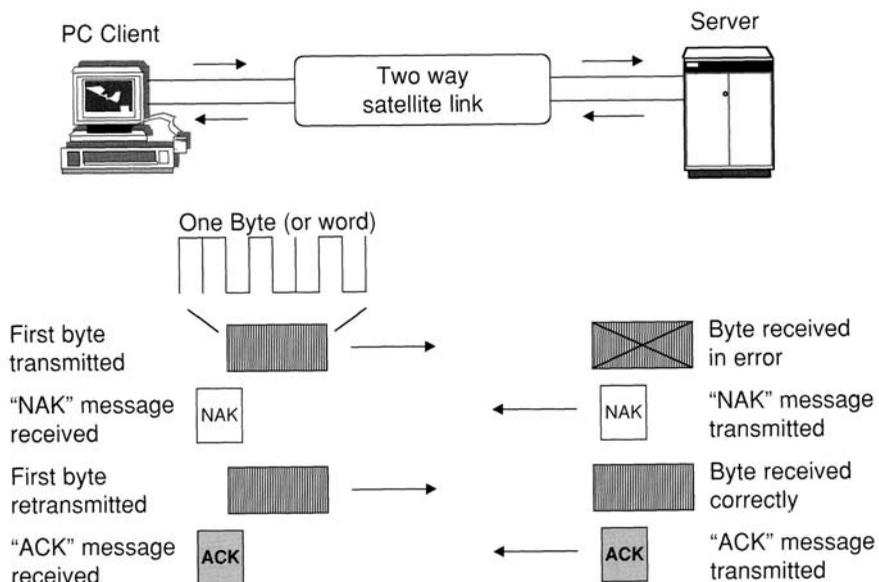

Figure 3.10   Inefficient "word-oriented" data protocols are subject to greater delay on a satellite link when errors in transmission are experienced.

in the form of a ladder diagram. As long as errors are infrequent and the resulting delay is short, the protocol provides a satisfactory service to the user. However, satellite delay in the presence of data errors could force the sending end to halt the transmission of new data for as much as 2 seconds as the necessary requests and retransmissions are executed. A terrestrial connection would not exhibit that difficulty, particularly for high-quality fiber optics.

Latency-related service quality can be improved by giving the sending end the acknowledgment that it expects even before the distant end has received the word or group of words. The piece of equipment that acomplishes the task of pacifying the sending end is called a protocol gateway. Another term that is used to describe this is protocol *spoofing*.

A protocol gateway corrects for most of the performance issues over a GEO satellite except the delay itself. By handling errors, the user experience can be nearly the same as that from an average-quality terrestrial line such as DSL or a partial T1. The basic arrangement and operation of a protocol gateway is illustrated in Figure 3.11 using the TCP/IP connection-oriented protocol from the Internet Protocol suite [2]. The approach taken in this illustration is based on the layered model of a data communications protocol. For TCP/IP, the standard layers are the Physical Layer, Link Layer, Network Layer, Transport Layer, and Application Layer. This is somewhat abbreviated from the seven-layers of the Open Systems Interconnection (OSI) model, which has largely been overtaken by TCP/IP. The left-hand and right-hand "stacks" represent the user client computer and server computer, respectively. The protocol gateway has two components as well—on the left is the one that is associated with the client and on the right is the one associated with the server. Communication between the client and gateway immediately attached to it uses the standard TCP/IP protocol; the same can be said of the server and its associated
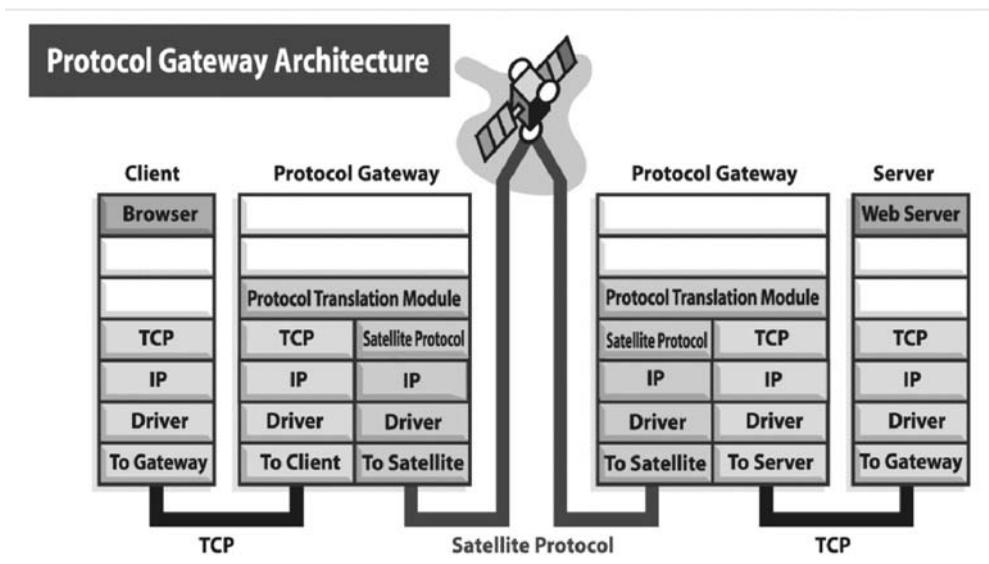

Figure 3.11   A protocol gateway converts between TCP/IP and a satellite protocol that is optimized for the delay and error rate properties of the satellite link [2].

gateway. Thus, the client and server are given a standard interface and the communication follows the well-known rules of TCP/IP absent the vagaries of the satellite link in the center.

To address the increased latency and errors over the satellite link, the two protocol gateways employ a special data communication protocol that is optimized for that particular environment. For example, data would be assembled into larger blocks than used by TCP/IP, and special software handles an efficient exchange with reduced impact of delay and errors. One of the issues with TCP/IP that is addressed in this manner is what is called the "slow start." When the sending end of a TCP/IP connection experiences poor throughput due to errors and retransmissions, it automatically drops the flow rate of data to a minimum. From there, it builds the rate back up and actually stops at half the rate that existed before throughput was disturbed by the link. Slow start is not as much of an issue on terrestrial circuits due to the shorter delay and possibly higher data rate capability. Over a satellite link, it can be devastating to the user experience.

So we see that the protocol gateway, which speaks the standard protocol (TCP/IP) to the end computers and an optimized satellite protocol over the radio link, provides the user with a good quality of service. The majority of satellite-based data communications systems, particularly VSATs, employ the protocol gateway principle.

Errors are corrected at the other end using forward error correction (FEC) coding. Another feature that adds to the reliability of end-to-end communication is the cyclic redundancy check (CRC). To apply the technique, the sending communication device performs a calculation on the data word that results in a CRC bit sequence. The receiving end makes a similar calculation on the incoming word and compares this version of the CRC sequence against the one attached by the sending end. If they do not agree, then the receiving end requests retransmission (or does not send an acknowledgment).

Recognizing that the spoofing or protocol gateway approach was only a provisional fix, software developers have considered modifications to TCP/IP to address the known problems with data communications over wireless communications systems in general and satellite links in particular. Modern protocols could make satellite links nearly as efficient as the best-quality terrestrial links for the transmission of high-speed data. The performance is further enhanced by increasing the bandwidth on the satellite link and hence, the throughput of data as provided on an end-to-end basis.

The packet transmission approach, discussed earlier in this chapter, provides another effective means to control errors on terrestrial and satellite networks. The first packet protocol specially designed for satellite links is called ALOHA (discussed in Chapter 5), developed by Dr. Norman Abramson at the University of Hawaii. In ALOHA, a given station transmits its packet of data immediately and in the blind. That ensures that there is the least possible delay through the satellite link, provided there is no overlap between two or more transmissions that happen to arrive at the satellite at the same time. Such an overlap is termed a collision, causing those particular packets to become corrupted. When the receiving station gets a good packet, it sends an acknowledgment message. The protocol of ALOHA provides for automatic retransmission if no acknowledgment is received, with the

transmitting stations each waiting a randomly selected time period to prevent a second occurrence. Most VSAT data networks employ versions of ALOHA and packet transmission to greatly improve throughput and responsiveness of satellite data transmission.

The issues of reliability and quality can, in balance, weigh in favor of satellite communications when the engineering, installation, and operation are done properly. The difficulty often is that the terrestrial network exists first, and transferring service over to a satellite network becomes a complex and difficult task. In cases where terrestrial service did not exist first or was totally inadequate, the new satellite network was very well accepted (INTELSAT and the Indonesian Palapa A system are good examples). Satellite broadband services using VSATs have likewise shown flexibility and reliability in data applications.

3.2   Point-to-Multipoint (Broadcast) Networks

Broadcast links take full advantage of the wide area coverage of the satellite's footprint. Figure 3.12 provides a typical architecture for how satellite broadcasting is accomplished, using one transmitting Earth station (the right-hand dish) and many receive-only (RO) Earth stations. The GEO satellite repeater retransmits one or more RF carriers containing the information to be distributed. This particular network is used to distribute TV programming to local rebroadcasting stations that serve viewers in urban areas.

As the number of ROs increases into the millions common in DTH applications, the optimum transmitter power to use in space becomes much larger than that permitted at C-band by the ITU. The BSS segment of Ku-band is available for such high-power broadcast applications, as discussed in Chapter 1. The cost of the more
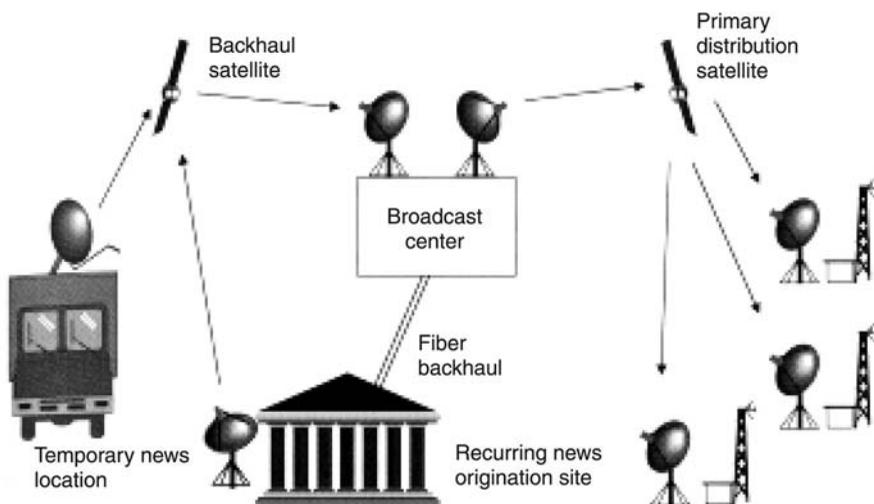

Figure 3.12   The use of satellite transmission in the commercial television industry for backhaul of event coverage and program distribution to affiliates.

expensive BSS satellite is shared among more and more users, who then save substantial amounts on the cost of their ground equipment. That is an economic tradeoff between the cost of the satellite and that of the ground segment, as discussed in Chapter 12.

Achieving point-to-multipoint connectivity exclusively through a terrestrial network is extremely expensive, since the cost of adding cable or microwave facilities to reach all service points is roughly proportional to the number of points. The only economy of scale in terrestrial broadcasing comes if many services can share a common local cable distribution network. Also, chaining terrestrial links together tends to be less reliable on an overall basis because a chain is no stronger than its weakest link. Cable TV systems, on the other hand, are very effective in urban and suburban environments because the cable infrastructure has the potential to serve a large and economically attractive community. In data communications, telephone companies use digital subscriber line (DSL) technology to impart broadband data over copper twisted pair.

This section reviews the typical satellite applications for the point-to-multipoint broadcast connectivity.

3.2.1 Video Distribution

Television or video service (which are one and the same) is perhaps the most popular source of entertainment and information for the public. The broadcasting industry has embraced satellite communication as the primary means of carrying programming from the program originator (TV networks, cable TV programmers, content producers, and program syndicators) to the final point of distribution (broadcast TV stations, cable TV system operators, home dishes, and PCs). In this section, we review the way in which programmers and distributors use satellites and Earth stations in their business.

3.2.1.1 TV Broadcasting

As an example of the importance of the current role of satellites, we begin with a review of the general characteristics of the TV broadcasting industry as it exists in North America. Broadcasting is the commonplace medium whereby local TV stations employ VHF or UHF frequencies to transmit programming to the local community. Beginning as an analog signal, TV is undergoing digital conversion to increase both the channel capacity and to offer high-definition (HD) video quality. The range of reception usually is limited by line-of-sight propagation to approximately 50 to 150 km. The Nielsen Media Research organization specifies the local coverage regions in the United States as the defined market areas (DMAs). There are approximately 200 DMAs in the continental United States, corresponding more or less to coverage of local TV stations and their respective advertising. To conserve frequency spectrum, the same channel is assigned by the government to another station some safe distance away. Individuals use directional antennas (yagis and reflector dipoles) to maximize signal strength and to suppress reception of unwanted distant stations operating on the same or adjacent channels. A given station trans-

mits only a single channel and hence, is constrained to offer only one program at a time. This limitation can change as digital television (DTV) is introduced.

Analog TV transmission employs one of the three global standards: National Television Standards Committee (NTSC) in North America and South America, Japan, the Philippines, and Korea; Phase-Alteration-Line (PAL) in much of Europe and Asia; and *Sequential Couleur a Memoire* (SECAM) in France, several other European countries, Russia, and many countries in Africa and Central Europe. DTV broadcasting adds more channels and features, such as data distribution and interactivity over the air. The FCC decided that U.S. TV stations must adopt the digital standard known as the Advanced Television Systems Committee (ATSC) or face the loss of their broadcast rights after approximately 2009.

Networks, Affiliates, and Independent Stations

National television networks (ABC, CBS, NBC, and Fox in the United States) provide programs to affiliated TV stations for broadcast over their assigned frequency channel either in real time or by replay from tape or hard drives. In the majority of other countries, all local broadcasting is via network programming. Stations in the United States also can obtain programming from the outside from syndication companies, which sell programs either individually or as packages. Networks around the world often purchase programs from syndicators, which themselves are networks. While most syndicated programs are in fact old network programs (reruns), syndicators often deliver new programs and movies.

Stations operate their own studios so that they too can originate programs, particularly local news, special events, and, most important to the success of the station, advertising. The revenues of the stations and the networks are derived primarily from the sale of advertising because individual viewers do not pay for the right to watch over-the-air television (except of course when they buy the advertised product or service). Subscription television (STV) employs scrambling to control viewing of the broadcasts and to ensure that monthly fees are paid to the service provider and ultimately the studio or other copyright holder. In the United States, STV was successful for only a short while between 1980 and 1982 until competition from videocassette recorders (VCRs) and cable television undermined their profitability.

TV programs also are distributed through other means with advertising again playing a major funding role. As discussed later in this chapter, satellite radio services are beginning to offer TV programs to mobile users. The same function is also appearing on third generation (3G) cellular networks. Underneath all is the Internet and the World Wide Web, offering an array of real-time and on-demand TV programs. DTH and DTV services are responding with feature-enhanced set-top boxes that deliver HD video from network and satellite feeds along with downloaded content and interactive services via a high-speed Internet connection to the home.

TV networks offer advertisers a nationwide audience, which is important to products like automobiles and pharmaceuticals. The revenues of the stations and networks are tied to the relative size of their respective audiences, which is evaluated by respected polling organizations such as Nielsen. Therefore, the programmers need to deliver programming of sufficiently high quality to attract the largest

possible audience. Their profitability is constrained, however, by the cost of produc-ing this programming and of delivering it to the affiliated stations.

Satellite Content Distribution
That brings us to the importance of satellites in providing the needed low cost and highly reliable means of delivering the content. A single satellite can employ point-to-multipoint connectivity to perform that function on a routine basis. To receive programming, every TV station in the United States owns and operates at least one receiving Earth station, and many own Earth stations usable as uplinks. To achieve very high reliability during an extremely high value (in terms of advertising dollars) event such as the Olympics or the Super Bowl, a network will "double feed" the program on a second satellite or a fiber optic link.

The Public Broadcasting Service (PBS) is the U.S.-based nonprofit television network that distributes programming to local public television stations. Most of the funds for PBS and the stations are raised through individual and corporate contributions rather than advertising. Stations also pay PBS and each other for program production and rights for broadcasting. That has allowed the development of a narrower slice of programming (i.e., not of mass appeal) that caters to an audience more interested in education, public affairs, and cultural events. Because of budget constraints, PBS was the first to adopt satellite delivery in 1976, using the Westar 1 satellite. Having demonstrated the benefits of satellite delivery, the commercial networks then began to move quickly in the same direction during the following years. By 1984, all network programming and most syndicated program-ming was being delivered by satellite. Prior to that, the networks had used satellite links over the INTELSAT system to provide coverage of overseas events.

The predominant frequency band employed for video in North and South America is C-band, for the simple reason that more ground antennas and satellites are available than at Ku-band. The program distribution satellite shown in Figure 3.12 is used to broadcast the edited program feed on a point-to-multipoint basis. The downlink is received at each TV station by its own RO Earth station and from there it is either transmitted over the local TV channel or stored on tape or hard drive. In the case of a live broadcast from the network studio, the signal is connected from the camera to the uplink Earth station and over the program distribution satellite. A video switching capability in the studio and at each TV station allows technicians to insert recorded advertising and computer-generated graphics. Even though most programs are recorded, it generally is more economical to distribute these programs by satellite to the TV stations where they are again recorded rather than mailing the tapes (a process called "bicycling") around the country. Since most commercial locations are now served with fiber, it is becoming feasible to deliver recorded programs by terrestrial means.

Whether the programs are live or recorded, the local TV stations are able to insert their own paid advertising in time slots left for that purpose by the network or the syndicator. Stations and networks have automated those functions to improve productivity and to reduce errors. The reason for the latter improvement is that, because commercial stations and networks are paid by advertisers, every lost or degraded commercial results in lost revenue. The technology to do that is an

outgrowth of the computer industry, where very high capacity hard disk drives and high-speed data transmission lines have been made relatively cheap and reliable.

Backhaul of Event Coverage
All sports events and much news coverage are brought back to the studio over a separate point-to-point satellite link called a *backhaul*. In the case of football games, for example, stadiums in North America have access via fiber or terrestrial microwave to a local Earth station, which can uplink the telecast to a backhaul satellite (shown on the left in Figure 3.12). The network or stations pay for the use of the satellite and uplinking Earth station by the minute or the hour. If it is not scrambled or encrypted, anyone with a receiving Earth station can pick up the backhaul, which does not yet include either the "commentary" or advertising spots that are inserted at the studio prior to re-uplinking to the program distribution satellite. If coverage is of a one-time event such as a natural disaster or an Olympic race, then a truck-mounted transportable Earth station is driven to the site and erected prior to transmission. The use of Ku-band (14/12 GHz) is particularly attractive for this type of rapid deployment service and is called satellite news gathering (SNG). A Ku-band SNG vehicle, shown in Figure 3.13, is much more compact and mobile than its C-band equivalent and can be operated almost immediately after it has been parked on location. In addition, the use of nonshared Ku-band frequencies eliminates any need for prior frequency coordination. If an SNG vehicle cannot reach the desired location, then a "fly-away" portable like that shown in Figure 3.14 can be used. Whether C-band or Ku-band, the time demands and
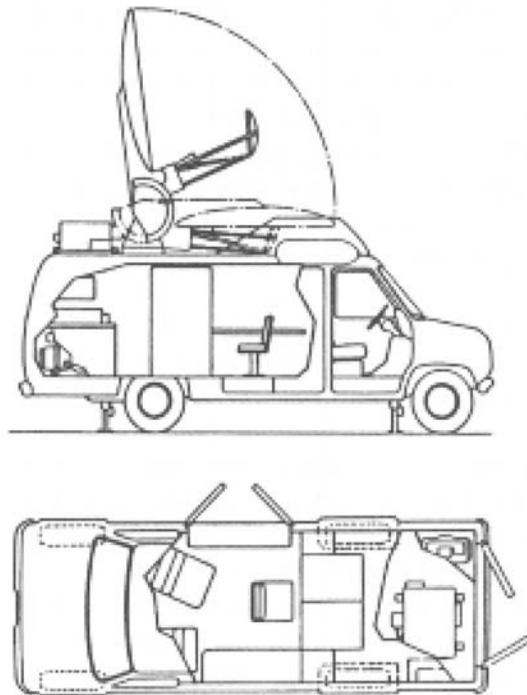

Figure 3.13   An example of an SNG vehicle.

Figure 3.14 A compact fly-away Ku-band video uplink Earth station available from Advent of the United Kingdom.

economics of event coverage can mean that a backhaul satellite link will be attractive where the distance to the studio is anywhere from 50 to 5,000 km. For example, a backhaul was used during the Los Angeles Olympics in 1984 to reach from Lake Casitas to Hollywood, a distance of approximately 90 km.

Ground Antenna Utilization
A network-affiliated TV station will use one fixed-mounted Earth station antenna to receive full-time network programming from the point-to-multipoint program distribution satellite. In addition, some "roving" among other satellites can be done with a movable antenna to pick up special programs provided by syndicators and to receive live coverage of sports events of interest to the local community (e.g., when the local baseball team is playing an away game in another city). Antennas used by the networks in backhaul service therefore would need to be movable since events and satellites change from time to time. Some of these movable antennas can be controlled by the network without intervention by the local station. The actual control signals are generated automatically by the broadcast center and transmitted digitally over the program distribution satellite feed in unused bandwidth.

3.2.1.2 Cable Television

The cable television medium has achieved widespread acceptance in major cities around the world, with 70% of North American households subscribing to cable service. Originally a means to bring over-the-air broadcasts into remote areas with otherwise poor reception, a cable TV system uses coaxial and fiber optic cable to connect to each home through a point-to-multipoint distribution network. The

arrangement of a typical local cable TV system is shown in Figure 3.15. The programming material is collected at the head end, which has high-gain receiving antennas to pick up TV signals with reasonably good quality. In fact, the original name for cable TV was CATV, standing for community antenna television. A studio may be provided at a point between the head end and the cable distribution network for playback of commercials, limited program origination and organization of video-on-demand (VOD) services. Unlike over-the-air broadcasting, viewers (called subscribers) pay a monthly feee for reception of the several TV channels delivered by the cable (many of which are ''free'' advertising supported by local and distant TV stations).

In the mid-1970s, Home Box Office (HBO), now a subsidiary of Time Warner, experimented with a closed programming service that was made available to cable TV systems on a subscription basis. The key to that success has been the control that cable programmers and the cable system operators have over the delivery of the program ''product'' to the home. More recently, telephone companies are entering the business using a combination of fiber-to-the-home (FTTH) and Internet Protocol TV (IPTV). As mentioned in Chapter 2, satellite distribution was adopted
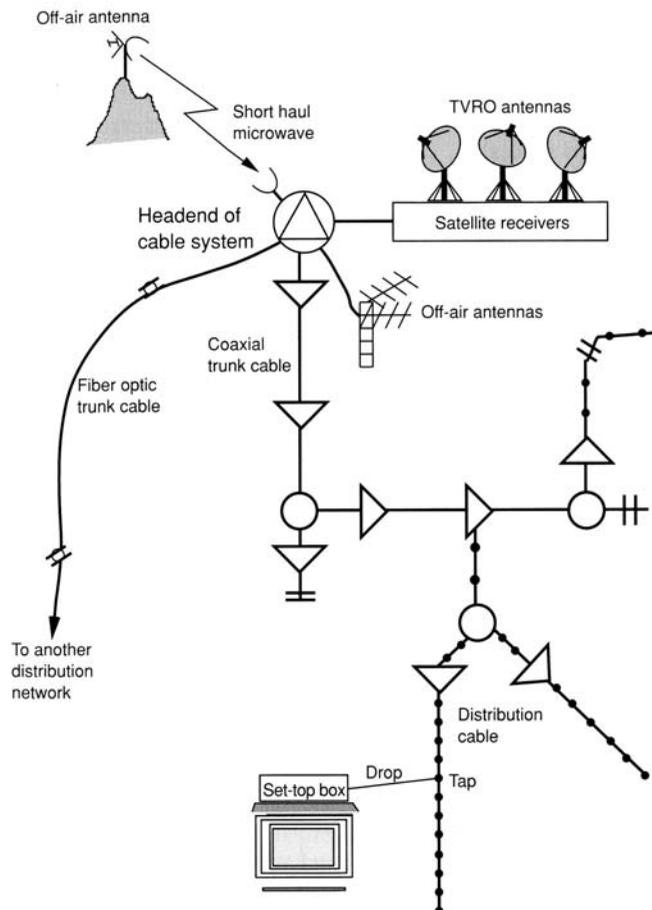

Figure 3.15 Typical layout of a cable TV system.

because of its low cost of making the program channel available throughout the country. Today, approximately one-third of all domestic or regional satellite transponders are employed for cable program distribution.

Classes of Cable Programming
Several classes of specialized cable programming exist, as summarized in Table 3.3. Many are available internationally over C-band and Ku-band satellites, from which they can be received through DTH and local cable services. The satellite-delivered channels are each focused into particular niches. HBO established the most lucrative niche, that of recent movies. In fact, prior to widespread use of VCRs, HBO and Showtime (a service of Viacom International) provided the only means to watch recent movies uninterrupted by commercials. To continuously receive those services along with the regular channels, subscribers pay an additional monthly fee, which is shared between the programmer and the cable system operator. The revenue of the programmer is then used like the advertising revenues of a commercial TV network, paying for the acquisition of programming and for its distribution by satellite.

Because of the marketing strategies of the movie studios that produce most of the product offered by the premium services like HBO and Showtime, there is a time delay (release window) of 6 months or more between when a movie is first shown in movie theaters and when it is made available to cable. Furthermore, the time delay is somewhat shorter with regard to the availability on DVDs for home usage. A VOD service allows the subscriber to receive a specific movie for a fee. Because the administration of this type of service is complicated, the ordering function usually is implemented through the settop box. HDTV has had a significant impact on cable TV, both in terms of its attraction to the consumer and the costs of providing programming and bandwidth.

Satellite Utilization in Cable TV
The satellite delivery system used by cable essentially is identical to that employed by the networks. Program distribution satellites are used to carry the cable programming channels to dishes at cable system head ends. A typical cable system head end in the United States will continuously receive 100 satellite delivered channels, coming from three to six different satellites. That means multiple receiving antennas are required. The cost of the receiving antenna is relatively low because of the

Table 3.3   A Selection of Cable Television Programming Networks That Originate in the United States

Type of Service	Examples
Pay movie (premium service)	HBO, Showtime, Starz, HDNet
Superstation	TBS, USA Network
24-hour news	CNN, CNBC, CSPAN, Fox News
24-hour sports	ESPN, Fox Sports
Music	MTV, The Nashville Network
Children's programming	Nickelodeon, Disney, Cartoon Network
Cultural programming	Arts & Entertainment, Food Network
Science and education	Discovery, The Learning Channel, HGTV
Home shopping	Home Shopping Network, QVC

small reflector size required (3–6m). Expense sometimes can be minimized by using a single reflector with several feeds, each aligned for reception of closely spaced satellites. An example of such a multifeed antenna is shown in Figure 3.16. In actuality, it is the cost of the indoor receivers and considerations as to physical space for antennas that dominate the economics of the cable head end.

To better control access to their product, the cable programmers scramble the programming at the uplink using the analog Video Cipher II system, developed by General Instruments, or a digital system like Digi-Cipher or the digital video broadcasting (DVB) standard. The receiving site at the cable head end then requires an appropriate integrated receiver-decoder (IRD) for each channel being recovered. This approach allows the programmer to certify the particular downlink because the descramblers can be activated (addressed) individually.

Some cable programming services cover remote events and therefore employ satellite backhaul. CNN and Fox News Channel rely heavily on backhaul to provide nearly around-the-world news coverage. Also, Fox Sports and ESPN channel distribute games of professional sports teams and are a major user of backhaul links. An important requirement in the use of satellites for cable is that the backhaul (which is occasional) be separate from the program distribution (which is full time and continuous). Cable systems themselves generally do not alter the programming coming from the satellite, and many cannot handle a break in transmission the way TV stations are equipped to do. Instead, the satellite signal employs imbedded "abc tones" or other triggers to control ad insertion at the cable head end.

3.2.2 Direct-to-Home Television

A brief history of DTH services was provided in Chapter 2. From the U.S. perspective, DTH appeared as a viable business in the early 1980s through the manufacture
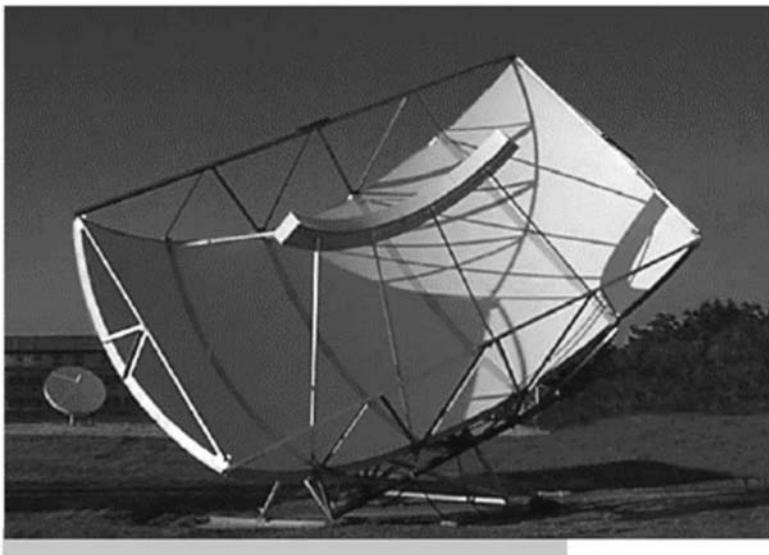

Figure 3.16   The Simulsat 7 multibeam Earth station antenna. (*Source:* ATCi.)

and sale of inexpensive C-band receiving systems to the public. Called backyard dishes, their attraction at the time was the cable programming available for free that the satellites delivered. By 1985, over one million backyard dishes were installed at homes, making the concept of direct broadcasting a reality at least to many U.S. viewers. The pay cable services began to scramble their signals in 1986, and a new phase of industry development was underway. Individuals could buy a Video Cipher II decoder and pay a monthly or annual fee to the cable programmers so their decoders would be authorized to unscramble the signals. The development of this new (but not unexpected) market was important to the success of the Ku-band direct broadcast satellite concept itself.

Nearly all DTH systems employ high power Ku-band satellites that use the BSS channel plan. Under that plan, each orbit position has 32 channels assigned to it. A medium-power satellite might transmit 16 channels at one time, thus requiring that two satellites be operated in the same slot. This approach has the added benefit of protecting the service against the failure of a single satellite. The full complement of 32 channels of about 27 MHz each can deliver a mixture of 150 and 250 compressed standard definition and HD digital video channels. The broadcast center in this case must be capable of either originating or retransmitting a substantial number of simultaneous video and audio feeds. Other features include subscriber control through conditional access and provision for customer service and billing. Subscribers can select PPV movies at any time, with the record of usage transmitted over a connection from the settop box to the PSTN. Integration of DBS with the terrestrial Internet was under development at the time of this writing.

3.2.3   Content Distribution Networks

The rapid growth of Internet applications brought with it a keen interest in using this medium to distribute live and recorded video, audio, image, and other content over point-to-multipoint satellite links. Thus was born the concept of the content distribution network (CDN). The satellite component is the standard Ku- or C-band area coverage repeater with a footprint to match the intended market. During the early 2000s, there were many concepts and startup companies that intended to pursue what was hoped to be the next big market in satellite communications. What happened instead is that CDNs evolved into a valuable niche to serve private network needs and as the means to deliver alternative forms of data over existing satellite broadcasting systems.

There are many potential applications for satellite-based CDNs but one of the most promising is used in the retail environment for digital signage. Figure 3.17 illustrates a digital signage CDN to reach customers' eyes and ears in retail branches such as stores, banks, or gasoline stations. On the left is shown the system to create and manage the content itself, which can be composed of a live TV signal from a real-time information broadcast; previously stored video, audio, or other content related to the products or services of this network; and other real-time feeds such as traffic and financial information. Obviously, there can be a lot of material to deal with and it needs to be selected, organized for display and scheduled for delivery. These are the functions performed using one or more computer workstations such as that illustrated.
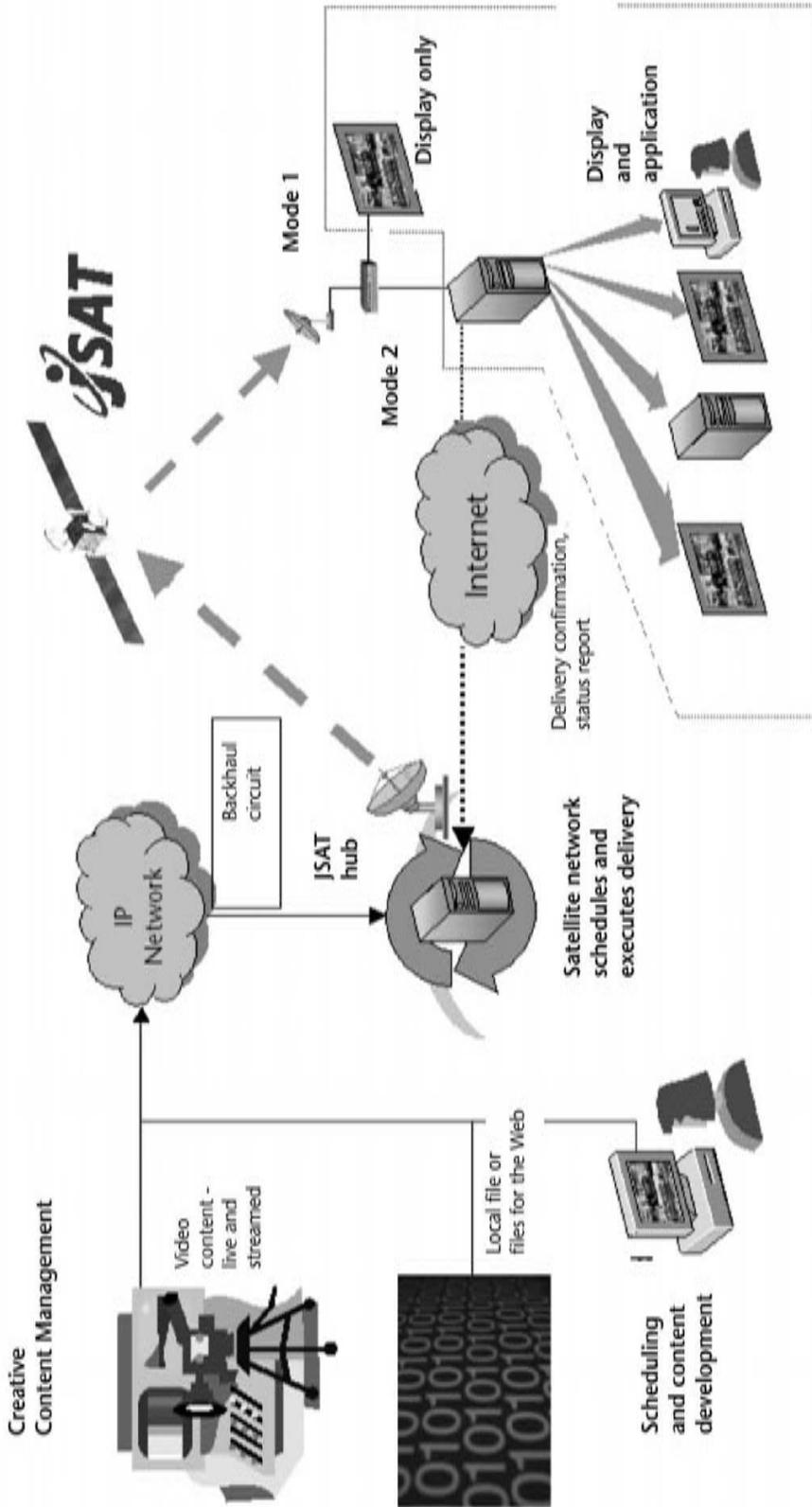
Figure 3.17   A typical DTH satellite broadcasting network. (*Source:* JSAT International.)

Once the content is properly composed and scheduled, it can move from the management site on the left to the uplink "hub" in the middle of the figure. The content is represented in digital form and can be transferred over a leased line or through the Internet provided that the quality of service is adequate for the type of content involved. The hub would be a normal video uplink if it were not for the presence of the computer servers and software needed to play out the content in the most appropriate manner. Systems such as this are available from a number of providers who also support the hub operator through software upgrades and training. The uplink signal to the satellite contains all of the content elements—some in real time and some transferred as files. The satellite simply broadcasts the content throughout the coverage area and it is received using low-cost TVRO terminals like the one shown on the right.

At the retail branch on the right, a standard IRD receives the broadcast and delivers the content over a local area network (LAN) to a variety of interconnected elements. Following either the MPEG-2 or MPEG-4 standard for real-time TV reception, the real-time content can be presented to the public on a large-screen display. This might be hung on the wall or suspended from the ceiling, and located where customers can easily view it. The remaining elements deal with files that come through the IRD in the form of IP packets. Control and scheduling data sent over the satellite with the content is used within the server computer on the right to render the material for viewing on other displays. Also indicated is a PC with a hard drive that allows a local employee or customer to interact with the file information as if in real time. Some potentially valuable applications are employee training and videoconferencing. With an appropriate LAN arrangement, any display can present a hybrid arrangement of real-time and stored content, and can replay material according to the schedule intended by the work station operator on the left of Figure 3.17.

This particular CDN is designed to provide a full range of applications in the retail environment, digital signage being one of them. Other uses of CDNs include software distribution for the enterprise, pay-per-movie transfer to head ends in cable TV, and digital cinema.

3.2.4   Mobile Satellite Communications

The majority of satellite applications discussed up to this point have been to serve fixed points of the ground. While fixed satellite services represent the majority of the commercial business, mobile satellite communications applications have risen to the point of being more than an industry niche. We discuss three main aspects of the mobile sector: interactive two-communications using the Mobile Satellite Service (MSS); Satellite Digital Audio Radio Service (S-DARS), and a hybrid system currently under development called ancillary terrestrial component (ATC). All of these employ the portion of the radio spectrum below C-band known as L-band and S-band. As such, they enjoy somewhat improved propagation in free space as compared to the higher frequencies that are subject to rain attenuation. Complementary portions of L- and S-band are likewise employed by terrestrial wireless services like 3G cellular (EDGE, CDMA 2000, and WCDMA), WiFi (IEEE 802.11x), and WiMAX (IEEE 802.16x).

3.2.4.1 Mobile Satellite Service

The evolution of the MSS segment is discussed in Chapter 2. Here, we focus on the architecture and applications associated with MSS. Owing to the interactive nature of MSS, the dominant use is for voice telephony in a manner similar to cellular telephone. The familiar names of Iridium, Globalstar, and Inmarsat represent the bulk of this business, but they are joined by a new operator in the Middle East called Thuraya. Whether using a low-Earth orbit (LEO) or GEO, these systems generally follow the structure shown in Figure 3.18, which consists of the satellites in the form of a constellation, the mobile user terminals of various types, and a fixed ground segment used to operate the system and provide access to terrestrial networks such as the telephone network and the Internet.

The satellite constellation of Inmarsat is the most mature and has seen four generations of GEO satellites. Their current design is called the Inmarsat 4, two of which were operational, one in the Atlantic and one in the Pacific, at the time of this writing (the Indian Ocean satellite is planned for service). Inmarsat 4 employs a large antenna in space and provides small spot beams to increase performance and capacity. User terminals are primarily of the laptop size, used in what Inmarsat calls the broadband global area network (BGAN). Thuraya employs GEO satellites of similar design although they have offered satellite telephone service to handheld terminals as well as data terminals. It is best practice for the user to determine the direction to the satellite and then to assure an unobstructed path in the particular direction. In this way, once the connection is established, it will most likely be maintained for satisfactory communications.

As is well known, Iridium and Globalstar satellites are in LEO and have orbit periods of under 2 hours. Since the satellites are in constant motion relative to the ground, there must be many satellites available to provide continuous service.
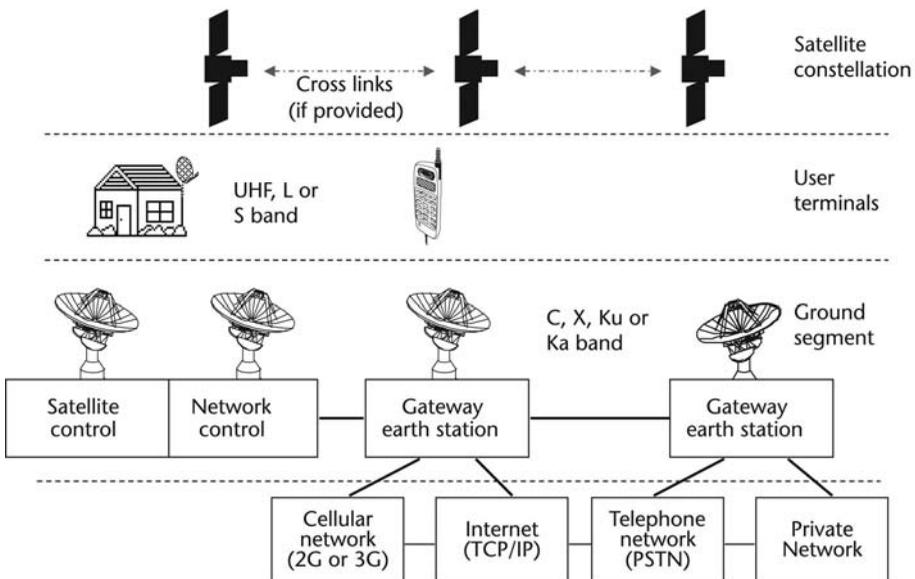

Figure 3.18 An example of a content distribution network in retail service, using a Ku-band GEO satellite. (*Source:* JSAT International, Inc.)

Iridium takes this one step further by using intersatellite links, indicated by the broken arrows at the top of Figure 3.18. Both Iridium and Globalstar were designed to service handheld satellite phones; like all other MSS systems, the user must maintain a clear, unobstructed path to the satellite to assure that the telephone call can be initiated and maintained for its expected duration. With LEO satellites in constant motion relative to the user, this gives the additional requirement that the user be out in the open, such is in a field or on a rooftop. Otherwise, the connection will be broken when the satellite/phone signals are blocked for more than a few seconds. Voice quality on the link is generally below that of a normal cell phone because of the reduced bit rate that can be provided. However, users generally accept the degraded quality because of the ability to communicate where no other reliable means is available.

The fixed ground segment in Figure 3.18 could consist of a single Earth station for a system composed of GEO satellites serving one country. For a global system like Inmarsat, there would be facilities across the globe to provide service in particular countries. Because of the intersatellite links, the Iridium system only requires one Earth station, which is what existed for some years following bankruptcy and reorganization under new owners. Globalstar does not have intersatellite links and so relies on gateway Earth stations for each country being served (and multiple Earth stations across large countries like the United States and Russia because of the range limitation of a single satellite). The fixed Earth stations can employ a variety of frequencies for the feeder and control links to the satellites, such as C-, X-, Ku-, and Ka-bands. GEO satellites can be served from limited motion antennas while the LEO satellites require full-tracking antennas on the ground (including three to four at the same location for handoff).

MSS systems have an association with global position services, like GPS and GLONASS. The mobile user may need to be located so that the appropriate charges for service are made. Also, the phone instrument itself can be used to accept the input from a GPS receiver and relay position data back through the satellite network to a central dispatch point. This allows the tracking of vehicles, ships, and aircraft on a reliable basis. The satellite system can also be used to improve the accuracy of position measurement by sending correction information to the mobile user.

3.2.4.2   Satellite Digital Audio Radio Service

S-DARS is in the process of joining DBS as a major segment of the satellite communications industry. In the United States, S-DARS has achieved nearly 10 million subscribers and continues to penetrate both the vehicular and home markets. This is probably because of the wide array of music and talk channels offered on a consistent basis throughout the country. Other S-DARS systems are operating in Asia and Africa, although neither has achieved the level of acceptance as in the United States.

The architecture of a typical S-DARS network in Figure 3.19 consists of at least two high-power satellites that downlink the broadcast at either L- or S-band. Two orbital arrangements have been applied: standard GEO and a highly elliptical inclined orbit (HEO) with a 24-hour period. As illustrated in Figure 3.20, the first generation Sirius Satellite Radio system has three satellites in HEO at 120-degree
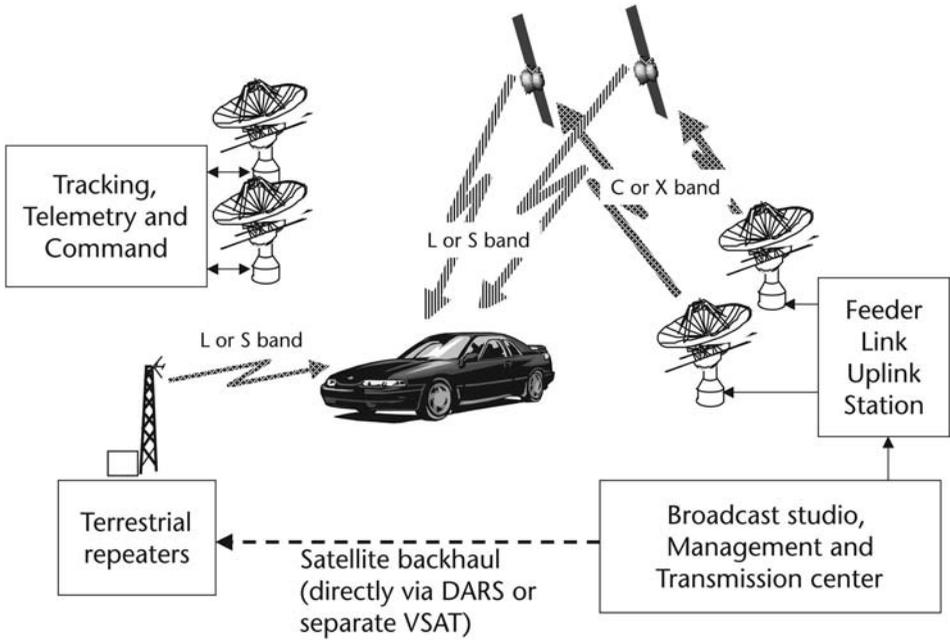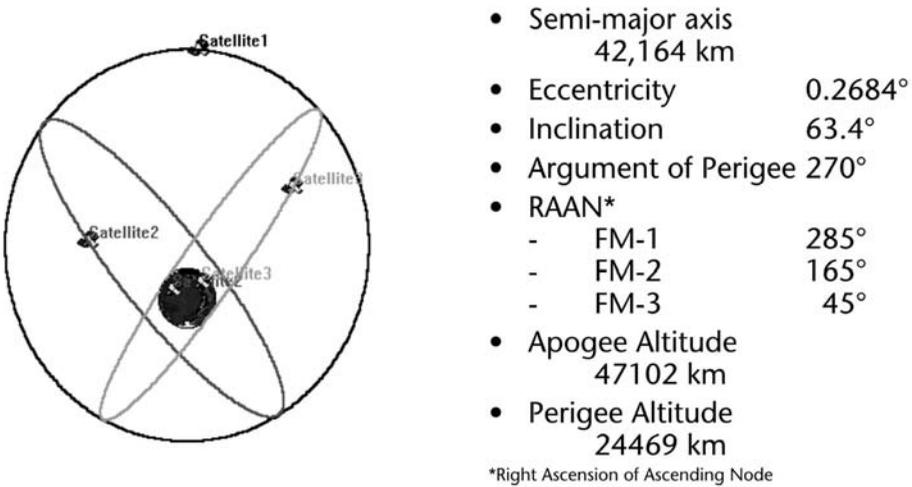
Figure 3.19   Mobile Satellite Service segments.

- Semi-major axis
  42,164 km
- Eccentricity   0.2684°
- Inclination   63.4°
- Argument of Perigee 270°
- RAAN*
  - FM-1   285°
  - FM-2   165°
  - FM-3   45°
- Apogee Altitude
  47102 km
- Perigee Altitude
  24469 km

*Right Ascension of Ascending Node

Figure 3.20   Basic arrangement of a Satellite Digital Audio Radio Service (S-DARS).

offsets that sequentially provide coverage of the United States at elevation angles greater than 60 degrees. The latter is illustrated with the 24-hour timeline to the right of the orbit diagram. In its second generation, Sirius has elected to launch a GEO satellite with substantially more power than the three HEO satellites, the intention to provide backup and to improve coverage to users who wish to receive a consistent signal through the window of a home or other building.

Having two satellites in view in the sky at one time provides a higher assurance of an unobstructed path. Propagation studies have shown that this is equivalent to approximately 10 dB of additional fade margin, which would not otherwise have been practical. Service into urban areas is further enhanced through the use of terrestrial repeaters that retransmit the same broadcast at much higher power than would be possible from space. Two approaches are used to backhaul the broadcast to the terrestrial repeaters: downlinking directly from the GEO satellites to the repeater, or in the case of HEO, using an independent FSS satellite transmission to low-cost dishes that transfer the signal to the repeater.

The content for S-DARS is assembled at the broadcast facility that houses many sound studios, downlinks for talk and news programming, and servers containing recorded music. The management of the system is automated to the point that a disc jockey need only show up at the appointed time in the proper studio and from there compile a broadcast. Business management functions for S-DARS are essentially the same as for DBS, building upon that model of providing service and distributing equipment to consumers. In addition, the S-DARS operators were successful in signing up the automobile manufacturers who preinstall the receivers in vehicles.

3.2.4.3   Ancillary Terrestrial Component

The previous two mobile satellite communications applications would appear to be different in terms of their target markets. MSS is aimed at users of two-communications services, including mobile voice and data, and S-DARS is for one-way broadcasting of music and other content. Both are valuable in their own right; however, the potential for marrying features of one with the other has become of great interest in the United States and Europe. Thus was born the ancillary terrestrial component (ATC). The basic principle of ATC is to offer two-way communications like MSS and to use terrestrial base stations to extend service into suburban and urban areas, like the terrestrial repeaters of S-DARS. Figure 3.21 provides an illustration of one such arrangement of a hybrid terrestrial/satellite wireless network employing two GEO satellites along with large terrestrial cell sites similar to what is found in a 3G cellular network. The key is to employ the same spectrum, L-band in this example, and a common technical "air interface" so that the user devices are inexpensive and easy to use.

The U.S. Federal Communications Commission issued its Second Report and Order, laying out the objectives for ATC:

> In deciding to permit MSS operators to integrate ATC into their MSS systems, we stated that 'permitting MSS licensees to enhance spectrum efficiency through ATC represents a superior choice to continuing with the regulatory status quo.' Further, we found that MSS/ATC would expand the consumer market MSS is capable of serving, leading to economies of scale and lower prices for consumers. Granting authority for integrated MSS/ATC would also allow MSS operators to offer the capability of receiving calls via both satellite and terrestrial radio links with a single dual-mode handset and a single telephone number, whereas current dual-capability MSS/Commercial Mobile Radio Service (CMRS) handsets require separate numbers for MSS and CMRS services. Moreover, MSS/ATC would eliminate operational
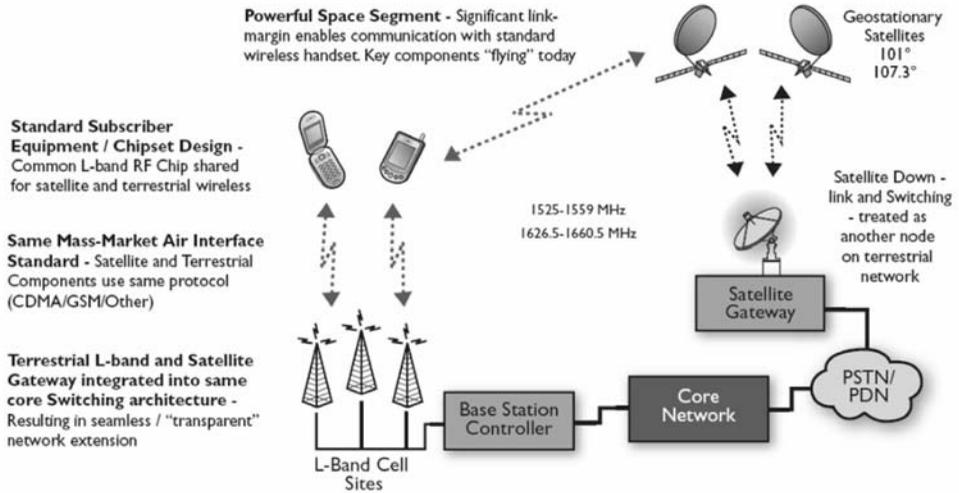
Figure 3.21   The first generation Sirius Satellite system employs three satellites that provide a continuous coverage of the United States at elevation angles greater than 60 degrees with three satellites operating [3].

and transactional difficulties and costs for MSS operators in negotiating separate terrestrial roaming agreements with various terrestrial CMRS operators within the MSS operators' service areas. Finally, we found that MSS/ATC would enhance the ability of the national and global telecommunications systems to protect the public by offering ubiquitous service to law enforcement, public aid agencies, and the public, and would strengthen competition in the telecommunications market.

Three U.S. companies are in the process of constructing very large GEO satellites to pursue this vision. To make this happen, there will need to be an extensive build-out of another terrestrial infrastructure to provide service where the satellite signal will not be adequate. Both S- and L-band frequencies have been authorized by the FCC for the United States. In Europe, there is the possibility that a system will be developed as well.

3.3   Point-to-Point Networks

Point-to-point connectivity finds many uses where good-quality terrestrial links are not available. With this type of connectivity a pair of Earth stations transmit RF carriers one to the other. Because both stations simultaneously receive the respective carriers, the link is full duplex. In other words, the users being served can talk or transmit information in both directions at the same time. The uplink section of the satellite repeater receives both tranmissions and, after translation to the downlink frequency range, transmits them back toward the ground. If the satellite provides a single footprint covering both Earth stations, then a given station can receive in the downlink its own information as well as that of its communicating partner. That supplementary ability provides a unique way for stations to verify the content

and quality of their own satellite transmission. (That would not be the case if the stations were in different beams provided by the same satellite.)

A popular application of point-to-point links is to backhaul telephone calls from remotely located cell towers.

The use of separate frequencies is called frequency division multiple access (FDMA). Alternatively, the Earth stations can transmit their data on the same frequency but separated in time. That technique, in which the transmissions are in terms of limited duration bursts, is TDMA. FDMA and TDMA represent a pair of alternatives that can isolate Earth stations that share a common transponder channel or piece of RF bandwidth. It is possible to combine the two multiple access modes, for example, TDMA/FDMA, where there are multiple TDMA networks, each operating in a different narrow frequency channel. The reason for doing it that way is to reduce the required uplink power or transponder bandwidth for a small TDMA network. A third technique, called code division multiple access (CDMA), permits the stations to transmit on the same frequency at the same time. All these techniques are reviewed in Chapter 5.

A satellite mesh network provides several links to connect multiple Earth station locations. The maximum number of possible links (a full mesh) between N Earth stations is equal to $N(N-1)/2$. Figure 3.4 illustrates a mesh network with N equal to 4. A point-to-point link that is continuously maintained is called dedicated, or preassigned, bandwidth. They typically are used for special applications like cellular backhaul or high-speed data links in computing networks.

3.4   VSAT Networks

VSAT networks have helped many businesses solve difficult network problems though the use of communication satellites. The basic principle it to start with data broadcasting to remote sites and to add a return link over the same satellite, bypassing the terrestrial network. As shown in Figure 3.3, multipoint interactive connectivity provides two-way communications because the remote VSATs receive the broadcast from the central hub station and can transmit back typically at a much lower throughput data rate. It is different from a point-to-point network because the remote stations cannot communicate directly with one another but must do so through the hub. The generally accepted principle is that the data broadcast, called the forward link, continuously transmits information to all VSATs, which in turn utilize burst transmissions, called return links, so that several can share a common frequency channel. There would likely be multiple frequency channels to aggregate sufficient traffic capacity to match the forward link.

The fact that VSAT return links are received only by the hub means that their transmit power is quite low. In addition, the forward link is at a high power level to permit reception by the small VSAT antennas. There is, therefore, an economical balance between the high cost of the hub (which is shared among the VSATs) and the low cost per VSAT (installed at each remote served location). The best applications of VSAT star networks are where there is need for remote locations to access a headquarters host computer or LAN and where remote to remote

communication is limited. Low cost VSATs of this type also offer economical access to the Internet.

The VSAT is relatively inexpensive and easy to operate, justifying its use on the customer premise. Figure 3.22 shows how a VSAT is used in a multipoint-to-point network; Chapter 9 contains a detailed discussion of the typical VSAT and hub equipment configuration. With an antenna size of 80 cm (30 inches) and a purchase price comparable to that of a PC, a VSAT gives full access to a private network involving thousands of locations. Such networks would be prohibitively expensive if implemented with dedicated point-to-point private lines, whether terrestrial or satellite.

The key to the effectiveness of a VSAT network is its interactive nature, allowing two-way communications from remote locations in the same manner as the terrestrial telephone network, provided, however, that a star arrangement is appropriate for the application. If a true mesh connectivity is required, each station must be capable of direct communication with every other.

Star networks are easily implemented using VSATs because the low sensitivity of the small dish is compensated for by the high sensitivity of the large antenna and high-power amplifier at the hub. That is why the first VSAT networks to be introduced in North America all employed hubs in a star architecture, which continues to be the majority of installations around the world.
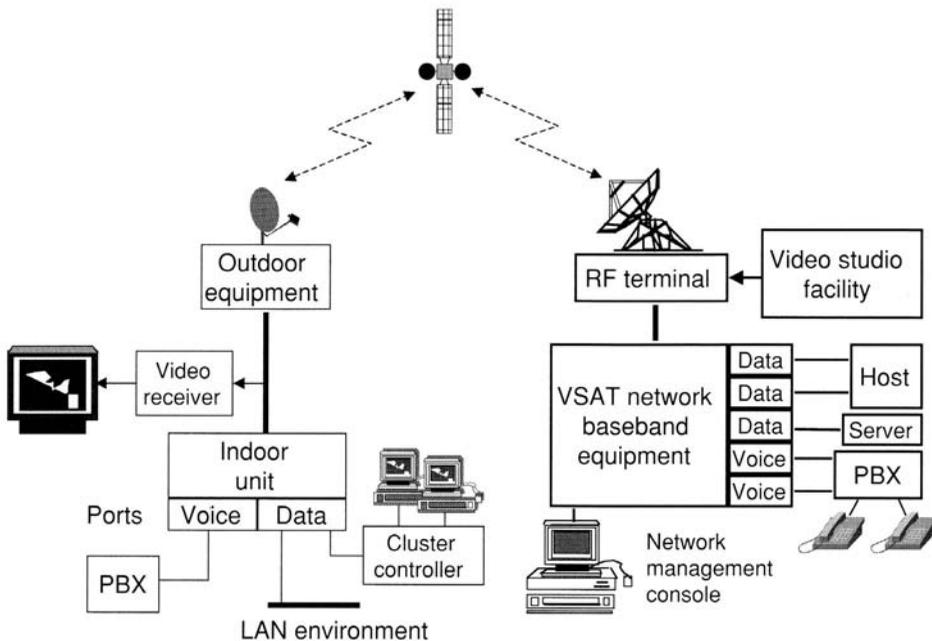
Figure 3.22   A hybrid terrestrial/satellite wireless network employing the ancillary terrestrial component (ATC) or complementary ground component (CGC) concept.

References

[1]   Elbert, B. R., *The Satellite Communication Applications Handbook*, 2nd ed., Norwood, MA: Artech House, 2004.

[2]   Palter, D. C., *Satellites and the Internet—Challenges and Solutions,* West Covina, CA: SatNews Publishers, 2004.

[3]   Briskman, R. D., et al., "S-DARS Broadcast from Inclined, Elliptical Orbits," IAF-01-M.5.04, *52nd International Astronautical Congress,* Toulouse, France, October 1–5, 2001.

Microwave Link Engineering

Microwave link engineering is the branch of communication engineering that deals with the analysis, design, implementation, and testing of radio paths in the atmosphere or in space, operating in the range of about 1 to 100 GHz. Satellite and other microwave radio links obey certain rather predictable laws of nature, allowing engineers to design satellite networks on paper (or, more likely, on computer) with good accuracy. An unobstructed line-of-sight path provides the most predictable transmission link; however, various forms of signal absorption (attenuation), blockage, and multipath propagation (caused by reflections) introduce time-varying fading, which can best be estimated using statistical techniques. Nevertheless, the service obtained from microwave links between the satellites and ground stations can be very acceptable, even when compared to terrestrial cable and radio (wireless) systems.

This chapter is intended for engineers and scientists new to the field and is perhaps the most mathematical chapter in the book. Nontechnical readers may wish to scan it rather than delve into the details. In general, microwave energy travels in straight lines through space in the same manner as a light beam. That should not be a surprise because both light and microwaves are forms of electromagnetic energy, light being at much higher frequencies or, equivalently, shorter wavelengths. Radio waves in general and microwaves in particular can propagate through space like light and heat, spreading out as they move farther and farther from the source. The waves travel through a vacuum at the universal speed of light, the same for all forms of electromagnetic energy. Microwaves can be kept from propagating freely by forcing them to travel in a closed metallic pipe called a *waveguide*. The physical similarity to common household pipes has led to the use of the term *plumbing* when referring to the waveguides found in Earth stations and satellites. At the input end of the waveguide would be found the transmitter, the electronic device that generates the microwave energy on the proper frequency, with the other end connected to an antenna such as a parabolic "dish." Waveguides are similarly used on the receiving end.

4.1 The Decibel

Microwave power is measured in watts, the same unit we associate with electric power used for lighting and motors. Link effects that either reduce or increase that power almost always are expressed in terms of the decibel (dB), which is why it

is important to introduce the concept early in the discussion. Converting a quantity to decibels simply means taking the logarithm to the base 10, or common logarithm, and multiplying the result by 10:

$$A \text{ (in dB)} = 10 \log_{10} (a)$$

where a is a real, nonnegative number. In actuality, a is a ratio, such as the power gain of an amplifier, that is,

$$a = P_{out}/P_{in}$$

A sample of decibel values versus their equivalent ratios is provided in Table 4.1. Recall that adding logarithms is equivalent to multiplying ratios. So, for example, adding 4.0 dB is equivalent to multiplying by a factor of 2.51. In the table, relative power changes are computed to the closest tenth of a decibel (i.e., 0.1-dB precision) because measurement accuracy on an actual satellite link is about 0.5 dB at best. Since the logarithm of the number 1 is equal to zero, a zero-dB change simply means there has been no change in the power level of the signal. You may recall that logarithms are exponents, and in the case of logarithms to

Table 4.1   Example of a Decibel Table, Indicating the Ratio, Percentage Increase, and Corresponding Decibel Value

Power Ratio	% Increase	dB
1.0	0	0
1.12	12	0.5
1.26	26	1.0
1.41	41	1.5
1.59	58	2.0
1.78	78	2.5
2.00	100	3.0
2.24	124	3.5
2.51	151	4.0
2.82	182	4.5
3.16	216	5.0
3.55	255	5.5
3.98	298	6.0
4.47	347	6.5
5.01	401	7.0
5.62	462	7.5
6.31	531	8.0
7.08	608	8.5
7.94	694	9.0
8.91	791	9.5
10.0	900	10.0
20		13.0
50		17.0
100		20.0
200		23.0
500		27.0
1000		30.0

the base 10, they are exponents of the number 10. The inverse of a decibel is expressed as:

$$a = 10^{A/10}$$

which again is a ratio.

A decrease would be represented by a negative decibel value. For example, reducing a power to one-half its value (e.g., multiplying it by 0.5) is the same as subtracting 3 dB (e.g., adding −3.0 dB). Another consequence is that a decrease of 1.0 dB (adding −1.0 dB) represents a decrease of 20%. Table 4.1 indicates that adding 1.0 dB is an increase of 26%.

An inexpensive scientific calculator is the most convenient device for converting back and forth between ratios of powers and decibels. Just be sure to use the common logarithm (log) button and the 10^X button, where appropriate. This is safer than trying to memorize Table 4.1 (although such knowledge is the sign of a "dB artist").

The following are a few simple examples of the decibel approach. A 3-dB difference in power level is nearly equal to a factor of 2, while a factor of 10 is exactly equal to 10 dB, that is, 10 log (10) = 10. In satellite communications, it is not uncommon to encounter a factor of 10^{-20} decrease in signal power. That can more conveniently be expressed in decibels as a 200-dB reduction, called a loss. Gains are added, while losses are substracted. When dealing with factors in an equation, it is convenient to first convert each factor to decibels. Then the equation can be solved by summing the decibel values instead of having to resort to multiplication. Likewise, division becomes simple subtraction, as in the example of the 200-dB loss. That is the basis for the link power balance calculation and the link budget, reviewed later in this chapter.

4.2   Propagation on the Earth-Space Link

The process by which radio signals reach the receiving antenna from the transmitting station is called radio wave propagation. Radio waves represent a part of the electromagnetic spectrum, ecompassing radio, infrared, visible light, ultraviolet, and X rays (given in increasing order of frequency or decreasing order of wavelength).

The part of the electromagnetic spectrum that is of interest in commercial satellite communications lies between 1 and 60 GHz, as illustrated in Figure 4.1.
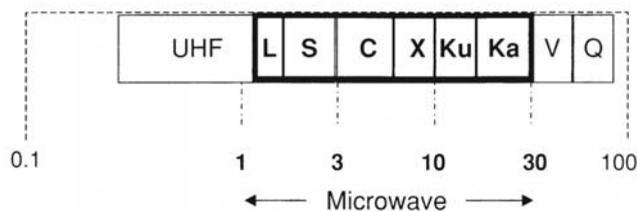
Figure 4.1   The microwave spectrum, indicating the approximate location of satellite bands.

The relative merits of each band segment were covered in Chapter 1. Using the popular letter designations for the various segments, the spectrum offers a total bandwidth of approximately 60 GHz. The frequencies below 30 GHz are those currently in use or under development; the frequencies above 30 GHz are a still untapped resource. RF bandwidth can be used and reused many times by multiple satellites and the radio beams that they can generate over portions of the Earth. The following sections review the propagation, generation, and reception of microwave signals, in preparation for studying the unique aspects of satellite communications presented in Chapters 5, 6, 7, and 9.

4.2.1 Basic Microwave Propagation

Once it is several emitter dimensions away from its emitter, electromagnetic wave energy propagates out radially. The spreading pattern decreases in intensity inversely to the square of the distance. All radio waves behave in this manner in free space, but various forms of matter produce interesting and potentially disruptive results when placed in their path. That is because microwave energy can be absorbed, scattered, bent, and reflected, as indicated in Table 4.2. As a result, there is additional loss of signal power, in decibels, which must be accounted for in the link design.

A question that comes to mind is, "How does such a wave, traveling at the speed of light, come into existence in the first place?" Because space consists of nothing, a radio wave does not propagate like the crest of a wave in the ocean. Instead, it usually is the result of the high-frequency vibration of electrons in a piece of wire or other conducting material of appropriate dimensions. Alternating current from a transmitter causes the electrons to vibrate back and forth; by not flowing continuously in one direction, the electrons lose most of their energy by throwing it off into space. That is illustrated in the drawing sequence provided in Figure 4.2.

A high transmitter power, say, 1,000W (1 kW), produces a correspondingly high level of microwave radiation from the antenna at the point of exit into space. That is the same principle behind a microwave oven, which uses a high-power

Table 4.2 Propagation Effects to Introduce Additional Attenuation of Microwave Signals as They Propagate Between the Transmitting Station and Receiving Station

Propagation Mode	Effect	Source Medium
Absorption	Conversion of microwave energy into simple heat	Air, water vapor, precipitation (rain and snow), nonmetallic structures, and foliage
Scattering	Multiple reflections and bending due to minute particles	Water vapor, precipitation, dust, ion plasma (the ionosphere)
Refraction	Bending of microwaves due a change in the refractive index	Air, ion plasma
Reflection	Total reflection of the microwave signal	Temperature inversion layers (ducting), metal structures, bodies of water, buildings, and flat terrain
Diffraction	Bending around physical obstacles	Metal structures, buildings, and partial terrain blockage
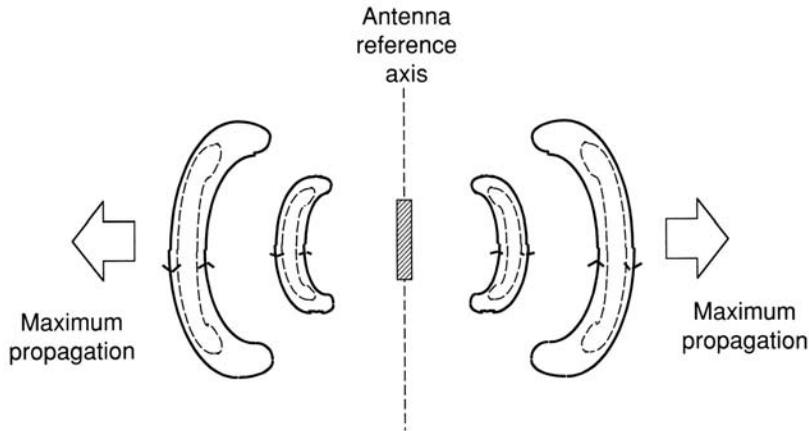
Figure 4.2   The fundamental concept of electromagnetic wave propagation from an elemental antenna. The packages of energy are separated by one wavelength in distance.

magnetron amplifier to generate an intense field for cooking purposes. The field causes the molecules in the food to oscillate at the same frequency; that generates the necessary heat to raise the temperature. Such intense radiation around a high-power microwave transmitting station can "bake" you if you come too close. At sufficient distance from a transmitting antenna, the microwave energy induces less heating. There is a defined area around the antenna, called the *near-field region*, where the energy has not quite coalesced into a clearly defined beam. Within the near field, microwave energy varies in intensity, depending where one stands relative to the antenna structure. At sufficient distance from the antenna, the radiation field is formed into the type of beam associated with the particular antenna struc-ture, called the *far-field region*. A parabolic reflector antenna has a sharply defined far field beam that is aimed at the satellite. The transition between the two regions is roughly defined by the following equation:

$$R_f \approx 2D^2/\lambda$$

where R_f is the approximate distance, D is the diameter of dish reflector in meters, and λ is the wavelength in meters. This equation indicates something unexpected: the distance to the transition increases with frequency rather than decreasing. For example, proper testing of a Ku-band reflector antenna demands that the transmitting (or receiving) source be farther away than if the same reflector were to be used at C-band. From the beginning of the far field, the radiated signal keeps on decreasing in intensity by the inverse of the square of the distance from the source (the same property is visible light from a point source such as a star). This particular concept is the basis for accurate link design.

4.2.2   Isotropic Radiator

The most fundamental type of radio antenna is the isotropic source, which is analogous to the lightbulb illustrated in Figure 4.3. At a fixed radius from an
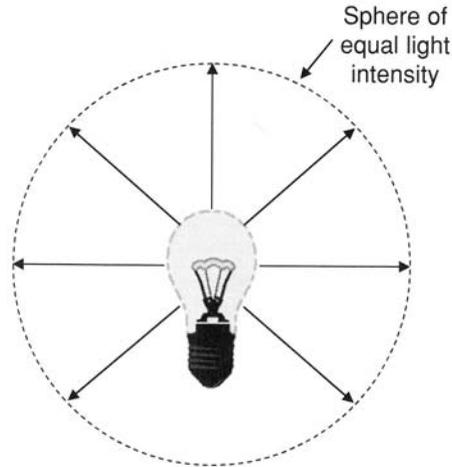
Figure 4.3   An isotropic source can be represented by a lightbulb radiating energy in all directions
with equal intensity.

isotropic source, which defines a sphere, the energy intensity is constant regardless
of direction. The area of the sphere of uniform received energy is equal to the
constant π (3.14159...) multiplied by 4 multiplied by the square of the radius, that
is, $A = 4\pi R^2$. It is common practice to measure the signal intensity at the particular
radius in units of watts per square meter, calculated by dividing the power of the
isotropic source by the area of the sphere in square meters:

$$P/A = P_t/4\pi R^2, \qquad \text{W/m}^2$$

The denominator, $4\pi R^2$, is sometimes called the spreading factor.

RF power driving the isotropic source produces a constant power density at a
fixed distance. That density decreases by $1/R^2$ as the point of reception moves
farther away from the source. Ignoring losses, it is theoretically possible to capture
all the transmitted power by collecting it with a closed surface around the source
(the physical size of the collector is immaterial as long as it closes on itself).

Receiving antennae, therefore, work by way of the area that they expose to
the RF energy radiating from the source. This is illustrated in Figure 4.4 by an
isotropic source that radiates energy equally in all directions, shown as equally
spaced rays. There are two receiving antenna surfaces of equal area: one at distance
R_2 is farther away from the source than the other at distance R_1. Notice how the
closer antenna intercepts considerably more power than the more distant area.
That illustrates how a radio signal becomes weaker as the receiver is moved farther
from the transmitter. It also demonstrates the important concept of capture area,
that is, the relationship between the effective area of an antenna and the strength
of the signal received by it. The challenge of antenna design is to maximize the
fraction of the energy that the antenna actually delivers from the reflector surface
to the receiver. The key parameter is the efficiency of the antenna, defined as the
ratio of the effective area (i.e., the area that would perfectly capture the same
amount of energy) to its physical area. Typical values of dish antenna efficiency
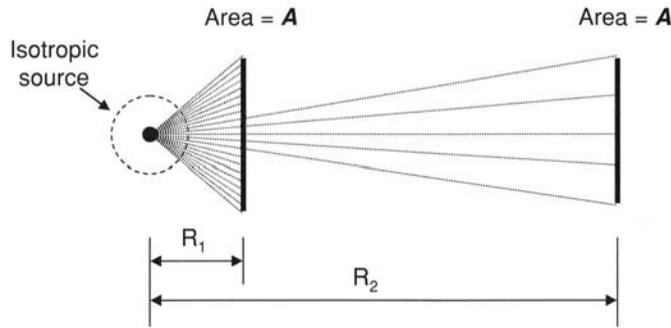
Figure 4.4   The radio energy captured by a fixed antenna area decreases as the distance to the
source increases.

are between 55% and 70%. Antennas that do not use reflectors but employ an
active receiving area to the incoming wave can achieve up to 90% efficiency.
Examples include horns and arrays of elemental antennas.

4.2.3   Directional Properties of Antennas

The isotropic antenna is neither practical (its ideal characteristic cannot be achieved
with a simple physical structure) nor particularly useful for satellites (due to its
low gain and inability to discriminate directions). What we rely on for receiving
and transmitting signals is the concept of directivity, which simply means that the
antenna has the ability to focus the energy in specific directions. Energy that would
have been radiated in unnecessary directions around the sphere is concentrated by
the structure of the antenna and redirected to increase the intensity in the desired
direction.

4.2.3.1   Directivity and Gain

The directivity in a given direction can be expressed as the ratio of the measured
signal to the maximum signal in the peak direction. The gain, on the other hand,
is an absolute measure, obtained by comparing the signal from the antenna to that
of an isotropic radiator. Our absolute measure of maximum gain could be as little
as 10 or greater than a factor of 1 million. When expressed in decibels, these
particular values would be 10 dBi and 60 dBi, respectively, where the "i" indicates
that the gain is with respect to our isotropic antenna. That implies that an isotropic
antenna has a gain of 0 dBi. There must be a fixed offset in decibels between the
absolute value of gain and the corresponding relative value of directivity, because
our definition of directivity is that it starts at a value of 0 dB at the peak and
follows the pattern in a negative-going sense for all other angles.

Another important property of an antenna is called reciprocity, which means
that the gain and the directivity are the same at a given frequency whether it is
used to receive or to transmit. That allows the antenna to receive with precisely
the same directional characteristics as it transmits. Figure 4.5 illustrates a simplified
antenna gain pattern for transmission or reception with a main beam (the region
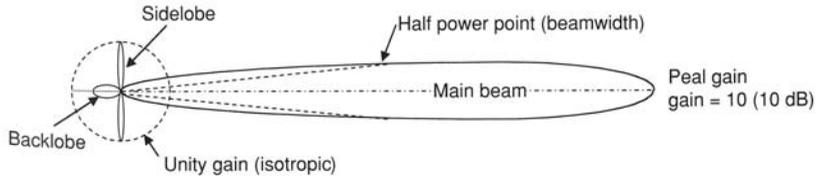
Figure 4.5   Gain of a high-gain antenna expressed as a power ratio and presented as a 360-degree polar plot.

of maximum directivity) oriented toward the right. For comparison, the uniform pattern of an isotropic source is superimposed to scale on the directional antenna pattern. In addition to a main beam, every real antenna operates in undesired directions, shown in the figure as a pair of sidelobes and a backlobe. The maximum gain, also called the peak gain, is indicated at the center, and the backward direction is indicated at ±180 degrees.

4.2.3.2   Beamwidth and Sidelobes

There are a number of other useful definitions of antenna performance besides the peak gain. The half-power beamwidth (often called simply the beamwidth) is the angular width of the main beam measured between the points where the power intensity is one-half that of the peak. An equally accurate name that is often used is the 3-dB beamwidth, since the half power point is where the directivity is 3 dB down. Assuming that the microwave link can still function with a 3-dB decrease in signal strength, the half-power beamwidth defines the range of antenna pointing (alignment angle) over which the antenna or satellite can move. It is a common practice, however, to allow only a 0.5-dB (12%) drop in signal power, which demands either tighter antenna pointing accuracy or satellite position control.

When installed at an Earth station, the antenna is attached to a mount, which may allow the beam to be repointed. Simple RO antennas used for direct reception usually are fixed to their mounts and need only be pointed toward the satellite when initially installed. Movable antennas usually have motorized mounts to allow remote repointing; that could also be part of an automatic tracking system to initially locate (acquire) a satellite and to keep the beam aligned with (track) the satellite. It is normal to refer to the elevation angle of the direction of pointing, where zero is for pointing at the horizon and 90 degrees corresponds to directly overhead (zenith). The mount also points the antenna in azimuth, that is, along the directions of the compass. Most mounts work directly in terms of elevation and azimuth, but some are polar and behave like an astronomer's telescope. The advantage of a polar mount is that it can repoint along the equatorial orbit plane by adjusting only one axis.

Presentations of antenna performance are called antenna patterns and are illustrated in Figure 4.6 (a hypothetical example) and Figure 4.7 (a real 10-m antenna) in terms of gain and directivity, respectively. Almost every antenna has a backlobe in the opposite direction from the main beam. However, as shown in Figure 4.6, the gain of the backlobe can be made to be less than unity, in this case, producing a negative gain of −2 dBi or lower. Sidelobes and backlogs are important
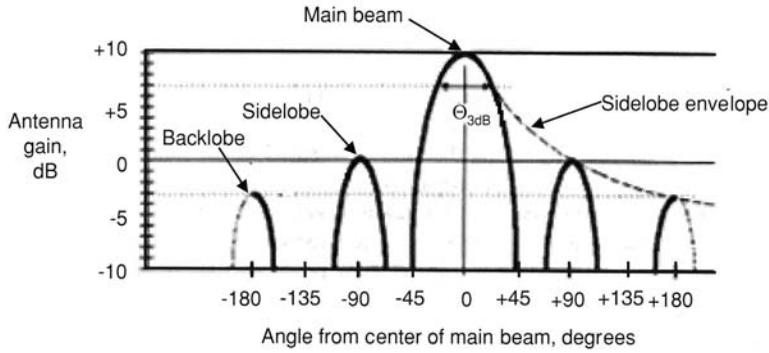
Figure 4.6 Gain of a directional antenna in decibels for all angles with respect to the peak of the main beam indicating the worst-case envelope of sidelobes.
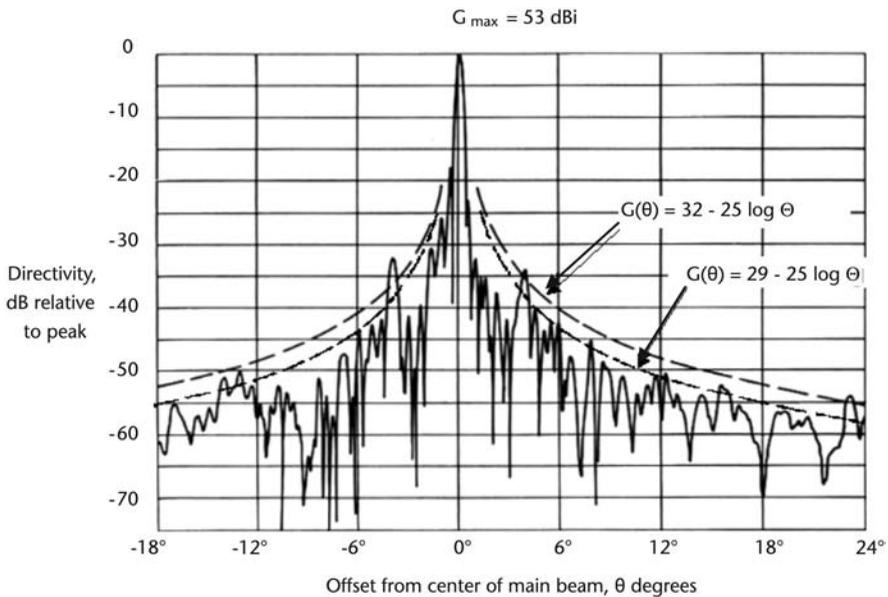
Figure 4.7 A comparison of the radiation (directivity) pattern of a real 10m C-band antenna with the ITU sidelobe envelope.

characteristics of Earth station antennas when one is considering transmit or receive interference. Figure 4.7 includes a smooth curve, called a sidelobe envelope, that defines a specification of maximum sidelobe gain. Two such envelopes are shown, the lower being a tighter specification that applies between 1° and 20° off the peak of the main beam.

4.2.3.3 Isolation

The directive property of an antenna dertermines how effective it will be for getting signal power from the source to the receiver. However, any link can be degraded by signals on the same frequency that enter the receiving antenna from a direction

other than along the main beam. Likewise, a transmitting station can cause degradation to other systems by sidelobe radiation. Any undesired signal that can potentially degrade reception is RFI. There is a whole field of engineering study that focuses on the identification of sources of RFI, the establishment of criteria for acceptable operation in the presence of RFI, and the development of techniques for countering its effects. Techniques such as beam shaping, cancelation, and shielding are effective in that regard.

The presence of RFI is a consequence of the fact that all radio communications systems reuse frequencies, that is, there will be more than one radio station operating on any particular frequency at any particular time. What keeps the RFI within acceptable limits is isolation that is either natural (e.g., geographical or angular/orbital separation) or artificial (e.g., measures such as RFI shielding or beam cancellation). Satisfactory operation of independent microwave users on the same frequency often requires cooperation (called frequency coordination) because ''one user's radio link is another user's RFI.'' Several satellites can operate in the geostationary arc in the same frequency band because each directional ground antenna can focus on one particular satellite, suppressing the RFI produced by adjacent satellites (see Figure 1.26).

For example, if we assume the receiving user's antenna has the characteristic given in Figure 4.6 and the interfering satellite is located 45 degrees away from the satellite that we wish to receive (the desired satellite), then the RFI is suppressed by at least 20 dB. Actual ground antennas used in FSS and BSS links provide 40 to 60 dB of peak gain with suppression at 45 degrees of 50 to 70 dB. At closer-in angles, the isolation may amount to only 20 to 30 dB, which is adequate to allow satellite spacing as small as 2 degrees. The first sidelobe, in particular, is normally the strongest, being typically 15 to 20 dB down from and within 1 degree (more or less) of the peak gain. The control of those sidelobes demands careful design and installation of such antennas. A useful specification in that regard is the sidelobe envelope (shown in Figure 4.7), which defines the worstcase potential for RFI. A standardized formula, which was adopted by the ITU as Recommendation S.465-5, provides a common standard for the larger ground antennas used at C and Ku bands. In it, gain at a particular off-axis angle is specified in the direction of a potentially interfering (or interfered-with) satellite by:

$$G(\Theta) \le 29 - 25 \log_{10} \Theta, \qquad \text{dBi}$$

where Θ is the offset angle between the direction of the main beam and that toward the interfering or interfered-with satellite, for angles between ±1 degree and ±37 degrees. Furthermore:

$$G(\Theta) \le 8 \text{ dBi, for angles between ±7 degrees and ±9.2 degrees}$$

For angles between ±9.2 degrees and 48 degrees:

$$G(\Theta) \le 32 - 25 \log_{10}(\Theta), \text{ dBi}$$

Lastly, $G(\Theta) \le -10$ dBi for $\Theta > 48°$.

4.2.4   Polarization (Linear and Circular)

Up to this point, we have dealt with radio propagation as if it were pure energy. However, there is a property of an electromagnetic wave called polarization, that depends on the orientation (or angle of rotation) of the transmitting antenna. You may be familiar with polarization as it relates to light. For example, true three-dimensional (3-D) movies utilize vertically and horizontally polarized light to simultaneously project appropriate left and right images on the screen. The images are separated by invisible polarizing grids imbedded in the lenses of the viewer's glasses, where one lens is aligned vertically while the other is aligned at a 90-degree angle (horizontally). Vertical and horizontal polarization, therefore, can ''reuse'' (i.e., use twice) a transmission path, such as the projection of a movie on a screen or, as described below, a radio path.

4.2.4.1   Linear Polarization

The concept of polarization discrimation in radio communications is illustrated in Figure 4.8. Shown at the top of the figure is a type of simple wire or rod antenna called a dipole. Electrical current from the transmitter flows along the rod first upward and then downward, oscillating at the frequency of transmission. At C-band downlink frequencies, the rate of oscillation is 4 billion times per second (i.e., 4 GHz). The alternating current in the rods produces an electromagnetic wave that propagates off into space (Figure 4.2). A dipole is not a true isotropic source because there is no radiation in the direction the rods point. Instead, what is formed is a doughnut-shaped pattern that is aligned horizontally. The electrical currents in the rods cause the electromagnetic wave to have its electric component to be lined up in the same direction, which is vertical (in the direction of the two arrows). This type of polarization is called linear polarization (LP) because the electric component has a fixed orientation. Horizontal LP is obtained when the dipole is rotated 90 degrees, so that the direction of the electrical current also is horizontal. Reception occurs when the electric component of the incoming wave produces a
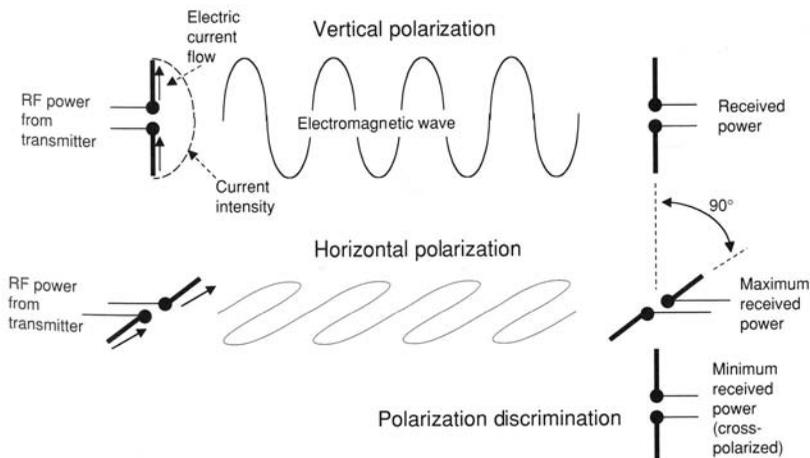
Figure 4.8   Properties of linear polarization as radiated and received by dipole rod antennas.

current in the receiving antenna, which cannot occur if the conductors of the receiving antenna are perpendicular with the incoming polarization. In the lower half of Figure 4.8, horizontally polarized transmitting and receiving antennas provide for a maximum amount of power to be carried (coupled) between them. A vertically polarized receive antenna, which is perpendicular to and therefore cross-polarized with the transmitter, minimizes the amount of coupled energy.

Microwave antennas usually make use of waveguide structures and solid reflecting surfaces because they are much more efficient and predictable than wires and rods. For example, if one encloses the dipole antenna in a metal box, it can feed a waveguide. Rectangular waveguide, illustrated behind the horn in Figure 4.9, is simple in design and efficient in the transmission of microwave energy and hence is very popular in satellite and Earth station design. The electric component is vertical for the orientation shown, extending between the upper and lower side walls (dimension b). The magnetic component lies in the plane of the top and bottom walls and is perpendicular to the electric component at every point along the waveguide. The horizontal dimension of these walls (dimension a) must be greater than one-half wavelength at the operating frequency. At any lower frequency, the wave cannot propagate because its wavelength literally cannot fit. This is termed waveguide beyond cutoff. The narrower dimension, b, is not as critical and is adjusted like a transformer to match the characteristic impedance of the waveguide to the antenna.

The relationship between horn antenna orientation and wave polarization is shown in Figure 4.9, which basically is a slightly flared piece of rectangular waveguide. Instead of electrical currents, the waveguide carries a bound electromagnetic wave with the electric components extending between the parallel walls. Creation of the guided waves in the first place is accomplished with the small monopole element inserted into the waveguide at the input end. Note how the electric components vary in intensity along the width of the horn, being maximum at the center (closely spaced arrows) and dropping to zero at the edges. That particular natural

Vertically polarized horn

Horizontally polarized horn
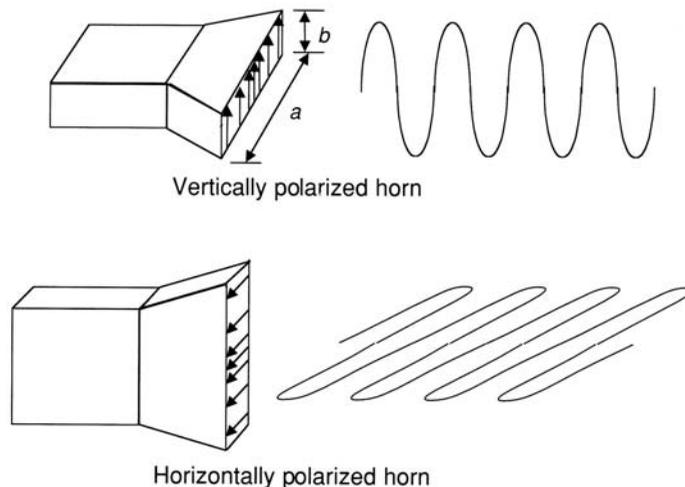
Figure 4.9 Linear polarization of microwave radiation from vertically and horizontally polarized feed horns.

property of a simple horn provides a radiation taper used in the design of horn-fed reflector antennas. In any case, the radiated electromagnetic wave is linearly polarized, just like that from a vertically oriented dipole. The polarization can be changed to horizontal simply by rotating the horn and waveguide by 90 degrees.

As with wire antennas, maximum coupling occurs when the transmitter and receiving horns are co-polarized. If one horn is rotated 90 degrees with respect to the other, minimum energy transfer (coupling) results and they are said to be cross-polarized. The performance of LP at intermediate points follows a simple law: The relative energy coupled is equal to the square of the cosine of the angle. That characteristic is plotted at the left of Figure 4.10 for the copolarized case, that is, for the level of signal received as the receiving antenna is rotated from maximum coupling to minimum coupling. An ideal LP wave and antenna are assumed, for which the coupling goes from a maximum of 1 to a minimum of zero. Maximum coupling always occurs at some angle, and minimum coupling is as close to zero as physically possible. Properly designed and installed antennas can deliver minimum coupling values of 0.0001, or −40 dB. Larger structures that are highly symmetrical in the direction of radiation can even exceed that value. However, due to slight misalignments and other imperfections, values in the range of 30 to 40 dB are more typical.

Probably the most important application of polarization is in frequency reuse, where two cross-polarized signals are transmitted at the same time on the same frequency. The right curve in Figure 4.10 shows how the level of the cross-polarized signal increases as the receiving antenna is rotated from 0 to 90 degrees. Notice how at 45 degrees both signals are at the same level, which results in no isolation between the signals on opposite polarizations. Figure 4.11 plots coupling in decibels, termed *polarization isolation*, between the desired and undesired polarizations. Maximum isolation occurs at zero offset angle, that is, where the receiving antenna is aligned in polarization with the transmitting antenna and the undesired polarization is nulled out (minimized). Alignment for maximum coupling is not particularly
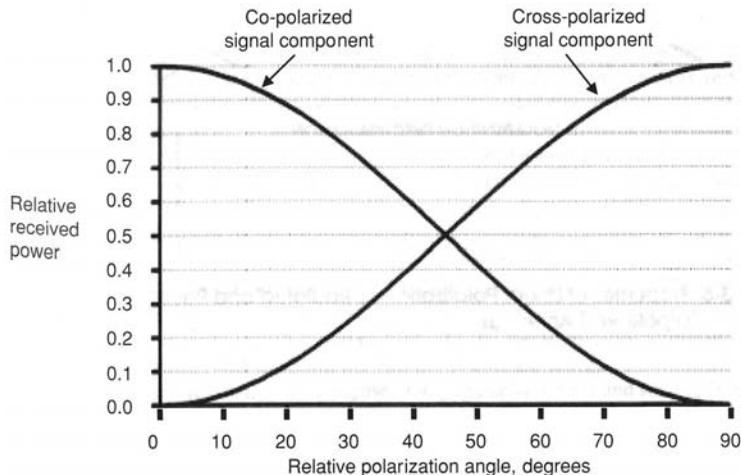
Figure 4.10 Relative receive power as a true ratio for the copolarized and cross-polarized signals in a linearly polarized link as the polarization angle of the receiving antenna is rotated.
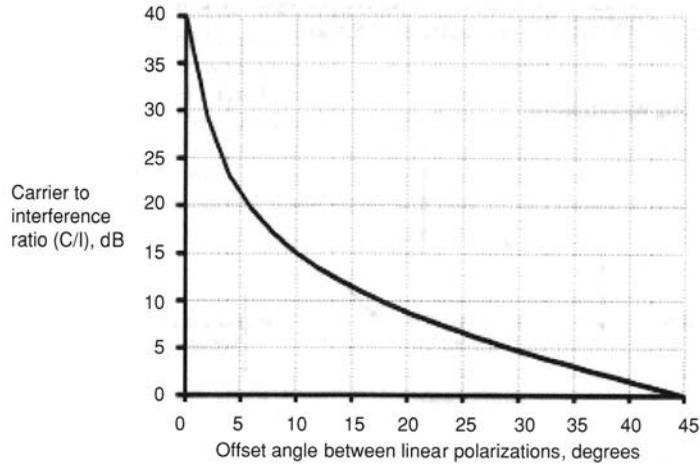
Figure 4.11   Polarization isolation in decibels for a linearly polarized RF signal as a function of the offset angle from maximum isolation.

critical (±5 degrees of error introduces very little loss of signal); however, the cross-polarization isolation is extremely critical, and precise alignment is a necessity. The maximum value of isolation of 40 dB in Figure 4.11 corresponds to the typical value described in the preceding paragraph. All FSS and BSS satellite communications systems use polarization isolation to increase the capacity of the particular orbit position. Therefore, users must employ antennas on the ground that meet polarization requirements; linear polarizations should be adjusted from time to time by rotating the feedhorn for minimum RFI. (The only way around the latter requirement is to use circular polarization, discussed in the next section.) Isolation can subsequently degrade from propagation effects, particularly Faraday rotation, which is a twist of linear polarization caused by the ionosphere. If the ionosphere caused a change of 6°, then the cross-polarization isolation would be approximately 20 dB. Such a low value generally is considered to be unacceptable.

4.2.4.2   Circular Polarization

Another type of polarization that is used in international, DTH broadcasting and MSS satellites is circular polarization (CP). Being circular in polarization, the receiving antenna is not aligned for the transmitting antenna's polarization. Rather, it is the inherent polarization of the signal, either right-hand CP (RCP) or left-hand CP (LCP), that must be supported by both ends of the link. An example of RCP is provided in Figure 4.12.

To understand the composition of CP, first observe that LP can be represented in the form of a vector, where the direction of the vector is in line with the electric component and its magnitude is proportional to the square root of the power of the signal. This vector is always perpendicular to the direction of propagation. CP is a particular combination of two equal LP waves that are cross-polarized with respect to each other (horizontal and vertical are two of the four vector states in Figure 4.12). If the two polarizations represent the same signal on the same frequency and they are precisely in phase with each other, the resultant vector lies at
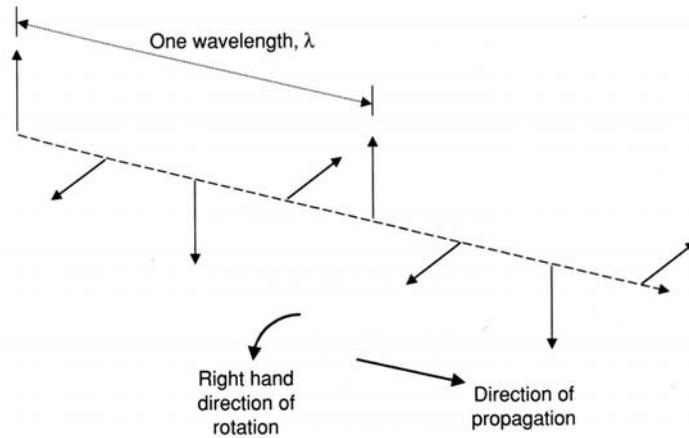
Figure 4.12 Rotation of the electric component during the propagation of a circularly polarized RF signal.

a 45-degree angle between the two polarizations with an amplitude that is $\sqrt{2}$ times either amplitude. In CP, however, the same two vectors are out of phase with each other by 90 degrees. That is accomplished by first splitting the transmit signal in two at the source antenna and then delaying one component by a quarter period before radiating the components through a dual-polarized antenna element. What is produced is a resultant vector that rotates like a corkscrew (Figure 4.12) as it propagates through space. The sense of polarization is either rotating to the right (RCP) or to the left (LCP), looking in the direction of propagation. That is determined by which linear component is delayed with respect to the other. Frequency reuse is possible because LCP and RCP waves are oppositely polarized with respect to each other, just like vertical and horizontal LP waves.

The advantage of CP is that the receiving antenna does not need to be aligned in polarization since the electric component rotates 360 degrees at a rate equal to the frequency of transmission. A receiving antenna could consist of combined vertical and horizontal LP elements, allowing the components to be coupled and recombined with the proper phase relationship. The sense of combining determines whether RCP or LCP will be recovered. Because most waveguides are linearly polarized, a device called a polarizer is used to make the conversion to CP. Also, the waveguide must be square or round to prevent cutoff at the operating microwave frequency. The most common approach is to generate two orthogonal linear polarizations, delay one of them by 90 degrees (one-quarter wavelength), and then add them. In general, the amount of polarization isolation thus obtained is not as high as that achieved by LP because of the imbalance between the two components that often results in practice. This imbalance produces an elliptical pattern measured in terms of the axial ratio. (CP has an axial ratio of 1, or 0 dB.)

4.2.5 Propagation Losses

The free-space path between transmitter and receiver may pass through a combination of ground obstructions and atmospheric phenomena. They can be divided into

those that are constant (and usually predictable) and those that are both time variant and random (which may be estimated statistically). Taken together, these losses reduce the strength of the received signal and can cause its level to vary over time even to the point where the signal fades away. This is illustrated in Figure 4.13, which shows received signal power (signal level) as a function of time during a period of rapid variation (fading). The mean signal level results from the constant and predictable losses on the link; the random losses cause the level to decrease and occasionally drop below the acceptable threshold. When the signal dips below that threshold, it could become unusable for communication purposes because of excessive noise and, in the case of digital links, an undesirable high error rate and ultimately a loss of synchronization as well. The difference in decibels between the mean level and the threshold is called the link margin, which is discussed at the end of this chapter.

Table 4.3 summarizes the primary losses that must be determined and accounted for in any microwave link design. The table compares typical downlink values for L-, C-, Ku-, and Ka-bands. Certain losses are critical in all satellite systems, while others are of concern only in specific cases.

4.2.5.1   Free-Space Loss

We already have covered in some detail the mechanism for free-space loss, which diminishes signal strength due to distance alone. For 1 m^2 of effective antenna area, the captured energy is simply

$$P_r = P_t / 4\pi R^2$$

where P_t = the radiated power in watts and R = the distance in meters.

The typical distance from a point on the Earth to a satellite in GEO is 36,000 km (22,300 miles).

This relationship, which should be recognized from the previous discussion of isotropic radiation, does not depend on frequency. It is a common practice to introduce frequency into the equation by splitting apart the propagation loss from the gain of the receiving antenna (here assumed to have an effective area of 1m^2).
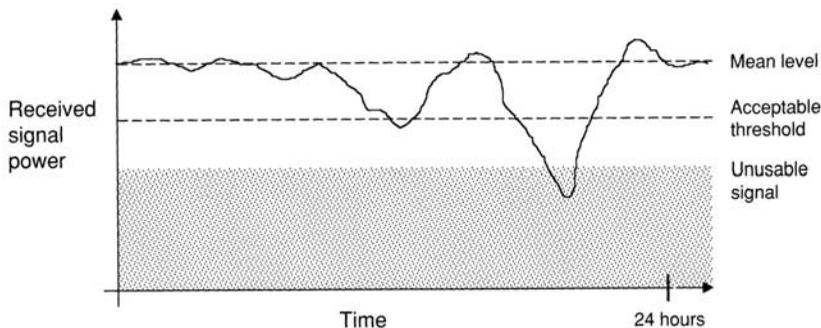
Figure 4.13   Variation of the received carrier power on a microwave link during a period of severe fluctuation (fading).

Table 4.3   Comparison of Significant Propagation and Fading Modes and Resulting Losses That Affect Microwave Links over GEO Paths (Values Are Rough Order of Magnitude, in Decibels)

Propagation Mode	Relative Importance on Satellite Links	L-Band (1.6/1.5 GHz)	C-Band (6/4 GHz)	Ku-Band (14/12 GHz)	Ka-Band (30/20 GHz)
Free space downlink	Dominating factor	187	196	205	210
Atmospheric	Relatively small and nearly constant at high elevation angles	0.1	0.2	0.3	0.5
Rain attenuation	Severe as frequency and elevation angle increases	0.1	0.5	2	6
Refraction	Significant at times at low elevation angles	6	3	2	1
Scattering	Produces local RFI				
Diffraction	Considered for mobile links without line of sight	6 to 12			
Ionospheric scintillation (multipath)	Occasional wide signal variation	3 to 6	1 to 3	<1	
Doppler	Frequency shift for moving vehicles or satellites				

The free-space path loss, which would be in the denominator of the preceding equation, is

$$a_0 = (4\pi R/\lambda)^2$$

where λ = the wavelength in meters and R = the path length in meters.

Substituting frequency for wavelength (e.g., $f = c/\lambda$) and converting to decibels, we obtain the following basic relationship for free space loss:

$$A_0 = 183.5 + 20 \log(f) + 20 \log(R/35{,}788)$$

where f is the frequency in gigahertz and R is the path length in kilometers. (The multiplier of 20 for the log accounts for the squaring of the terms within the brackets.)

This formula is plotted at three downlink frequencies as a function of path length in Figure 4.14. The figure illustrates the basic free-space loss that must be countered as a function of the particular orbital configuration. In particular, we see a fundamental difference of 30 dB between GEO and LEO satellite systems.

4.2.5.2   Absorption

RF energy is absorbed to some extent (and converted into heat) as it passes through clear air, water vapor, and smog. The impact of frequency of operation is shown in Figure 4.15 for an Earth station at sea level and transmitting or receiving at an
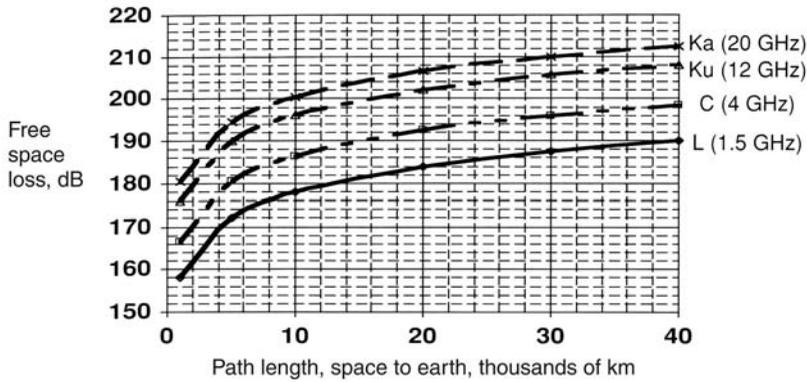
Figure 4.14 Free-space loss, A_0, as a function of orbit altitude and frequency band.
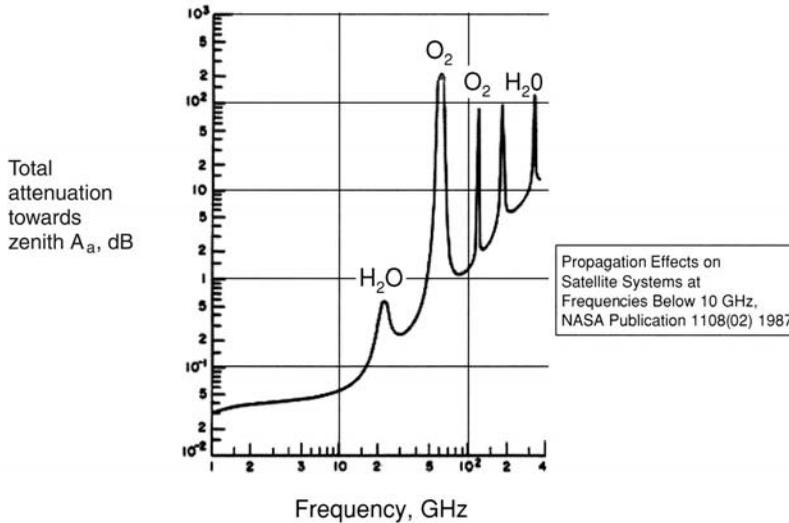
Figure 4.15 Typical relationship between atmospheric absorption and the microwave frequency of transmission, expressed in decibels.

elevation angle toward the zenith ($\Theta = 90°$). The absorption essentially is constant and will be under 1 dB at C- and Ku-bands. The only place on the ground where $\Theta = 90°$ is at the subsatellite point, on the equator directly below the satellite. At other locations the elevation angle takes values down to perhaps a practical low of 5°. At a particular frequency shown along the X-axis, the atmospheric absorption for elevation angles less than 90° would increase due to the greater thickness of atmosphere. Below 10°, absorption is highly sensitive to multipath due to ground reflections and atmospheric ducting. The following formula indicates how atmospheric absorption changes for elevation angles greater than 10°:

$$A_a \approx \frac{8a_O + 2a_W}{\sin \theta}$$

The two parameters, a_O and a_W, represent constituent absorptions due to oxygen and water vapor, respectively. We can view the numerator as being the total attenuation indicated on the Y-axis of Figure 4.15.

As frequency increases above 15 GHz, the constituents of the atmosphere reach individual points of resonance, and absorption can become very high, even total. The bands of fequencies around 22 and 66 GHz correspond to resonances for water vapor and oxygen, respectively, and are not employed for either uplinks or downlinks. Direct links between satellites, called intersatellite links (ISLs) or, alternatively, cross-links, bypass the atmosphere and hence may utilize the absorptive bands.

4.2.5.3 Rain Attenuation

After free-space loss, the most detrimental effect on commercial satellite links above C-band is rain attenuation, which results from absorption and scattering of microwave energy by rain drops. That loss, which increases with frequency, was discussed in Chapter 1 in the comparison of frequency bands. Rain attenuation is not predictable on an instantaneous basis, but statistical estimates can be made that allow links to be designed. Obviously, dry seasons and regions of the world with low rainfall would not suffer greatly from this phenomenon. However, links in regions with heavy thunderstorm activity—and hence rainfall—should be provided with greater link margin, or service might not be maintained with sufficient availability to satisfy commercial requirements.

Intense rain is contained in rain cells, which have somewhat limited geographic size. The statistical relationship between local rainfall and the resulting attenuation caused at a particular microwave frequency is a complex combination of several factors. The dimensions of the rain cell vary based on the rain rate, measured typically in millimeters per hour. At a particular rain rate and cell size, the attenuation increases with path length through the cell. Because a cell is shaped like an oblate spheroid (i.e., flat like a hamburger), the path length is inversely related to the elevation angle. Thus, rain attenuation varies in much the same manner as does atmospheric attention in the absence of rain. Other factors include gaseous water vapor, clouds, and ice particles. The work of combining all of these factors is largely done and available to us in the form of rain propagation models. The ITU has adopted the DAH model, which was developed by the research team of Dissanayake, Allnut, and Haidara. Figure 4.16 provides in summary form some rain attenuation data as a function of frequency and for a temperate climate such as that of Western Europe or the northeastern United States. Comparable data for a tropical region would be two to three times greater in decibels. The best approach is to use a software tool like SatMaster (discussed later in this chapter) that uses the DAH model to calculate rain attenuation for a specific location.

As indicated in the figure, the lower the elevation angle from the ground, the greater is the amount of attenuation from a given amount of rainfall. Heavy rainfall also alters the polarization of the signal because atmospheric drag causes raindrops to flatten and not be perfectly spherical. That reduces cross-polarization isolation between linear polarized transmissions. For example, the maximum isolation in Figure 4.11 decreases at Ku-band in heavy rain from 40 dB to approximately
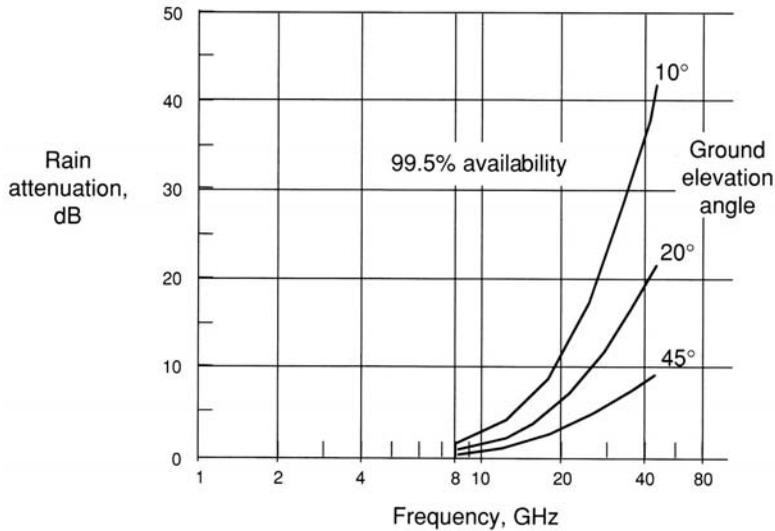
Figure 4.16   Rain attenuation versus frequency and elevation angle for a temperate climate.

25 dB. That would not be detrimental to most transmissions if appropriate adjustments are made in the link design. Depolarization in rain is not particularly harmful at C-band and the lower parts of Ku-band, but it can greatly reduce isolation at frequencies above 15 GHz. It also turns out that depolarization of CP transmissions is somewhat more severe than for LP.

The first step in using a rain model is to identify the particular rain region where the service must be provided. The map in Figure 4.17 shows different regions, each characterized by a type of climate and associated rain properties. Tropical regions are the worst, due to a higher incidence of thunderstorms, which contain the most intense rain cells. The shading and numbers designate the relative impor-
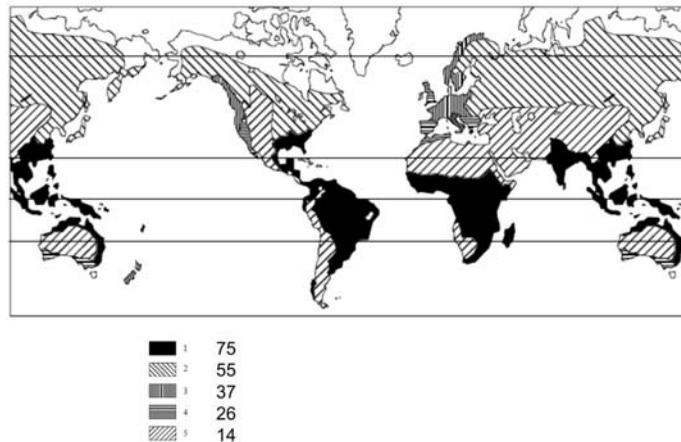
Figure 4.17   Regions corresponding to the five rain-climatic zones, indicating the rain rate in mm/hour to be exceeded for 0.01% of the time [2].

tance of rain in the service design. This presentation is not detailed enough to indicate higher rain rates in the most intense areas of Brazil, Indonesia, and central Africa, where they reach 145 mm/hr for 0.01% of the time. Note that the attenuation is approximately proportional to rain rate.

4.2.5.4   Refraction

The lower portion of the atmosphere, called the troposphere, decreases in density upward from the Earth's surface. As illustrated in Figure 4.18, electromagnetic waves are bent as they pass through the medium; as a result, the satellite has a virtual position slightly above that of its true position. That characteristic is taken into account in terrestrial microwave system design by increasing the Earth's radius by a factor of 4/3 in the profile map used to plot the line-of-sight path. In satellite links, the bending is less significant because of the higher angle toward the satellite.

Unstable atmospheric conditions like temperature inversions, clouds, and fog produce discontinuities and fluctuations in what otherwise would be a uniform air density distribution. The consequent random, temporary bending can cause significant signal fading akin to rain attenuation. The effect is more pronounced for paths that are nearly parallel to the Earth (line-of-sight microwave and satellite links at low elevation angles). That is also a worry for making precise gain measurements of an Earth station antenna when the reference transmitting source is on a ground-based tower. Generally, refraction by itself does not impair typical satellite links because the amount of bending is small relative to the beamwidth of the satellite and the Earth station antennas.

The ionosphere, the layer of ionized particles at around 150 km, completely reflects frequencies between 0.1 and 30 MHz under certain conditions, producing the "sky wave" effect, which allows short-wave transmissions to cover long dis-
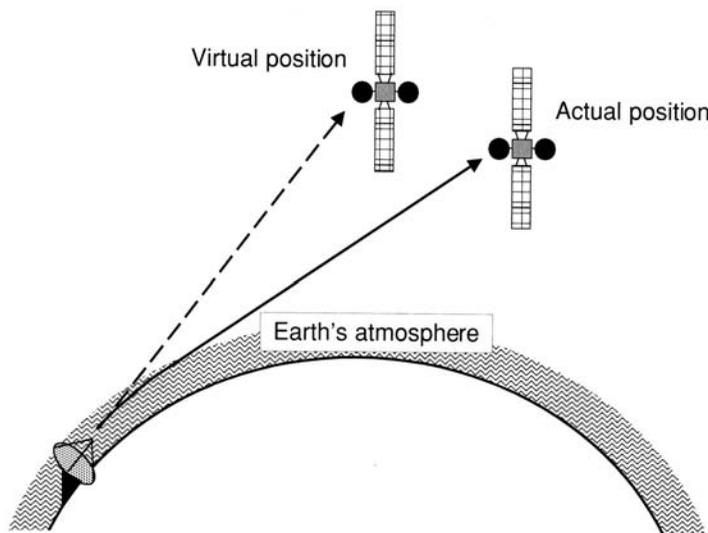
Figure 4.18   Refraction is the bending of a microwave signal as it passes through the layers of the atmosphere.

tances. In addition to refraction, the ionosphere rotates the linear polarization of microwave transmissions (Faraday effect), although that can be compensated for at frequencies above 3 GHz by appropriate adjustment of feed-horn angle at the receiving Earth station. This is complicated by the fact that the angle of rotation varies from day to night by as much as 6°. Below 3 GHz, Faraday effect produces one or more full rotations, which is why CP is preferred. Faraday rotation increases during periods of high sunspot activity, according to the 11- and 22-year solar cycles. Faraday effect decreases with frequency and can be ignored at Ku-band and above.

4.2.5.5 Scattering, Diffraction, and Shadowing

It was mentioned previously that water droplets scatter microwave signals. That would reduce the direct-path power level, spraying some of it back toward the source. At an Earth station site, the occasional scatter can place RFI in the direction of terrestrial microwave receivers that otherwise would have been adequately protected.

Diffraction, on the other hand, occurs when microwaves encounter and bend over a physical obstacle such as a building or mountain. The principle behind diffraction is illustrated in Figure 4.19, which shows how an Earth station might be shielded from local microwaves by an intervening hill or mountain. The heights, H_1 and H_2, introduce shielding but also provide the possibly diffraction path for microwave signals in both directions. If the hilltop just happens to touch the line-of-sight path between the Earth station and local microwave antennas (e.g., H_1 and H_2 are zero), a diffraction loss of precisely 6 dB is introduced. As the obstacle begins to block the path, the amount of diffraction loss increases until there essentially is total blockage.

The line-of-sight nature of microwave transmission means that blockage may reduce signal strength below the threshold level. For frequencies above 3 GHz, such blockage must be assumed whenever there is an obstacle that results in a situation like that illustrated in Figure 4.19. The situation at lower frequencies such as L- and S-bands might be somewhat more favorable because of the significantly larger wavelength involved. Depending on the type of ground antenna in use, it might be possible to receive adequate signal strength from reflections off of buildings and hills. It has been proposed that partially blocked and reflected signals
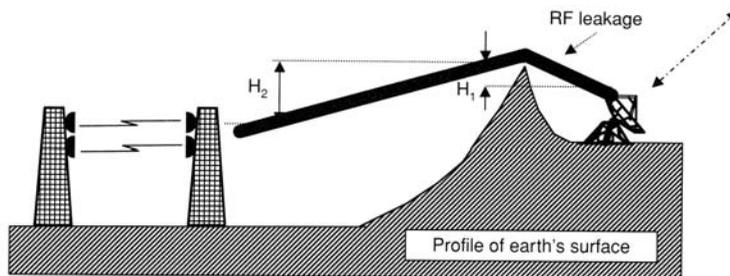
Figure 4.19 The use of terrain shielding to block RFI between a terrestrial microwave link and an Earth station.

be used to directly communicate or to augment communications for mobile users either with handheld satellite phones or in vehicles. This is certainly done in terrestrial cellular systems, but they have the advantage of much higher link margins from closely spaced base stations. Using a GEO or LEO satellite, it is not feasible to obtain such high margin and so one cannot expect reliable communications under shadowing conditions. The one area where some shadowing could be tolerated is where it is produced by trees that are relatively far apart. Even here, heavy tree trunks will block signals much the same as a building.

4.2.5.6   Multipath

Multipath is the result of the same RF signal taking both a direct path and a slightly longer refracted path, the latter arriving at the receiving antenna delayed in time from the former. It is illustrated in Figure 4.20 for refraction caused by discontinuities in the ionosphere. Multipath also occurs under shadowing conditions. It will be shown that both signal cancellation and enhancement can occur. The twinkling of a star is a multipath phenomenon, where two light rays, one direct and the other bent back in the troposphere, reach the eye and combine to produce a variation in light intensity. The twinkling occurs when the variability in the air causes the refracted path to change over time.

Due to the fact that rays are wave phenomena, the direct and refracted paths combine using vector addition, as illustrated in Figure 4.21. Shown at the right are the two signal paths reaching the same antenna, wherein they combine to form a resultant signal (that is because both paths contain the same signal, although the refracted path, being longer, introduces a delay and consequent phase shift). The two drawings at the left of Figure 4.21 represent the extremes: enhancement at the top and cancellation at the bottom. The direct path is shown as the vertical signal vector of constant length. Added to that on a vector basis is the refracted signal (the dotted vector), which can arrive at any random relative phase angle, as indicated by the circle about the tip of the direct path vector. The resultant received signal vector, shown as a heavy arrow, is only slightly affected by the weak refracted
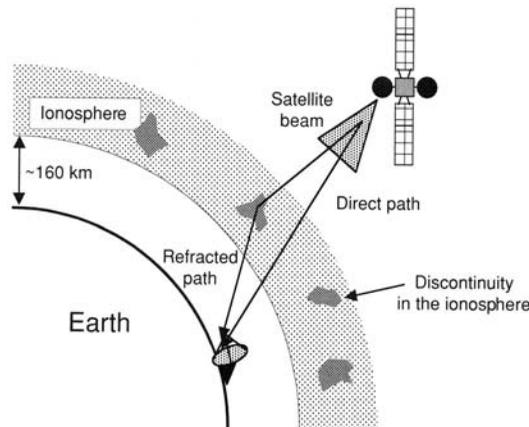
Figure 4.20   Multipath propagation caused by ionospheric scintillation.
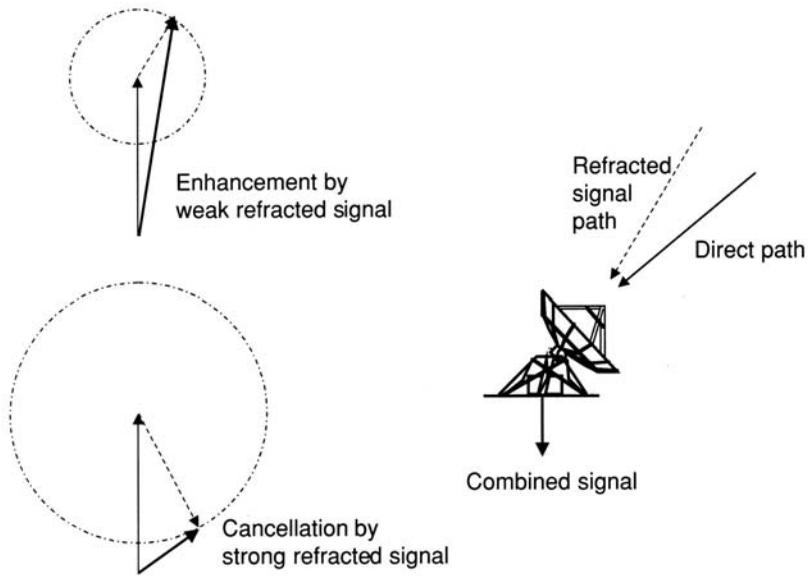
Figure 4.21   Multipath signals produce enhancement or cancellation of the direct path signal when received through a common antenna, as illustrated by vector diagrams.

signal. An increase or a decrease in strength occurs, depending on the relative phase angles.

If the refracted signal is comparable in strength to the direct path, cancellation occurs when the relative phase is 180 degrees. There also can be significant amplification (up to 6 dB) when the two vectors instantaneously combine in phase. Fortunately, the unstable conditions that produce this in the ionosphere only occur during two periods annually, around the equinoxes. Earth stations near the geomagnetic equator experience the most severe scintillation fades during those periods of activity. Ionospheric scintillation is not constant, with the result that the received signal both decreases and increases rapidly. The fact that the combining is frequency dependent produces selective fading, in which a spectrum of signals fluctuates widely. That is in contrast to shadowing due to terrain blockage, termed flat fading because it is not frequency selective.

Finally, multipath is an influence on L- and S-band mobile links because the user antenna is generally not very directional so can receive signals from several directions at the same time. Generally, the direct path dominates, but the reflected path or paths can produce pronounced fading. Terrestrial mobile radio services do not depend on the line-of-sight path and actually utilize multipath to provide coverage in areas of heavy terrain blockage. There, as much as 30 dB of margin is afforded by the close range between mobile user and lease station.

Satellite links, on the other hand, do not normally generate 30 dB of margin and so cannot depend heavily on multipath as a positive factor. Rather, signals reflected off buildings and local terrain features more likely will cause fading. The basic problem is illustrated in Figure 4.22 for a mobile receiver traveling along a highway. Two paths reach the mobile user: the direct line-of-sight path and a path that is reflected off the highway itself. They produce either fading or enhancement,
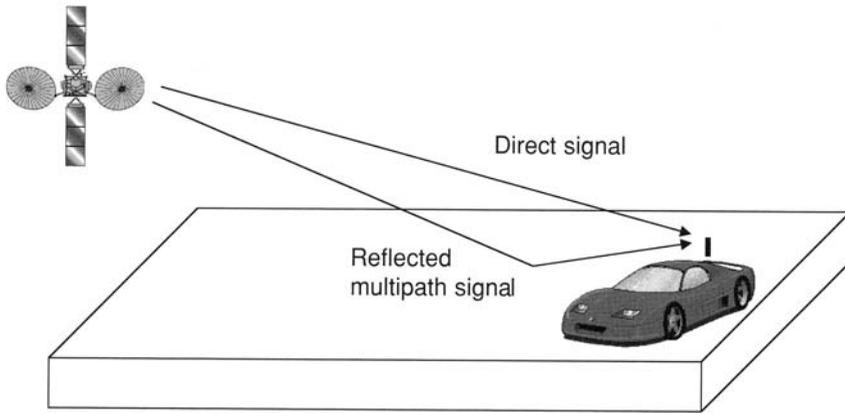
Figure 4.22   Generation of mobile multipath on flat, smooth terrain.

which is nearly constant as long as the satellite remains in the same position relative to the vehicle. The amount of loss depends on the reflection coefficient experienced by the reflected signal. Alternatively, reflection off a fixed feature, such as a hill or a building, produces time-varying fading as the vehicle moves relative to the reflection point. The rapidity of the fading depends on the speed of the vehicle. The design of MSS links is, therefore, the most challenging, because they are subject to greater variation and uncertainty.

4.3   Microwave Transmitters and Receivers

The basic elements in an end-to-end satellite communication link are illustrated in Figure 4.23. In the figure are the transmitting Earth station, which establishes the uplink path, a simplified satellite and its microwave repeater, the downlink path, and a receiving Earth station.

   The entry and exit points to the propagation medium are provided by the transmitting and receiving antennas. As previously discussed, antennas convert electrical energy at microwave frequencies into electromagnetic waves, and vice versa. A transmitting Earth station consists of equipment that impresses (modulates) the information to be sent on an RF signal called the carrier, translates it to the appropriate frequency and amplifies it to a high-enough power level to provide an adequate uplink. A receiving Earth station works in the exact opposite direction. Deferred to Chapters 6 and 9 are detailed descriptions of the electronics of the satellite and the Earth station, respectively.

4.3.1   Transmitting Station

Figure 4.24 portrays a single transmitting chain of a generic microwave station, with the signal input at the left and the RF output from the transmitting antenna at the right. The signal to be transmitted consists of information in electrical form, such as one or more voice channels for telephone service, digital data in the form
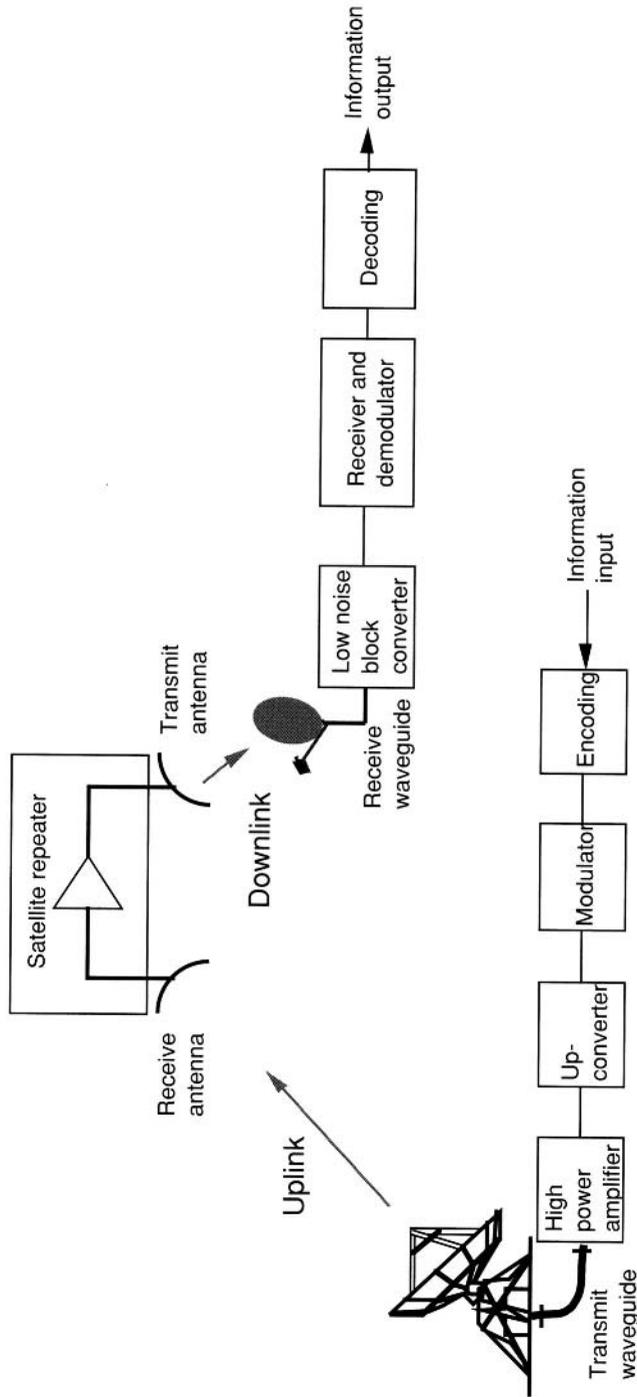
Figure 4.23   The main elements of a satellite link on an end-to-end basis.
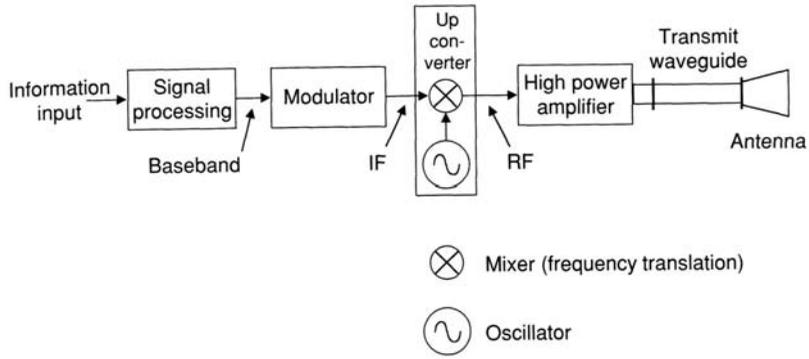
Figure 4.24 Simplified block diagram of a microwave transmitting station capable of baseband signal processing, modulation, frequency translation, and high power amplification.

of a high-speed bit stream, or a composite video signal such as that delivered from a video tape recorder. In modern satellite systems, analog information forms like voice and TV are first digitized and then compressed to reduce the required bandwidth. From that point, the link is digital in nature.

4.3.1.1 Encoding and Modulation

In the encoding element of the station (the first element to be encountered), the digital information is prepared for transmission through the link. Encoding consists of one or more numerical processes that better match the data to the specific characteristics of the satellite link. Table 4.4 summarizes the most common forms

Table 4.4 Typical Forms of Encoding Applied to Microwave Links over Communications Satellites

Encoding Technique	Application	Benefits	Impact on Data Throughput
Forward error correction	Error rate reduction	Lower error rate, or reduced power requirement (E_b/N_o)	Increases the output bit rate to accommodate redundant bits
Compression	Reduces the total bit count, either lossless or lossy	Less data to send or store—better utilization of the link and system	Increases throughput; may reduce quality and introduce delay
Encryption	Information security	Makes data private and difficult to corrupt	Generally, no increase in data rate; involves complex management and could introduce delay
Protocol adaptation	Data communications networks, typically using the Internet Protocol suite (TCP/IP)	Improves the user experience by countering satellite link impairments, such as bit errors and delay	May increase actual throughput; introduces complexity; tied to the type of information transfer or application

of encoding along with their benefits and characteristics. Forward error correction (FEC) is one type of encoding that is always applied as it can only improve the quality of transmission in terms of the bit error rate (BER). A reduction in BER can be taken either as a quality enhancement or to reduce the amount of transmitted power (alternatively, the antenna size). Modern FEC techniques use combinations of individual error correcting codes through the process of concatenation.

In many cases, FEC is incorporated into the modulation and demodulation function, which are performed by the modem. Compression techniques fall into two categories: lossless compression, which only removes useless bits that convey no information, and lossy compression, where information is removed but is intended to provide sufficient quality to satisfy the user and application. Information that starts out in analog form, such as voice, image, and video, can employ lossy compression as users are accustomed to some amount of impairment. However, the amount of compression deserves careful review by the intended audience to assure that the service is not compromised as to acceptability. Encryption is highly desirable on satellite links as it is a relatively simple matter to intercept satellite transmissions and potentially to introduce either bogus data or to otherwise disrupt information transfer. Protocol adaptation was discussed in Chapter 3 and is very effective for reducing the perceived delay of Internet services over GEO satellites.

The output of the encoding stage is referred to as the baseband. It is the purpose of the modulator to take the baseband and apply it to an RF carrier. The type of modulator used in a microwave station often handles a wideband baseband input such as that obtained from a high-capacity data stream measured in megabits per second. In SCPC service, the output of the modem is kept on to allow a continuous stream of data to be uplinked to the satellite. TDMA operation, on the other hand, requires that the modem transmit in noncontinuous bursts. That is because the RF channel is being shared by multiple Earth stations that transmit in different time slots.

The opposite of modulation is simply demodulation, which is the process whereby the baseband is removed from the carrier. The demodulator intially must acquire the incoming carrier, demodulate the bit stream, and then produce the baseband in a form that can be used by the decoder. The process is repeated for each received transmission, particularly for TDMA operation. The decoding circuitry corrects a majority of the errors produced by noise and interference on the uplink and downlink.

4.3.1.2   Frequency Conversion and RF Amplication

The RF carrier coming from the modulator typically is not at microwave frequencies but rather is centered within a standard frequency channel, the intermediate frequency (IF). Most transmitting and receiving stations use 70 MHz as the IF, allowing modulators and demodulators to be conveniently interchanged and interconnected by patch cords and coaxial switches. In low-cost consumer equipment, the IF is internal to the unit and hence need not follow any particular standard. Another point is that there are cases where the RF bandwidth is larger than 140 MHz, making it unfeasible to use 70 MHz as the IF (since the bandwidth would extend below zero frequency). This can be overcome by directly modulating a carrier at

the microwave frequency of transmission or, more likely, by using an adequately high IF, such as 140 MHz.

For digital modulation, the bandwidth is directly proportional to the input data rate:

$$B = a_m \cdot R$$

where R is the data rate in bits per second and a_m is a constant determined by the particular type of digital modulation and FEC. For example, QPSK is a popular form of digital modulation used on satellites, in which case a_m is nominally equal to 0.6. The proportionality constant, a_m, also depends on the order of the modulation. For example, going from QPSK to 16 QAM cuts a_m in half, to 0.3. Employing FEC with a code rate of 3/4 increases a_m to 0.8 and 0.4 for QPSK and 16 QAM, respectively.

The function of the upconverter is to translate the carrier without modification from IF to the desired microwave frequency of transmission. Within the upconverter are a microwave mixer and a local oscillator (LO). Translation is governed by a simple mathematical relationship that states that the output frequency equals the sum of the input IF and the frequency of the LO. For example, if the IF is at 70 MHz and the LO is at 6,030 MHz, the output RF is 6,100 MHz. (Although in theory the difference between the LO and IF could be employed, this is not done in practice due to more complex filtering.) From a practical standpoint, the RF usually is assigned by someone else and the IF is fixed. Therefore, the LO frequency must be selected properly to put the RF carrier in the right place (e.g., on the assigned frequency, which is 6,100 MHz in the example). Modern upconverters are frequency agile, which means that the LO can be tuned for a different RF channel from the front panel in much the same way as a TV set or car radio. The particular type of LO is called a frequency synthesizer because the frequencies are generated digitally using computational techniques. A nonsynthesized type of agile LO could cover the frequency range but would not be adequately stable over time and temperature. The desired frequency would be selected either from the front panel using an input keypad or remotely by computer over a data line.

The design of the IF aspect of Earth stations has evolved to a more flexible concept using a wideband IF nominally within the L-band within the broad range of 0.5 to 2.5 GHz. This is a continuous spectrum that is not directly radiated and so does not interfere with MSS services within this range. Using low-cost coaxial cable (coax), this L-band IF carries 500 MHz of bandwidth between the outdoor antenna system and indoor electronics used to select the specific transponder and frequency, and for carrier demodulation and baseband processing. The appropriate receiving device at the antenna is called a low-noise block converter (LNB), which contains a low-noise amplifier (discussed later in this chapter) and a block downconverter. On the transmit side, there would need to be an upconverter and high-power amplifier; if the transmit power required is less than about 10W, then it is possible to obtain both functions within what is called a block-upconverter (BUC).

4.3.1.3 High-Power Amplification

The last active element of the transmitting station is the high-power amplifier (HPA). Because all processing and frequency translation have been accomplished

in prior stages, the only function of the HPA is to increase the power of the microwave carrier from the low output of the upconverter to the power level needed to achieve satisfactory uplink operation. The HPA must have sufficient bandwidth to operate at the assigned microwave frequency and cover the active RF bandwidth of any anticipated carrier type. Examples of typical microwave HPA devices and their respective power capabilities are shown in Figure 4.25. Vacuum-tube HPAs, which do not require water cooling, offer the highest power. They include the klystron power amplifier (KPA) and the traveling wave tube amplifier (TWTA). Low to moderate power is affected by solid state power amplifiers (SSPAs), which are designed for a variety of power levels.

The KPA has a capability measured in kilowatts and is popular for video uplinks. Within the klystron microwave tube there is a resonant waveguide cavity that is tuned to the specfic frequency of operation. The operating bandwidth of a KPA is in the range of 50 to 100 MHz, making it necessary to retune the internal structure to change transponders. That difficulty is overcome with the TWTA, another class of microwave amplifier that happens to be very common on satellites as well (see Chapter 6). TWTAs have a higher-percentage bandwidth, which is the ratio of usable bandwidth to the center operating frequency, than KPAs and so can transmit over a total bandwidth of 500 MHz at C-band and as much as 2,000 MHz at Ku- and Ka-bands. Practical TWTAs can be found with power outputs of from 50W to as much as 800W, although 10-kW water-cooled TWTAs were used in early INTELSAT Earth stations. Being vacuum tubes, the TWTA and the KPA require sophisticated high-voltage power supplies, and both employ heated cathodes to emit electrons for use in the process of amplification. In the case of HPAs used in Earth stations, it is unavoidable that both types of high-power tubes will wear out and need to be replaced after a few years of operation. TWTAs used in satellites must have longer lifetimes, lower mass, and greater dc-to-RF efficiencies
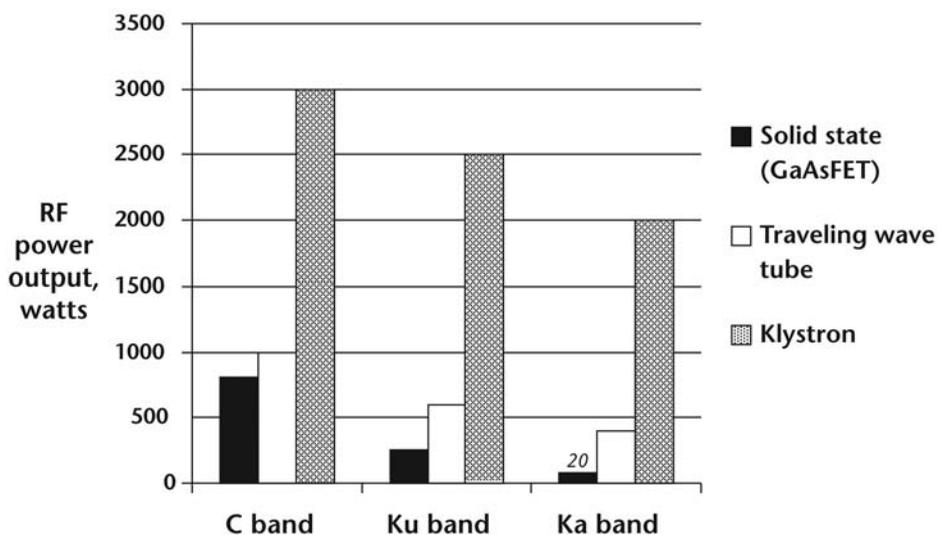
Figure 4.25 RF output capability at C-, Ku-, and Ka-bands of standard uplink microwave power amplifiers.

and so have to be designed differently. That also results in lower power outputs, generally under 300W at the time of this writing.

In low-power applications, such as the VSATs mentioned above, SSPAs can be employed for stable, long-term operation without maintenance. A detailed discussion of SSPAs can be found in Chapter 6. Figure 4.25 presents SSPA power levels available in 2008: up to 800W at C-band, 250W at Ku-band, and 20W at Ka-band. The basic building block of the SSPA is the gallium arsenide field effect transistor (GaAsFET), and the overall power capability is determined by the number of individual transistor stages that can be effectively paralleled. Bandwidth of SSPAs is somewhere between those of the KPA and the TWTA, based on the biasing and tuning of the transistor stages. That generally cannot be altered after the SSPA is manufactured.

4.3.2   Receiving Station

The reverse process found in the receiving station is illustrated in Figure 4.26. Because the microwave signal collected by the receiving antenna is weak, it first is necessary to raise the power to a level that can be accommodated by the processing elements. That is performed by the low-noise amplifier (LNA), whose gain must meet the requirements described above. However, the internal noise contribution of the LNA must be held small enough, or the weak signal input literally can be buried in noise. The rest of the elements perform functions that are inverse to those of the transmitting station. If an integrated downconverter is appended to the LNA, the resulting unit is called a low-noise block converter (LNB).

4.3.2.1   Low-Noise Amplifiers

It is possible to express noise in terms of an equivalent noise temperature in Kelvin (K). The Kelvin scale has gradations the same size as degrees C (centigrade) but begins at the noiseless state of absolute zero and measures the average random energy of motion of electrons within the receiver electronics. According to the theory, the random energy of the electrons is proportional to the noise power that overlays the desired signals within the passband of the amplifier. A device cooled to 0K (an impossibility) would not contribute noise to the system.
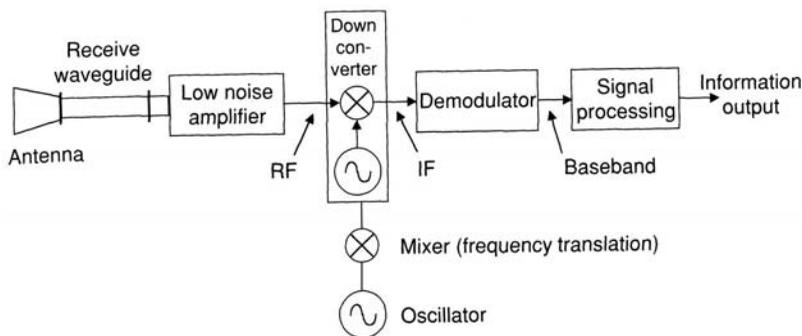
Figure 4.26   Simplified block diagram of a microwave receiving station capable of low-noise amplification, frequency conversion, demodulation, and baseband signal processing.

All LNAs and LNBs found in satellite communications Earth stations today use low-noise transistors such as gallium arsenide field effect transistors (GaAs-FETs). These are also referred to as high electron mobility transistors (HEMTs), which are more efficient and produce less noise than their silicon bipolar transistor counterparts (although silicon transistors are more rugged and static discharge resistant). Most transistors used in large-scale integrated circuits are FETs as are microwave amplifiers of various types. Indium phosphide is another material used in HEMTs and is gaining in popularity due to its improved performance over GaAsFETs. Typical values of LNA or LNB noise temperature at C-, Ku-, and Ka-bands are 30K, 60K, and 90K, respectively. Another important property of the LNA and LNB is its gain, which can be selected in the range of 50 to 70 dB. This involves one or two low-noise HEMT stages followed by one or more medium-power amplification stages that can be either HEMT or silicon. An LNB also incorporates the block downconverter stage that employs a wideband mixer and a local oscillator. The latter has a significant impact on the performance of the Earth station, particularly for reception of narrowband signals that cannot tolerate frequency instability or the oscillator impairment called phase noise.

It is common practice in Earth station design to place the LNA as close as possible to the feed and thereby minimize the input loss. That has two benefits: (1) it improves the effective gain of the receiving system because any loss on the input must be subtracted from that of the antenna, and (2) reducing this input loss also decreases the effective noise of the receiving system. The latter effect is discussed in later chapters.

4.3.2.2 IF-to-Baseband Chain

As discussed in Chapter 1, the satellite's downlink frequency usually is offset from its uplink frequency to prevent self-interference on the satellite (the exception is for single-band systems like Iridium that sequentially transmit and receive packets on the same frequency). To receive the carrier, the frequency of the LO of the downconverter would differ from that of the upconverter by the fixed frequency offset (2,225 MHz for standard C-band) introduced by the satellite repeater. To complete a previous example, the downconverter LO should be set to the (6,100 – 2,225) – the IF.

The demodulator is tuned to the IF to produce the baseband from the received carrier and sometimes provides FEC decoding. Along with the information are versions of noise and interference picked up over the link. Specifically, the received digital carrier on the downlink is demodulated to deliver a digital bit stream with the digitized version of the desired information. The demodulated performance is measured in terms of bit error rate. The bits are processed further to reduce the error rate, decompress the underlying information, and, if appropriate, convert the data back into the original analog form. Greater detail on digital modulation characteristics is contained in Chapter 5.

4.3.3 Definition of a Transponder

The satellite communication industry has long used the term *transponder* in reference to a defined RF channel of communication within the communication payload.

The term itself is a contraction of *transmitter-responder*, originally referring to a single-frequency repeating device found on aircraft. The purpose of the aircraft transponder is to add the identification of the aircraft and actively enhance the power to be reflected back to the radar transmitter. A satellite transponder is entirely different because it is more of a transparent microwave relay channel, also taking into account the need to translate the frequency from the uplink range to the downlink range.

We can better define a transponder by examining the two different payload configurations shown in Figure 4.27. The single-channel repeater (shown in A) does just what its name implies: provide a single channel of transmission within the satellite. As shown at the top of the figure, the entire uplink band is translated in the downconverter and applied to a single power amplifier. As a consequence, this amplifier must accommodate the entire uplink bandwidth, which would amount to 500 MHz at C- or Ku-band. That tends to limit the power-handling capability, particularly if multiple carriers are amplified by the common output stage. The first generation of LEO and GEO satellites like Telstar and Syncom, respectively, had repeaters of this type, capable of carrying one TV channel each.

The transponderized design (shown in B in Figure 4.27) breaks up the downlink range into individual frequency channels. Figure 4.28 suggests how eight transponders would be divided from the uplink bandwidth, indicating the usable transponder bandwidth and a guard band. Note that the uplink section still carries the entire bandwidth of 500 MHz and is shared by all transponders. After the downconverter, there is a bank of RF filters, called the input multiplexer (IMUX), that separates the transponder channels on a frequency basis. The output of each IMUX filter contains one transponder bandwidth, to be amplified in an individual power amplifier (PA), either a TWTA or SSPA, depending on the frequency and power required. On the output side, the RF power of the PAs is summed in a passive power combiner called the output multiplexer (OMUX); it is composed of specially designed low-
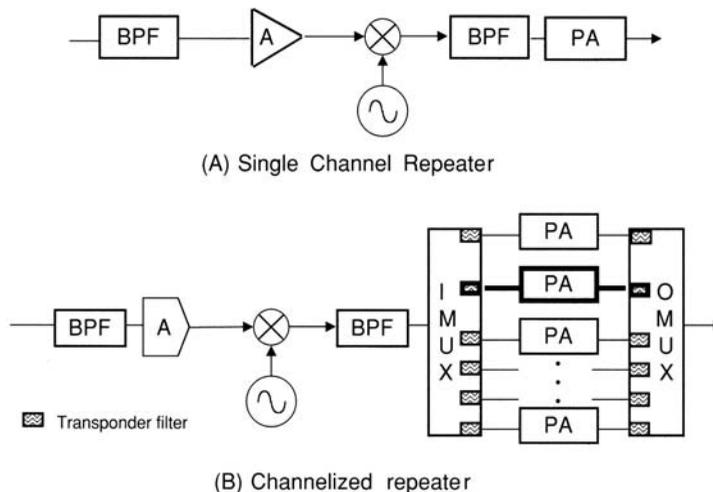
Figure 4.27   General arrangement of (a) a single-channel repeater and (b) a channelized repeater with multiple transponders.
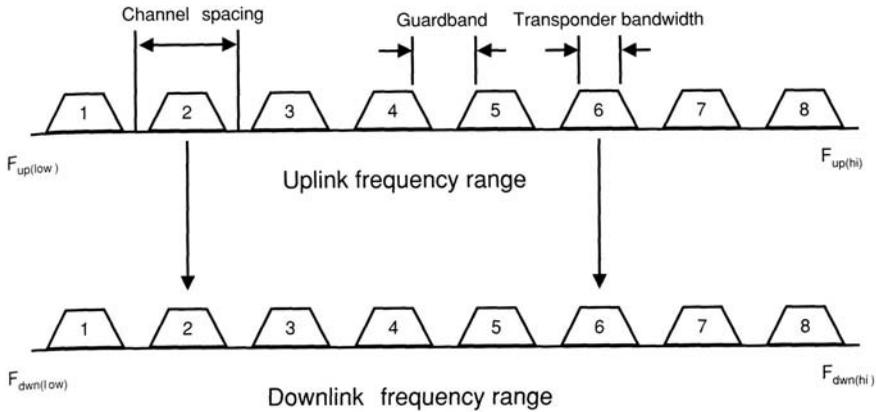
Figure 4.28 Hypothetical frequency plan illustrated for eight transponders in one polarization.

loss waveguide filters. A single waveguide transfers all the power from the OMUX to the antenna system.

The transponder, then, is a combination of elements within the payload. On the input side, it represents a share of the common uplink and receive equipment within the repeater. We are able to identify specific equipment for each transponder on the downlink side, consisting of the input filter, power amplifier, and output filter. Not shown in the figure are the necessary spare active elements (redundancy) to ensure continuity of service in the event of amplifier or receiver failure. Also not shown at the input to each PA is a level controlling channel amplifier or linearizer often found on modern satellites.

The fact that a transponder can be assigned to a particular user application network has caused them to be rented or sold like condominium flats. In actuality, it is the microwave channel of communication bandwidth that the transponder lessor or purchaser acquires.

4.4 Overall Link Quality

One of the more complex problems in microwave link engineering is knowing precisely how a particular signal will be affected by the noise, which is random in nature, and by interference from other radio carriers. This section reviews some of the factors that determine the actual quality and threshold for digital receivers. A clear understanding of the threshold of the link allows the engineer to predict how much margin will be available to overcome fading from the various sources discussed at the beginning of this chapter. It is presumed that the circuitry involved is properly impedanced matched, so that signal reflections do not unduly distort the signals.

4.4.1 How Noise and Interference Affect a Microwave Link

Both noise and interference will degrade service quality and if not controlled will at times render the link unusable. Interference, as explained previously, is often

due to the RF transmission of someone else. On the other hand, noise is totally random in nature because it results from the random motion of electrons or other elemental particles in the environment or the receiving equipment. Pure white noise is most common on satellite links and produces random voltage fluctuations whose probability follows a normal (Gaussian) distribution. The term *white* refers to the fact that the frequency spectrum does not have discrete components at specific frequencies but rather is a continuum of frequencies, like white light. Ideal white noise, running from zero to infinite frequency, is physically impossible since it would represent infinite power. Noise is sufficiently white if it is constant over the bandwidth of the signal in question.

The resulting density of the noise power (noise power density) in watts per hertz is proportional to the equivalent noise temperature in Kelvins, that is,

$$N_0 = k \cdot T$$

where k is Boltzmann's constant (i.e., $1.3803 \cdot 10^{-23}$ W-sec/K or -228.6 dBW-sec/K when expressed in decibels).

This simple relationship demonstrates why the noise performance of an LNA of a receiving Earth station is rated in terms of an equivalent noise temperature. The actual noise power that affects a given signal is that which lies with the signal's bandwidth, B, that is,

$$N = k \cdot T \cdot B$$

The effect of white noise on a modulated carrier containing binary data can be shown using time waveforms, presented in Figure 4.29. The noise-free case is shown in Figure 4.29(a) in the form of digital information impressed on a carrier. The transition from binary 1 to 0 occurs where the phase of the sinewave reverses. That type of modulation is called phase shift keying (PSK) because the shifting (flipping) of phase by 180 degrees is the means by which information is transmitted. It is the job of the demodulator to detect the phase reversal to convey to the user
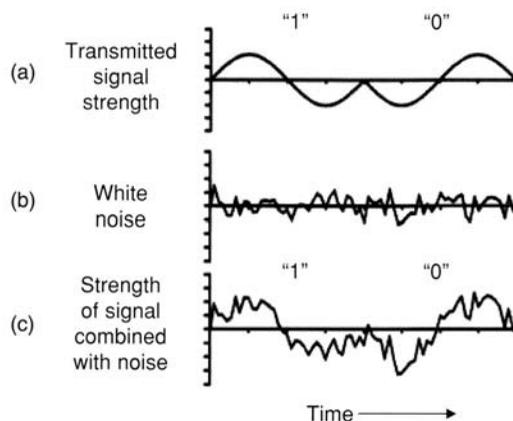
Figure 4.29   (a–c) A digital phase-shift keyed signal combines with white noise to produce a distorted signal waveform that must be detected by the receiver.

the proper bit sense. The PSK signal is sent through the link and enters the receiver along with white noise inherent in the electronics. That noise is depicted in Figure 4.29(b), where the mean voltage is zero and the standard deviation of the noise is approximately one-half the carrier amplitude. Because power is proportional to the square of the voltage, the true ratio of carrier power to noise power is 4, or equivalently 6 dB. The rather high relative noise (low C/N) level yields the sum of carrier plus noise shown in Figure 4.29(c), where the carrier appears to be somewhat obliterated by noise spikes.

A well-designed demodulator can detect the transition from the 1 to 0 state in most cases. The noise can occasionally reach a voltage greater than the signal amplitude; if it has the opposite sense to the transmitted signal at the precise moment of testing, it can cause the demodulator to make the wrong choice. The assumed binary digit will then be incorrect (i.e., a 1 instead of a 0 or vice versa, depending on what was sent). It should now be possible to visualize that the strength of the signal relative to the noise will determine the rate at which errors in detection are made. This is a significant area of engineering study in the communications field. Some of the most important applications of digital signal processing (DSP) allow low-cost sending and receiving equipment to identify errors and reverse them prior to delivery to the user. Special codes that provide such forward error correction (FEC) accomplish this in practice.

Wideband interfering signals can be treated as white noise and often are. However, if the interference is narrower in bandwidth than the desired signal, more sophisticated analysis techniques or even direct measurement must be employed. As that type of investigation is beyond the scope of this book, we will assume that the interference can be converted into an equivalent amount of white noise and added on a power basis to the other noise within the bandwidth of the desired signal.

4.4.2   Carrier-to-Noise Ratio

The strength of the carrier relative to the noise and not the absolute level of what comes out is what determines the quality of transmission. This can be understood by realizing that the power level of the baseband signal from the receiving station's signal processor can be raised by simple power amplification. However, the noise that is present also will be amplified. Only by suppressing the noise at the input to the receiver can the quality of transmission be improved. That is why the true performance of the link is measured by the ratio of RF carrier power to noise power. A variant used in digital transmission is the ratio of the energy per bit to the noise power density, E_b/N_0. It will be shown that there is a simple constant adjustment factor between C/N and E_b/N_0, that is,

$$\frac{C}{N} = \frac{E_b}{N_0} \cdot \frac{R_b}{B}$$

where R_b is the information bit rate, and B is the carrier bandwidth.

A receiving system showing a carrier, white noise, and a single source of interference is presented in Figure 4.30. The frequency spectrum is similar to the
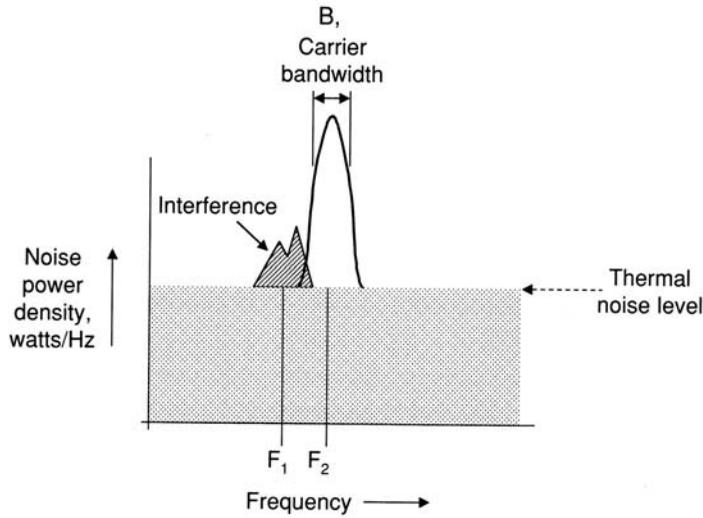
Figure 4.30   Frequency spectrum showing the desired carrier and the interfering carrier and white noise, as seen on a spectrum analyzer.

display on a microwave spectrum analyzer, a useful piece of test equipment. Spectrum analysis is vital to microwave communications because it provides one of the few windows for viewing the reception process. The figure shows a constant spectrum of white noise, providing what is referred to as a noise floor. The vertical scale is linear, measuring power density in terms of watts per hertz. Piercing through the floor are the desired carrier occupying bandwidth B and centered at frequency F_2, along with an interfering carrier at frequency F_1. The total noise power, N, over the bandwidth of the carrier is the product $N_0 B$. An actual spectrum analyzer would have a vertical scale measured in decibels to simplify the measurement of C/N. In that case, the C/N in decibels is equal to the difference between the measured decibel levels of total carrier power and noise power within the bandwidth B. While bandwidth can be taken into account by subtracting the quantity $10 \log B$ from the spectrum analyzer reading, it is also possible to present the spectrum analyzer's measurement bandwidth to eliminate the need for manual correction.

Another correction to the measurement of C/N on a spectrum analyzer results from the fact that this instrument actually displays the sum of carrier power and noise power, since the noise is present within the carrier bandwidth. This correction has the following form:

$$\frac{(C + N)}{N} = \frac{C}{N} + \frac{N}{N}$$

$$= \frac{C}{N} + 1$$

$$\therefore \frac{C}{N} = \frac{(C + N)}{N} - 1$$

This formula is performed on true ratios of powers, and not in decibels. To perform this calculation, take the reading in decibels, convert to a ratio, subtract 1, and then convert back to decibels. For example, if $(C + N)/N$ equals 8 dB (a factor of 6.3), then the actual C/N equals 5.3 as a ratio or 7.2 dB.

The type of interference shown in Figure 4.30 is not noiselike and therefore cannot be analyzed in a straightforward manner. One consideration is that only the upper corner of the interference actually affects the carrier. The ratio of carrier power to interference power can be taken from the spectrum analyzer display, but its effect on the reception process also depends on the frequency offset $(F_2 - F_1)$ between the two carriers and their relative bandwidths. If due to a malfunction, the interference may rove through the bandwidth for the carrier, affecting it as well as other carriers on adjacent frequencies.

4.4.3 Link Budget Analysis

The key station elements and propagation phenomena having been described, it is now possible to review the analysis and prediction of link performance. Figure 4.31 shows a simplified microwave link with a key parameter (in dB) indicated for each element. The transmitter can be characterized by the HPA power output, with the transmitting waveguide introducing some loss as it carries the power to the transmitting antenna. The electromagnetic wave propagates outward from the antenna into the medium, where it is subjected to various losses, the free-space loss being dominant. The small amount of signal power gathered by the receive antenna is carried through waveguide to the LNA of the receiver. It is possible to characterize the receiver by the minimum acceptable threshold, which takes into account the RF noise as well as the quality desired by the user.

4.4.3.1 Power Balance Equation

The simple mathematical relationship shown at the bottom of Figure 4.31 is the power balance equation. The parameters of the link are actually factors; however, it is a common practice to express all of them in decibels because that reduces the
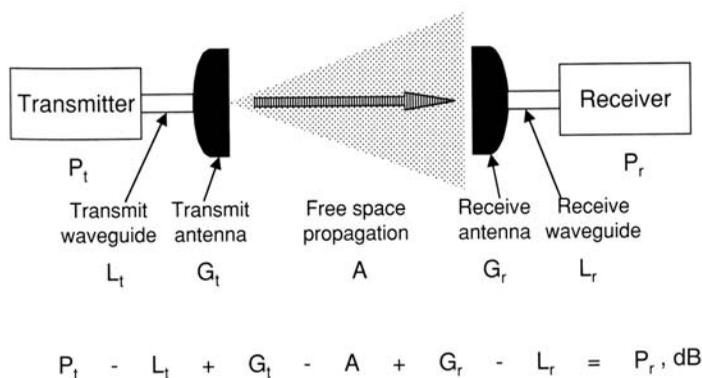
Figure 4.31   The relationship between the critical elements of a microwave link and the power balance equation.

analysis to addition and subtraction. The power balance in decibels can be stated simply: The power received equals the power transmitted plus all gains and minus all losses. Transmitter power is expressed in decibels relative to 1W (dBW). For example, 1W is by definition 0 dBW; 2W is 3 dBW; 10W is 10 dBW; 100W is 20 dBW; and so on. To preserve proper dimensions, there can be only one dBW term on each side of the equation (the rest are dimensionless dB terms). The power balance equation usually is arranged in a table called a link budget, in which each paramter of the link is provided with its own explanation and quantity. One approach to the problem is to set up for each case to be analyzed, using a spreadsheet program like Microsoft Excel. There are specialized link budget programs on the market such as SatMaster that simplify and standardize the process [3].

4.4.3.2   Typical Link Budget (Ku-Band)

A typical link budget for a Ku-band satellite downlink is presented in Table 4.4. The first entry is the RF power output of the transmitter, expressed first in watts and then converted into dBW. Transmit waveguide loss of 1 dB (26% power reduction) is subtracted, while the gain of the spacecraft antenna in the direction of the receiving Earth station is added. A detailed discussion of spacecraft antenna design and performance is given in Chapter 7. It is customary to show a subtotal at this point called the effective isotropic radiated power (EIRP), indicating how the satellite is performing compared to an isotropic source with 1W of RF drive power (i.e., 0 dBW). The value of 53 dBW shown in the table is typical for current high-power Ku-band satellites in the BSS. For a large country such as the United States, China, Indonesia, or Australia, this can be delivered by a 250W TWT.

With the satellite in GEO at an altitude of 36,000 km, the link budget contains a single entry of –205.6 dB of free-space loss. In contrast to spreading loss (e.g., $4\pi R^2$), path loss includes λ as a component, having the arithmetic form:

$$A = (4\pi R^2)(4\pi/\lambda^2)$$
$$= (4\pi R)^2/\lambda^2$$

converting to dB, $A = 92.5 + 20 \log F + 20 \log R$ where F is the frequency in GHz and R is the range in km.

Other propagation losses, such as rain attenuation, that are random in nature are evaluated separately and compared against the overall link margin. Since this is a clear sky link budget, we include an allocation of 0.1 dB for atmospheric loss. The next two items relate to the receiving Earth station: the peak gain of an assumed 45-cm antenna and 0.5 dB of waveguide loss. The combined power balance yields a received power of –120.5 dBW at the input to the Earth station LNA and receive electronics. For that hypothetical link, assume we have a receive system noise temperature of 140K and a signal bandwidth of 27 MHz. For the corresponding received noise power of –132.8 dBW (e.g., kTB), the downlink C/N result is 12.3 dB.

Uplink noise within the spacecraft receiver contributes a lesser amount to the total link noise and is ignored in the present example. However, the standard

approach is to prepare a separate link budget for the uplink, following approximately the same format as Table 4.5. Then the overall link C/N is determined using the following formula:

$$C/N_t = (N_d/C + N_u/C)^{-1}$$

calculated as ratios and not decibels. The final step is to convert back to decibels. Above, the subscripts t, d and u indicate total, down- and up-link components, respectively.

The next three lines allow us to translate the expected clear-sky C/N into the corresponding value of E_b/N_o. The process for doing this is to first take the ratio of the carrier bit rate, here assumed to be 30 Mbps, and the carrier bandwidth. That turns out to be 1.111, which corresponds to 0.5 dB of adjustment. This is subtracted from C/N to obtain the E_b/N_o for the downlink under clear sky conditions.

4.4.4   Link Margin

The final and perhaps key concept in microwave link engineering is that of link margin, which is nothing more than the excess power in the carrier relative to the threshold value. It is possible to measure the threshold in the laboratory under controlled conditions. First, the link performance without fading can be determined using the power balance equation and spectrum analysis. The margin, then, is the difference in decibels between the minimum value and the threshold value. For example, if the minimum acceptable (threshold) E_b/N_0 is 8 dB, the link margin is $11.8 - 8 = 3.8$ dB. After the link is operational, the margin can be verified by manually decreasing the transmit power to the point where the received baseband

Table 4.5 Typical Ku-Band Link Budget for the Downlink to a Small DTH Receiver

Parameter	Relative Value (Expressed as a Gain)	Absolute Value	Units
Transmit power (250W)		24.0	dBW
Transmit waveguide loss	−1.0		dB
Transmit antenna gain	30.0		dBi
EIRP		53.0	dBW
Free-space loss	−205.6		dB
Atmospheric loss	−0.1		dB
Receive antenna gain (45-cm diameter)		32.7	dBi
Receive waveguide loss	−0.5		dB
Received power (clear sky)		−120.5	dBW
Bandwidth (27 MHz)		74.3	dB(Hz)
System noise temperature (140K)		21.5	dB(K)
Boltzmann's constant		−228.6	dB(Hz*K)$^{-1}$
Received noise power		−132.8	dBW
C/N_d	12.3		dB
Bit rate (30.0 Mbps)			
Ratio of bit rate to bandwidth (1.111)	0.5		dB
E_b/N_o (downlink)	11.8		dB

signal is barely acceptable. The recorded change in transmit power in decibels is essentially the link margin.

Adequate link margin allows the link to deliver exceptionally good quality (low bit error rate) under mean signal, clear weather conditions. However, when either the uplink or the downlink is experiencing heavy rain or mulipath fading, the available link margin determines how often and for how long the link will drop below threshold. The criterion often used is the availability, that is, the percentage of time the link is above threshold. Typical satellite links operate in the range of 99% to 99.95%, demonstrating the high reliability of line-of-sight paths between the satellite and its associated Earth stations.

References

[1]   Final Acts of WRC-77, 2. Radio Propagation Factors, International Telecommunication Union, Geneva, 1997.

[2]   ITU Radio Regulations, Vol. 2, Appendix S7, Edition of 2008.

[3]   Sat Master software, available from Arrowe Technical Services (http://www.satmaster.com/).

Modulation, Multiple Access, and Impairments

Digital communication formats and transmission systems have long been used over commercial satellites. To understand how a satellite link actually transfers information between Earth stations, we must delve into the twin topics of modulation and multiple access. We will quickly see that these relate directly to the analog or digital format being applied. Also important are the impairments to digital transmission produced by nonlinearity and nonideal frequency response of the end-to-end link.

The discussion that follows is reasonably technical and detailed, but the topics of modulation and multiple access are extremely important to the cost-effective use of communications satellites. In most networks, the efficiency of transmission is increased by combining several channels of information with a process called *multiplexing*, creating a *baseband* spectrum in electrical form. *Modulation* is the process by which the baseband is then impressed on a carrier. The arrangement and the sequence of these functions for analog and digital formats are shown in Figure 5.1. In Chapter 4, we touched on the role of modulation in transmitting user information with adequate circuit quality on a microwave link. The type of modulation as well as the specific parameters employed also determine how much traffic the satellite repeater can carry in aggregate. Also, the types of connectivity and flexibility achieved in serving multiple earth stations depend to some degree on the modulation system. Those aspects of RF transmission and link connectivity are referred to as *multiple access*.

5.1 Digital Baseband Signals and Hierarchies

Information in digital form can be represented as a stream of bits, that is, 1s and 0s. A digital baseband, therefore, is a digital stream that is to be transferred by a communication link or network. The bandwidth of the baseband is defined from approximately 0 Hz on the low end and up to a highest frequency somewhat greater than the bit rate. For example, if the bit rate is 10 Mbps, then the baseband ranges from 0 to approximately 13 MHz. Therefore, the bandwidth to convey the data is directly proportional to the bit rate of the information itself. The actual highest baseband frequency is determined by the filtering of the spectrum to control bandwidth and signal distortion, as discussed later in this chapter. These guidelines allow for the efficient transfer of the data over wires or coaxial cable. Generally
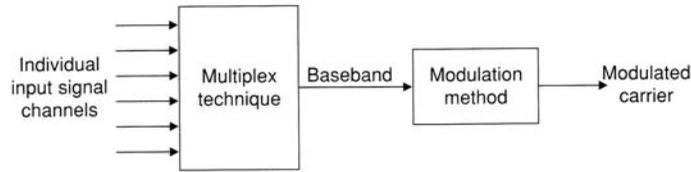
Figure 5.1 Multiplex and modulation methods for use with analog and digital information.

speaking, it does not matter if the link is terrestrial or satellite, provided that the baseband is delivered with an acceptable error rate and delay. If the link is interactive or two way, the error rate quality and delay (including its variability) can be significant factors in determining if the service is acceptable. Those are the most important technical issues in the design of a digital satellite communications network, particularly when we compare it with a comparable terrestrial service.

Digital capacity is defined according to one of the standard digital hierarchies. Table 5.1 presents two digital hierarchies in use around the world: the North American digital hierarchy and the European digital hierarchy, also called the CEPT hierarchy (CEPT is the abbreviation for the French term for European Commission of Post and Telecommunication). Both hierarchies use the basic building block of the 64-kbps channel. Each is referred to as a plesiochronous digital hierarchy (PDH), meaning that they are nearly synchronous. That can be seen from the fact the digital rates above 64 kbps are not integer multiples but include some additional bit rate capacity inserted at each level. For example, the T1 rate of 1.544 Mbps is higher than the precise value of 1.532 Mbps, which is exactly 24 times 64 kbps. The extra bits are inserted to provide synchronization information and to allow for maladjustments in clock rates, which were endemic in earlier networks.

Table 5.1 The Plesiochronous Digital Hierarchy (PDH) as Employed in North America and Japan (1.5 Mbps) and Europe, South America, Asia, and Africa (2 Mbps) (Entries Expressed in kbps)

Level	North America (1.5 Mbps)	Japan (1.5 Mbps)	Europe, South America, Asia, and Africa (2 Mbps)
1	1,544	1,544	2,028
2	6,312	6,312	8,448
3	44,736	32,064	34,368
4	97,728	139,264	

Because of two developments, such poor synchronization is largely a thing of the past. The first development is the production of low-cost atomic clocks employing a rubidium standard that are used to synchronize the local and national networks within a region and globally as well. Complementing that is the second development: the use of fiber optic communications with its inherent stability and quality of transmission. Highly synchronized digital transmission systems reduce wasted bandwidth and provide the means to increase the bit rates that users can employ. Today, the synchronous digital hierarchy (SDH) is the digital hierarchy of choice, both in public networks and in private installations as well.

In North America, where the standard was first introduced, we have the synchronous optical network (SONET, pronounced "sonnet"); in Europe the standard is simply called SDH. A comparison of the standards is presented in Table 5.2. The principal differences between SONET and SDH relate to terminology, not technology. They use different naming conventions and start at different basic line rates (which are multiplied up). The basic rate for SONET is 51.84 Mbps; we see that SDH's basic rate of 155.52 is exactly three times 51.84 Mbps. As it turns out, all practical SONET equipment begins at the OC-3 level, which, at 155.52 Mbps, happens to be precisely the same as STM-1, the basic line rate of SDH. In contrast, PDH equipment interfaces with user devices such as multiplexers, video teleconferencing equipment, and corporate access to the Internet. SONET is the principal digital transport medium of the Internet backbone in the United States. Satellite systems may have to accommodate both PDH and SDH forms of digital baseband transmission to properly interface with the multitude of application systems and networks.

5.1.1 Digital Information Sources and Bandwidth Requirements

The range of digital information sources that provide input to satellite communication networks has grown rapidly over recent years. It begins with text and simple graphic images and extends through standard definition and high definition television, and streaming broadband and interactive multimedia. In this section, we merely wish to identify the more familiar and promising of what can only be described as an opening cornucopia.

The basic digital information inputs are classified according to the following criteria:

Table 5.2 SONET (North America) and SDH (Worldwide) (Entries Are Expressed in Mbps)

SONET Level	SONET Designation	Bit Rate (Mbps)	SDH Level	SDH Designation
STS-1	OC-1	51.840		
STS-3	OC-3	155.52	1	STM-1
STS-9	OC-9	466.560		
STS-12	OC-12	622.080	4	STM-4
STS-18	OC-18	933.120		
STS-24	OC-24	1,244.160		
STS-36	OC-36	1,866.240		
STS-48	OC-48	2,488.320	16	STM-16

- Bandwidth or data rate requirement at the user interface, usually a peak or maximum value in kilobits, megabits, or even gigabits per second;
- Variability of this data rate, that is, whether it is constant, time varying, or bursty;
- Acceptable end-to-end delay (latency), which is more important for interactive services and less important for one-way information distribution on a broadcast, multicast or narrowcast basis;
- Communication connection paradigm, namely, dedicated bandwidth, circuit-switched or packet-switched.

Table 5.3 presents traditional telecommunication networks and services, evaluated against the parameters listed above. Those four criteria are probably the most important, but there are many more that can be critical to the success of a particular network or service. The secondary factors, such as call setup time, ease of use, availability, and cost per bit or minute of use, are important differentiators in the competitive commercial marketplace.

The entries in Table 5.3 are suggestive of the options available over traditional networks that are predominantly terrestrial in nature. Satellite communications must, of necessity, reflect and accommodate these as they represent the mainstream. There is a heavy dependence on fiber-optic systems and their inherent large bandwidth capacity and high quality of transmission. Accomplishing this over satellite links is not often practical, as discussed in Chapter 4. However, more and more data communications are finding their way to wireless networks and these are subject to many of the transmission issues that we associate with satellites.

In the past, terrestrial networks have employed standard transmission schemes such as time division multiplexing (TDM) and Asynchronous Transfer Mode (ATM). These share bandwidth among several different communication channels, but accomplish this in fundamentally different ways. In TDM, the data rate for each channel is fixed and dedicated to it; a variation of TDM called statistical multiplexing allows some variability among the channels as long as the maximum total is not exceeded (see Figure 3.5). ATM is a packet-switching scheme with a fixed-length packet called a cell. Each 53-byte cell transfers only 48 bytes of data, and so it takes many cells to service a particular demand. The switching "fabric" associated with ATM is based on very fast hardware technology and fiber optic lines, and so the latency associated with this packet approach is quite low.

TDM and ATM are prevalent in a variety of applications, but may be replaced ultimately by Internet-based protocols and services. This is because the Internet offers the greatest versatility in serving users, who employ wire cable, fiber optics, wireless systems, and satellite links. The only data communication scheme that accommodates all of this so far is the Internet Protocol family (and in particular TCP/IP). The dominance of TCP/IP results from its widespread use within LANs and WANs, both in terms of use in the enterprise environment and at home. Essentially every modern computer from to the PCs to largest server in a server farm provides the foundation, as do the public networks that interconnect the computers and applications.

Table 5.3 Comparison of Information Requirements for Traditional Telecommunication Networks

Type of Information or Service	Bandwidth Requirement (kbps)	Variability of Data Rate	Latency Requirement (One Way)	Connection Paradigm	Traditional Network
Telephone (POTS)	16 to 64	40% duty cycle for voice	Between 150 and 500 ms	Circuit switched	PSTN
Fax	9.6 to 64	Continuous but time varying	Between 150 and 500 ms	Circuit switched	PSTN
Local area network (LAN)	10 Mbps to 1 Gbps	Variable and bursty	<1 second, depending on application	Packet switched	Ethernet (IEEE 802.3)
Internet access	14.4 to 2,048	Highly variable to bursty	Between 150 ms and 2 seconds, depending on content	Circuit switched (virtual circuit)	PSTN; dedicated bandwidth T1/E1
Video teleconferencing	128 to 2,048	Continuous but time varying	Between 500 ms and 1 second	Circuit switched (virtual circuit)	T1/E1, Internet
Data file transfers, small large	64 to <15,000 45,000 to 1 Gbps	Continuous during transmission	Between 500 ms and 10 seconds Up to 2 hours	Any	Internet SDH (fiber), Gigabit Ethernet
Digital TV SD	Approximately 2,500	Continuous but time varying	Up to 10 seconds for broadcast or narrowcast	Dedicated bandwidth	Satellite area coverage transmission; local radio broadcast and cable systems
HD	10,000	Continuous but time varying	Up to 10 seconds for broadcast or narrowcast	Dedicated bandwidth	Satellite area or spot beam coverage; local radio broadcast or cable systems

5.1.2 Analog-to-Digital Conversion

The process by which a continuous information signal is converted into a stream of digital data is called A/D conversion. Quite obviously, the reversal of digital data to analog information is D/A conversion. A/D conversion is necessary to convert analog signals, which generally are associated with some natural form of activity or communication (e.g., voice or image), into a bit stream that can be applied to a digital network. Many user devices such as personal computer audio processors ("sound cards") and video codecs perform this function as a first step prior to delivery of the information to the Earth station. In applications such as MSS telephony and DTH reception, the user terminal must perform the appropriate direction of conversion.

The first step in A/D conversion is to sample the analog signal at equally spaced points in time. Electronic devices that perform that measurement and conversion operate at high rates, corresponding to twice the highest baseband frequency. The voltage range being sampled is divided into narrow bands called quantization levels, illustrated in Figure 5.2. (This linear quantization scheme is not representative of real-world A/D systems which use some form of audio compression.) The logic of the A/D converter puts out a numerical code, typically in the binary system.

The A/D conversion process introduces some inaccuracy into the end-to-end link. Called quantization error, it is the small amount of information loss due to the use of steps (as opposed to transferring a continuous range of voltage). The basic relationship for quantization error, measured by the ratio of signal power to quantization noise, is:
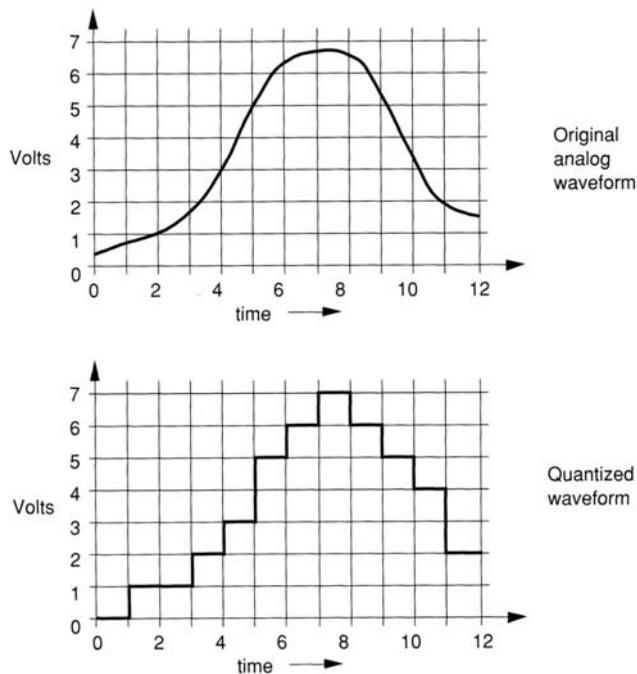
Figure 5.2 Comparison of the original analog waveform with the quantized version, based on 8 levels (or 3 bits).

$$S/N_q = 2^{2M}$$

where M is the number of bits per sample. The same formula [1] can be expressed in decibels as follows:

$$S/N_q = 2M \cdot 10 \log 2 = 6.02 \cdot M$$

The importance of having enough bits per analog sample is evident in this relationship. The main features of the A/D and D/A converter that affect end-to-end performance include the following:

- Resolution, specified in terms of the number of bits in the digital code used to represent each sample. An 8-bit code, for example, can represent $2^8 - 1$ (i.e., 255) different levels. The more bits in the code, the more steps and hence, the smaller the range of each level. That improves the resolution and hence the accuracy (including S/N_q) with which the sample can be reproduced on the other end.
- Speed, based on how fast the device can perform a conversion.
- Sampling rate, the number of times per second the analog signal is sampled and converted into the code word. The minimum sampling rate is determined by the baseband bandwidth of the analog input. For Nyquist-type A/D converters, the sampling rate is at least twice the highest baseband frequency. An oversampling A/D converter operates at an elevated sampling rate, in the range of 2 to 64 times the highest baseband frequency. That facilitates other digital processing steps, such as compression.
- Linearity, the deviation from precise recreation of the input signal over the entire range of minimum to maximum signal voltage range. That, in turn, determines the reproduction fidelity of the combined A/D and D/A process.
- Dynamic range, also related to linearity and resolution, indicates the spread of analog voltage levels that are tolerated with acceptable fidelity.

This is possibly more detail about A/D conversion than most readers would care to know. However, it does bring to light the criticality of choosing the right parameters for the conversion, lest the result be unacceptable to users. In the case of telephone service, the standard sampling rate is 8 kHz (8,000 times per second), which is slightly more than twice the highest signal frequency (i.e., 3,400 Hz). With regard to quantization, good telephone quality is obtainable with 128 levels (seven bits when coded in binary). Standard pulse code modulation (PCM) adds an eighth bit for synchronization, yielding a transmission speed of 64 kbps per telephone channel. Companding in the form of nonequal levels is used to help overcome some of the quantization error and channel noise by expanding low audio volume on transmit and compressing them on receive.

Another important consideration in satellite network design is the number of times that an analog information source might be exposed to the A/D and D/A process. The example of an end-to-end communication link in Figure 5.3 shows how an analog input can experience four conversions on the way to the distant
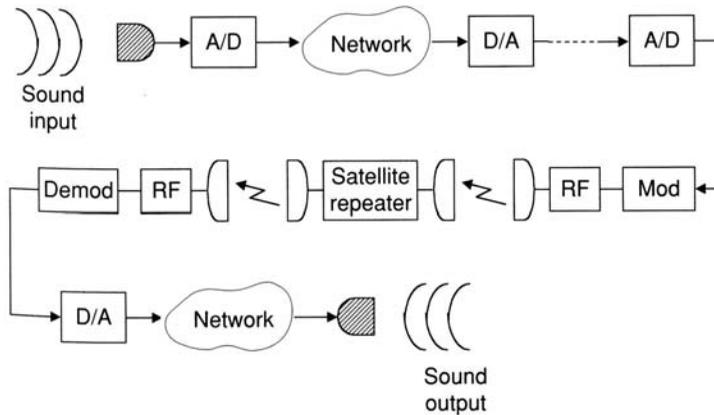
Figure 5.3 Example of an end-to-end communication link, indicating two pairs of A/D-D/A conversion.

end user. Each pair of conversions introduces distortion in the form of quantization error (due to the selected number of bits per sample) and nonlinearity. Therefore, the final product may not provide the desired quality. What sounded or looked acceptable after a pair of conversions (e.g., A/D followed on the other end by D/A) might indeed be considered of poor reproduction after two or more such pairs. While it is desirable to preclude more than two pairs of conversions, that may end up being outside the control of the satellite network designer.

5.1.3 Compression

The number of bits per sample can be reduced by taking advantage of predictable aspects of a particular signal type such as voice or video. Removing excess bits in the sample effectively compresses the signal bandwidth by allowing the resulting data to be sent at a lower speed. That requires that the compression equipment have intelligence in the form of a microcomputer or DSP and an algorithm for controlling it. In the following discussion, we assume that the information already has been digitized at the source.

There basically are two classes of compression: lossless and lossy. In lossless compression, the reduced bit stream still contains enough of the original information content to allow the decompressor to recreate the input without distortion. It is said to be a reversible process because the output is, for all intents and purposes, exactly the same as the input. This is the one case in which compression and decompression can be done over and over without any additional distortion. To make that work, the input signal usually must be constrained to fit within a certain bandwidth and amplitude range, such as is common in the PSTN.

An example of near-lossless compression is adaptive-differential PCM (ADPCM), a standard voice-encoding technique that cuts in half the number of bits per sample while maintaining the quality of voice. The speech is reproduced without distortion even though only half the bits are transferred. On the other hand, there does appear to be a tradeoff in that the speed with which inband data can be sent is slightly reduced. Another type of lossless compression called run-

length encoding replaces a string of identical characters with a short sequence that represents the string.

That brings us to the more interesting topic of lossy compression, which, as the name implies, does not precisely reproduce the input information. Why would anyone accept lossy compression for a commercial service, satellite or otherwise? Well, it turns out that every means of communication is lossy, and we are all accustomed to "filling in" missing details. Our brains are very good at doing that, which explains why we believe that the typical magazine color image looks like a photograph. However, if one looks closely, the actual image is not at all of photographic quality. Even greater loss of quality can be seen in newspaper printed pictures, especially ones in color. The same principle applies to sound. We certainly can enjoy a conversation over the telephone or a music broadcast on AM radio. Both of these limit the bandwidth of the information to something substantially less than what the source generates and what the average human is capable of hearing.

Lossy compression is used in perhaps its most effective manner in the typical JPEG image file (type .JPG) in wide use over the Internet. Images that were compressed with JPEG by a factor of, say, 40 are definitely not perfect but nevertheless provide utility and even enjoyment. Of course, JPEG includes the option of lossless compression, but that would not be as effective in reducing the quantity of data in the compressed image.

Compression of motion pictures combines the properties of lossless and lossy compression and must deal with images, motion, and sound all at the same time. As discussed in our previous work [2], the MPEG gave users and service providers a workable approach to information compression and integration. The amount of compression typically is adjusted to the type of image and the needs of the user. With more and more compression, the picture quality can look choppy, particularly when there is rapid motion. In video teleconferencing applications, compression down to 384 kbps usually is acceptable. However, standard definition TV is more demanding, and reduction to about 1.5 Mbps is considered acceptable using currently available processing algorithms. That was not always the case; in the first edition of this book, we indicated that the industry standard for digitized commercial TV was at least 45 Mbps. Colors can be reproduced with excellent quality no matter how much compression is used, simply because hue information occupies very little bandwidth to start with. Advances in the compression of full-motion video are anticipated as faster digital processors and more sophisticated algorithms are developed.

5.1.4 Error Detection and Correction

Noise, interference, and distortion on a digital communications link will on occasion confuse the receiving data processing equipment, causing bits to be misread. We measure the result through the bit error rate (BER), which is the ratio of error bits divided by the total bits during a particular time interval. The effect of BER can be relatively minor, as in the case of an audible "click" in voice communications, or catastrophic if the bits represent numerical data such as computer programs or important statistics. As discussed in Chapter 3, most Internet connections over

satellite employ the TCP/IP protocol that guarantees delivery of all of the data. TCP provides automatic retransmission of bad packets, which results in unacceptable delays with high BER. A protocol gateway can improve delay performance, but a strong FEC code can yield big improvements in throughput and latency because it reduces the need for TCP to activate retransmission in the first place.

Consider the following example of a satellite link operating at 10 Mbps and a BER of 10^{-6}. This means that, on average, there will be ten errors per second, resulting in a potential need for error recovery by TCP every tenth of a second. This would have a very serious effect both on the throughput and latency since TCP would slow data transfer to a crawl. By reducing the BER to 10^{-9}, which is easily accomplished with currently-available FEC, the errors occur an average of 100 seconds apart, which is about one every minute and a half. Combined with the function of the protocol gateway or accelerator, error would be sufficiently mitigated to make the user experience nearly the same as that of a T1 circuit.

To control errors in the communications link, there are techniques for preparing the data prior to transmission. In FEC, data are coded to expand the number of bits, and a sophisticated decoder is used by the receiver to reverse (correct) errors that occur along the link. This is a difficult concept to explain or understand without resorting to the mathematics of statistical communication theory and information theory [3]. Nevertheless, the following paragraph reviews a simple example that would not actually be used in practice but that identifies some of the concepts.

Suppose the transmitting end codes the binary 1 into the 4-bit sequence 1111 and the binary 0 into the sequence 0000. The 4-bit sequence is called a code word. This simple code increases the data transmission rate by a factor of 4 since the same information goes through the channel in four times the time. The receiver is designed to be able to recognize that any of the following sequences with no errors or one error started out as 1: 1111, 1110, 1101, 1011, or 0111. It would then deliver a 1 to the user. A 0 bit being coded and sent would work in a complementary manner (e.g., by reversing the 1s to 0s and vice versa in the example). This coding scheme can handle one bit error per code word. If a received code word has two 1s and two 0s, there were at least two errors in reception. In that case, the coder cannot determine the correct bit sense. The way to handle that in the decoder would be to make a guess, recognizing that the choice would have a 50% chance of being correct. The code is a simple example of error-correcting codes that are more powerful in their ability to correct errors caused by the link. Also, actual codes are more efficient in their use of the channel, increasing the data rate by a factor of 2 or less.

Two classes of FEC codes are applied in the majority of cases: block codes and convolutional codes. In the block code, a fixed length of input bits from the user side (say, k bits in length) is translated into a somewhat longer fixed length of a code word (n bits in length). Thus, there is a one-to-one mapping between any possible k-length input word and the corresponding n-length output word. Information theory, credited to Claude Shannon, shows that the error correction performance of the block code improves as the code length increases. A popular example of a block code is the Red Solomon code (RS code), first applied to digital TV over satellites and later to terrestrial DTV networks as well. It uses a data

word block of 188 B, corresponding to MPEG 2 format, and a code block size of 204 B. Hence, R-S of this type is designated (204, 188).

The convolutional class of FEC codes is also heavily applied to satellites and other wireless media. In convolutional coding, the input bits are not grouped in blocks but rather are run through a computation device that, in turn, outputs another stream of bits at an elevated bit rate. The increase in data rate is again related to the ratio of $n/k < 1$, causing the bandwidth on the link to be increased by the inverse ratio. An example of the type of computational device is a tapped shift register such as that shown in Figure 5.4. An important design parameter in convolution coding is the constraint length, which is the quantity of input bits that influence the encoder output of a single bit. Turbo codes are another category that can be very effective over power-limited satellite links. A popular form is referred to as the turbo product code (TPC).

For data communications, a typical objective is one bit error per 100 million bits transmitted, expressed as 10^{-8}. The channel efficiency of error-correcting codecs (often included in the high-speed modem in the Earth station) is specified by the coding rate (R), which is equal to the ratio of information bits to coded bits on the channel, or simply k/n. As shown in Figure 5.5, R determines how much the error rate on the channel is improved (reduced). For example, a rate $R = 1/4$ convolutional codec can reduce the error rate on a typical satellite link by three orders of magnitude (e.g., from 10^{-2} to 10^{-5}), which is equivalent to a 2-dB to 3-dB reduction in the demodulator threshold for a constant error rate. Evaluating FEC performance involves trading off the threshold (power) reduction against the increased bit rate (bandwidth) imposed by the code. Even greater benefit was obtained by applying RS and convolutional coding in the same link (concatenated coding). The only penalties with using that type of coding are the increase in bandwidth on the RF channel (because the data rate on the satellite is increased by a factor equal to the inverse of the coding rate) and the complexity and cost of including a codec in the RF modem. Two other block codes, BCH and low-density parity check (LDPC), are gaining favor. The performance of many FECs can be improved by the technique of interleaving where bits are shuffled according to a known pattern to reduce data loss from burst errors.
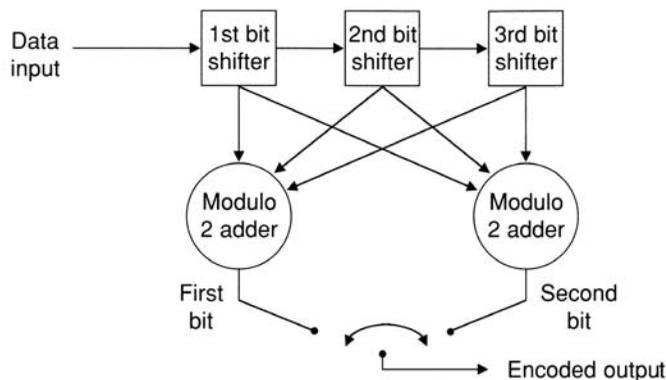
Figure 5.4   Convolutional coder block diagram.
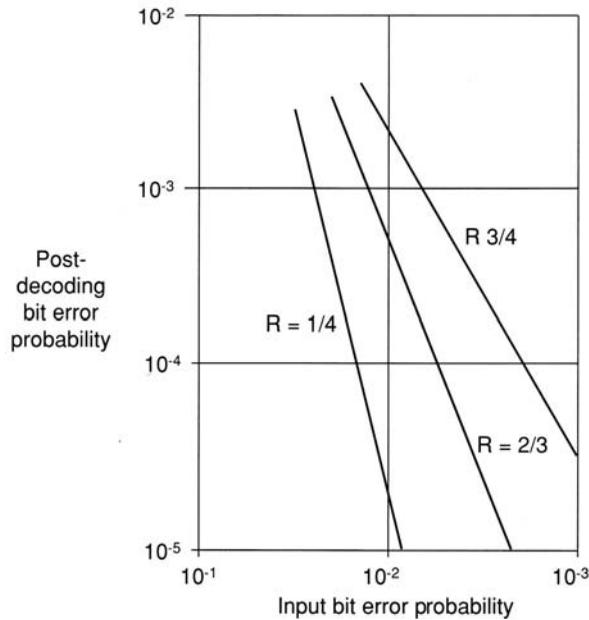
Figure 5.5 Improvement in BER performance versus code rate for coded binary data channels.

5.1.5 Scrambling and Encryption

While satellite communication offers many benefits from its wide-area coverage capability, it brings with it a concern for the security of the underlying information. Commercial users worry that their valuable data can be intercepted and used for undeserved gains. Also, government agencies involved with public safety and other security matters cannot allow senstitive information to fall into the wrong hands. There is also the concern that someone might inject false data into an environment and cause difficulties and even total failure of an important operation of some type. For all those reasons, designers and operators of satellite networks must make proper provision for security. As with any type of countermeasure, the steps taken should be in proportion to the threat, to minimize the spent resources and to not unnecessarily interfere with normal day-to-day use of the system. The following sections provide some guidelines and technological options for addressing those concerns. However, this subject is extremely complex and considers also that security must deal with intentional misuse and intrusion, for which clever humans are ultimately to blame. For those reasons, security technology must constantly be reviewed to counter threats that can become more sophisticated as the opportunities (and rewards) for abuse tend to increase. Another issue is that security technology and encrypted information may be subject to import and/or export restrictions and controls in particular countries.

5.1.5.1   Security Risks in Satellite Communication

If we were to take no security measures to protect our information, then the fact that a satellite broadcasts information over a wide area allows unintended users

with the proper receiving system to get any of our information. It is the same with local radio broadcasting, but the satellite multiplies the opportunities. That, of course, is an advantage in services designed to reach the largest possible audience. However, there are a variety of services in which the network provider must control access. The analog form of TV transmission is the most vulnerable because those transmissions are easily monitored with inexpensive receiving equipment. Digital transmission provides a certain degree of security because a somewhat higher barrier is presented by the greater cost and complexity of the receiver. However, the digital set-to-box is widely available and many services are not encrypted, or free-to-air. We cover many of the commercial solutions offered on the market at the time of this writing. This is not exhaustive and, as discussed below, the nature of the threat escalates as our ability to counter it improves.

The following are representative of some reasons why network operators require security:

- Preventing piracy of the information content. This is the most fundamental issue in broadcast application, leading first to the use of scrambling and subsequently to encryption.
- Controlling the customer base, also called conditional access. The network operator has the power to turn individual service on and off either through the network itself or through a separate "back channel."
- Protecting intellectual property (IP) rights, derived from the fact that satellites are used to distribute information such as TV, music and computer programs that are the property of someone else. The rights of the IP holder must be protected, for which conditional access and either scrambling or encryption are appropriate.
- Simplifying pay-per-use services to the end user. For example, PPV TV and information file transfer are popular applications of satellites but demand the type of control cited above.
- Providing levels of service and privacy for sensitive information. This is not a case of requiring payment but requiring that the information be blocked for reasons of competitive sensitivity or even national security.
- Theft of service and preventing disruption of the normal operation of the network. This is a concern for all telecommunication services and will depend on the nature of the satellite network.

5.1.5.2 Scrambling

For some commercial applications, enough of the security requirements given in Section 5.1.5.1 can be provided by scrambling the information according to a predefined pattern. Scrambling is achieved by altering the normal analog waveform or bit pattern so that it cannot be used unless the receiver can restore the original through foreknowledge of the scrambling pattern (Figure 5.6). Basically there are two ways to do that. The cheapest is to simply add an interfering or jamming signal into the baseband and then to filter it out at the receiving end. Alternatively, the signal itself can be altered in time or frequency according to some pattern, such that only authorized receivers can reverse the process to produce a proper signal.
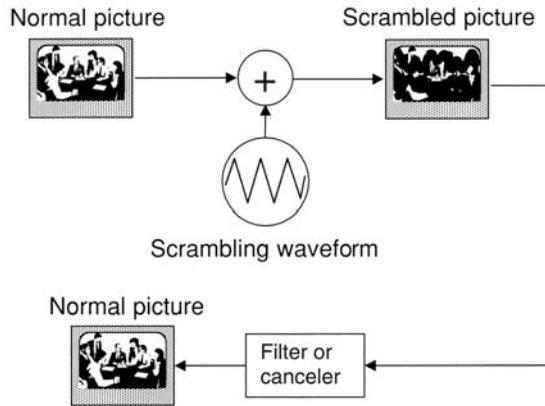
Figure 5.6 The principle of video scrambling as applied to TV security.

The additive technique is found in older analog cable TV systems, where costs are held to the lowest and people are already paying something for cable access. The jammer can be a simple sinewave or repeating bit pattern, either of which is easily filtered out or canceled with an active circuit within the home set-top box.

There are two elements to scrambling technology: the technique for scrambling the information, using appropriate sending and receiving devices, and a key to the sequencing of the scrambled bits and pieces. The devices can be offered through the commercial market, but they will not descramble unless the proper key is inserted into the receiver. The key mechanism is in the form of a digital code that, in turn, represents the scrambling sequence. If the scrambling sequence repeats itself frequently (which is determined by the technique and the length of the key), then one can analyze enough segments to obtain the key. That can be avoided if the key is long enough so that the repeat happens perhaps only once a week or year. Then it is simply a matter of changing and then redistributing the key to all authorized receivers. This principle is used in all encryption systems as well.

5.1.5.3 Digital Encryption

The technologies associated with the various forms of digital encryption have been with us for centuries. Encryption is, by definition, the use of ciphers, which are nothing more than codes. However, simple codes, in which a given letter in the alphabet or a number from 0 to 9 is converted into another on a one-by-one basis, are not particularly secure because they can be easily broken through analysis of the text. The better approach was invented in Nazi Germany in the form of the Enigma machine, a device that gave a rolling code, in which the conversion from letter to letter changed with each input. Thus, the letter A might be coded as a 6 the first time and as a Y the next. The decoding process would have to follow the path of the rolling code to recover the information. That is applied in some scrambling systems.

Over the years, more sophisticated and efficient encryption techniques were developed and reduced to standardized computer code or electronic circuitry. As with scrambling, there is the encryption technique and the key. There are really

two classes of techniques: private key encryption, of which the Digital Encryption Standard (DES) is the most common example, and public key encryption, which is embodied in the RSA standard (named after its inventors, Ronald L. Rivest, Adi Shamir, and Leonard Adelman). For most satellite communication applications, private key encryption is most popular. The DES standard can be utilized in the United States, where it was developed under government sponsorship.

The adequacy of DES was challenged when it became the foundation of Wired Equivalent Privacy (WEP), the encryption technique most common in IEEE 802.11 (WiFi) networks. Simple 64 bit and even 128 bit encryption keys have proven ineffective against attacks from hackers using commercial computers. Techniques that frequently change the key thwart hackers because they are unable to derive the key before the next change is effected.

For all intents and purposes in commercial satellite communications, public and private key encryption are highly adequate to protect user data from intrusion and unauthorized interception. Even without encryption, it has been argued that the requirement to use a specialized receiver provides a low level of security. That argument does not hold up in attractive consumer services with low-cost receivers, such as DTH TV and S-DARS, so encryption becomes the norm. The techniques and the equipment do their jobs very well, but the problem comes from the need to adequately protect the keys. Because of the repeating nature of the encryption techniques and the tendency of pirates to be able to find out what they want, keys are changed routinely and often. They usually are distributed over the same satellite link that is to be protected. If pirates have been able to overcome the key once, they are in the position to acquire the new keys when those keys subsequently are broadcast over the satellite.

Auxiliary systems like smart cards and special chips add another level of security on top of encryption. An authorized user needs three things: the decryption device, the key, and the smart card. If the system is compromised, the operator distributes new cards (which is cheaper than distributing new encryption devices). Another defense is to require users to hook their equipment to the Internet so the operator can transfer information such as the new key or an encryption technique upgrade. As is evident, security is a process of implementing a technique that is adequate until broken, then strengthening the system in response to the increasing capability of the hacker or pirate. It is, after all, a game of cat and mouse, in which the mouse has all the motivation to break through and obtain the information (TV programs, computer programs, and the like) without paying the originator. Of course, the pirate must pay to play the game, so there is a cost nevertheless.

5.2 Digital Modulation

Digital modulation involves transferring the bit stream to an RF carrier, for which techniques like QPSK and 16QAM are popular. This discussion is an introduction to the theory and the practice of digital modulation design, with emphasis on the typical satellite link. The steps in the process include generating a bit in the form of a pulse that meets certain criteria for bandwidth and shaping; transferring those shaped pulses onto a carrier, typically using some form of PSK; allowing for

impairments, noise, and interference along the link; and providing a suitable demod-ulator capable of recovering the pulses from the carrier and cleaning them up to reproduce a digital bit stream with minimum errors due to the previous elements. Readers wishing a more detailed discussion suitable for making specific design decisions should review [4]. Nontechnical readers not needing an understanding of modulation theory can skip this section entirely.

Nearly all digital modulation relies on the basic characteristic of a pulse, shown in the time and frequency plots in Figure 5.7. The pulse has a fixed duration of T seconds, which means a continuous stream of such pulses would have a rate $1/T$ bits per second. If $T = 0.001$ second (1 ms), then $1/T = 1,000$ bps (1 kbps). The frequency spectrum of the pulse, shown in the figure, is a maximum at zero frequency and passes through the X-axis at a frequency equal to the bit rate ($1/T$, or 1,000 Hz for the previous example). Successive sidebands have zero values at multiples of $1/T$. The bandwidth of the main lobe is proportional to $1/T$, the bit rate. The formula for the spectral shape of pulse amplitude is simply the ratio

$$\sin(\pi FT)/(\pi FT)$$

That formula is plotted more precisely between −4 and +4 in Figure 5.8 and converted in relative power in decibels in Figure 5.9. The shorthand way of referring to the formula is $\sin(x)/x$. The value at zero frequency, which is actually referenced to the carrier center frequency, is not zero even though $\sin(0) = 0$. In fact, the value of $\sin(0)/0$ is 1, because for small angles $\sin(x) = x$. The important relationship here is that a fixed pulse length produces maximum energy at zero (e.g., carrier center frequency) and an extended frequency spectrum.

Because of bandwidth taken by sidebands, the spectrum of the pulse should be filtered to minimize the occupied bandwidth. That is illustrated in Figure 5.10, where the attenuation characteristic of a low-pass filter is shown superimposed on the spectrum of a rectangular pulse. After filtering (the heavy line), the output spectrum is confined within the bandwidth of the pulse spectrum's main lobe. Due to the same mathematical relationship that causes a rectangular pulse to have an

Time waveform of a single rectangluar pulse
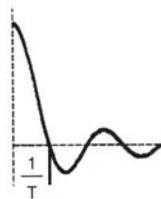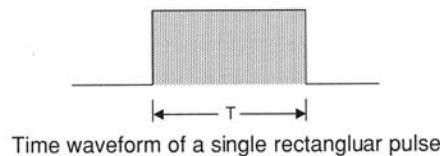
Baseband frequency spectrum of a rectangular pulse

Figure 5.7   Characteristics of a rectangular pulse in the time and frequency domain.
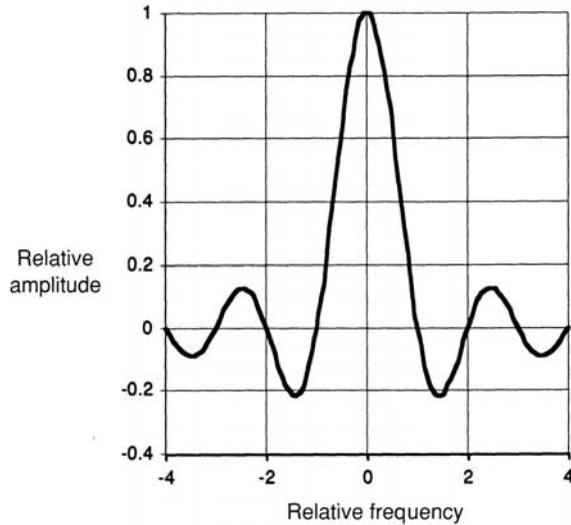
Figure 5.8   Amplitude versus frequency spectrum of a PSK carrier with rectangular pulse modulation.
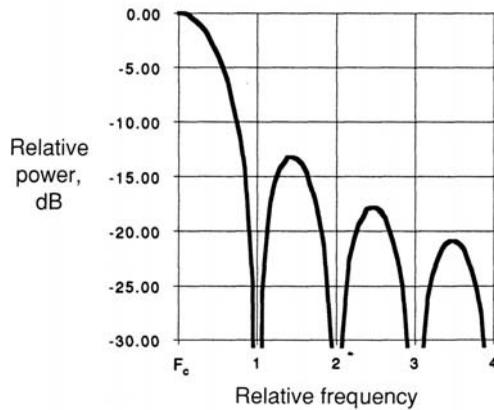
Figure 5.9   Power spectrum of rectangular pulse modulation shown on a decibel scale.

infinite spectrum, a limited spectrum alters the pulse so that it has infinite duration in time. Instead of a rectangular shape, a filtered pulse has a time waveform with an oscillating tail. A stream of filtered pulses contains residual tails, which increase the chance for errors. That is comparable to increasing the link noise level or adding interference; hence, the presence of tails is called intersymbol interference (ISI). If too little of the pulse spectrum is passed by the link, the ISI can so degrade reception that satisfactory operation is impossible. That, of course, is why adequate modulation bandwidth must be allowed.

   The error-causing effect of ISI can be reduced by aligning the pulse tails so their zero crossings occur at times when the pulse main lobes are at their peaks. That can be done by proper filter design and demodulator timing. Figure 5.11 shows an ideal rectangular spectrum and a spectrum shaped like a cosine curve (called the raised cosine spectrum). Both occupy approximately the same band-
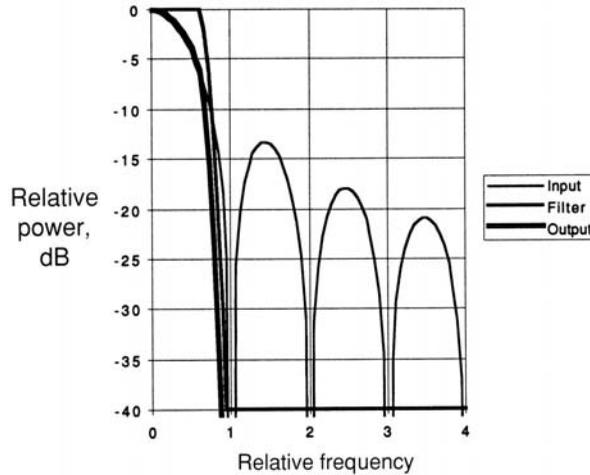
Figure 5.10   Baseband frequency spectra of a rectangular pulse before and after filtering to reduce the transmission bandwidth.
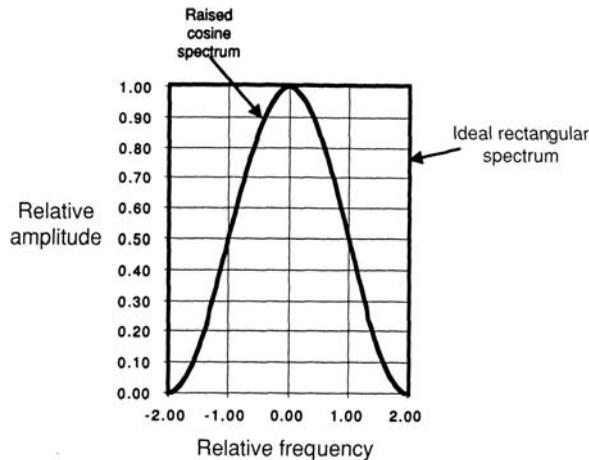
Figure 5.11   Frequency spectra of bandwidth-limited pulses using rectangular and raised cosine shaping.

width. The corresponding pulse waveforms in time are plotted in Figure 5.12, clearly showing the reduction in the oscillating tail produced by spectral filtering. Both pulse waveforms have zero crossings at the times corresponding to main lobe peaks ($-3, -1, -1, 1, 2, 3, \ldots$), thus precluding ISI. In addition, the raised cosine spectrum pulse has a much weaker tail, making it less sensitive to timing errors in the demodulator. This type of shaping is one technique used to efficiently reduce the modulated bandwidth of a digital carrier. Typically, the necessary RF bandwidth of a properly filtered pulse is $(1 + \alpha)$ times the bit rate (expressed in Hertz), where α is the roll-off factor. Typical values of α are in the range of 0.2 to 0.4.
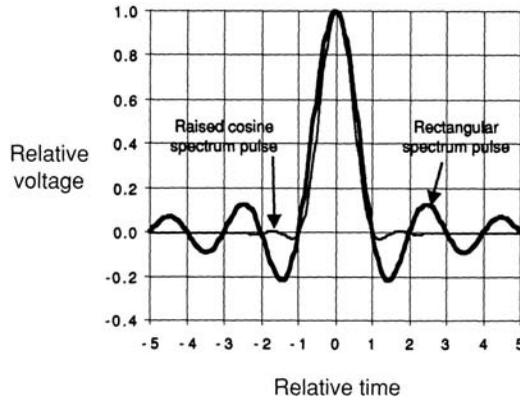
Figure 5.12 Comparison of pulse waveforms with limited bandwidths for rectangular spectrum and raised cosine spectrum shaping.

5.2.1 Frequency Shift Keying

An early form of digital modulation used in radio communications was FSK. Illustrated in Figure 5.13, FSK is obtained by switching between two discrete frequencies: F_1 and F_2, which correspond to the ''1'' and ''0'' states, respectively (or vice versa). That aspect causes FSK to consume twice the bandwidth of PSK. The principal advantage of FSK historically was that simple hardware could be used in the modulator and the demodulator. That lack of sophistication also means that FSK requires more signal power for the same performance in terms of error rate, which would have to be reflected in the link budget analysis. FSK was used in the first telephone modems, which were capable of between 600 and 1,200 bps. There may still be a role for more efficient FSK waveforms that overlap the spectrum pairs.

5.2.2 Phase Shift Keying

The technique of PSK, discussed in Chapter 4, is the most common system for transmitting and receiving high-speed digital information over satellites. PSK
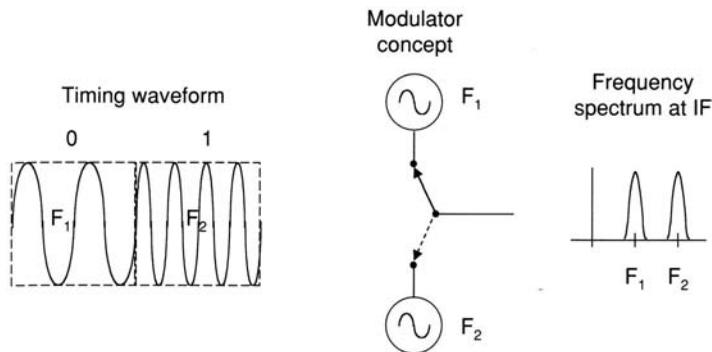
Figure 5.13 FSK as shown in time and frequency can be generated from two oscillators of different frequencies.

modems operate at rates between 64 kbps and 1,000 Mbps. BPSK uses two phase states: 0 and 180 degrees; QPSK uses four states split into pairs: 0 degrees and 180 degrees (the in-phase, or "I," component) and 90 degrees and 270 degrees (the quadrature, or "Q," component). Four possible phase states are illustrated in Figure 5.14 using vector diagrams, each representing a two-bit combination. A property called orthogonality is used in QPSK, wherein the I and Q components are separated by 90 degrees from each other, shown in Figure 5.15 by orthogonal vector pairs. That is similar in effect to the use of cross-polarization isolation to achieve frequency reuse (see Chapter 4). QPSK transmits 2 bits per symbol, meaning that the vector position in Figure 5.14 represents 2 bits of digital information. As a result, QPSK is efficient in the way bandwidth and power are used. Adding another bit per symbol forms eight-phase PSK, a technique that increases capacity but subjects the signal to greater degradation from noise and interference and the repeater impairments.
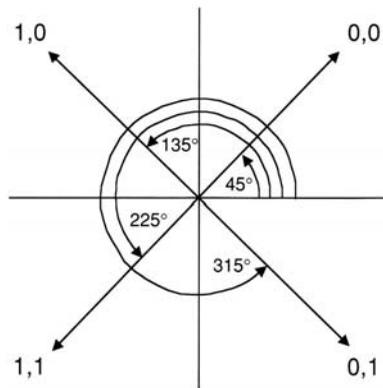
Figure 5.14   QPSK modulation vector diagram, indicating the four vector (phase) states and the corresponding encoded bit patterns.
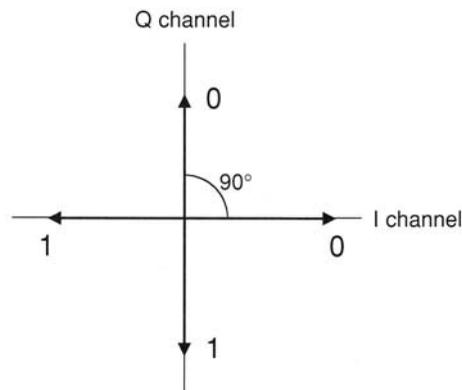
Figure 5.15   Orthogonal properties of the I and Q channels in QPSK.

5.2.3 Amplitude and Phase Shift Keying

QPSK is an efficient modulation method principally in terms of power; however, there are circumstances where its occupied bandwidth exceeds what is available or affordable. A solution may be found through a more complex waveform that comes under the general class called bandwidth efficient modulation (BEM). A point to keep in mind is that while BEM may be efficient in terms of bandwidth, it is not in terms of power. It will be necessary to compensate for this trade-off by using more satellite EIRP or a larger Earth station antenna (i.e., higher G/T on receive).

The most common form of BEM is amplitude and phase shift keying (APSK), in which both the phase and the amplitude are varied to carry the information bits. Figure 5.16 presents the phasor diagram, also referred to as the signal constellation, for the version of 16APSK adopted by the DVB-S2 standard. In it we see two concentric circles, the first identical to a QPSK waveform and the outer circle being divided into 12 sectors (i.e., 12-phase PSK). In total, there are 16 possible signal states to correspond to 4 bits of data. This is indicated in the figure where there are 16 different combinations of four bits, with only one bit difference between adjacent locations in the signal space.

This waveform uses a shift between two amplitude states and up to 12 different phase states. It was devised to improve performance in a repeater channel of the type found on bent-pipe repeater satellites, as discussed at the end of this chapter. A 32APSK waveform that adds an outer ring with 16 additional phase states produces five bits per signal state.
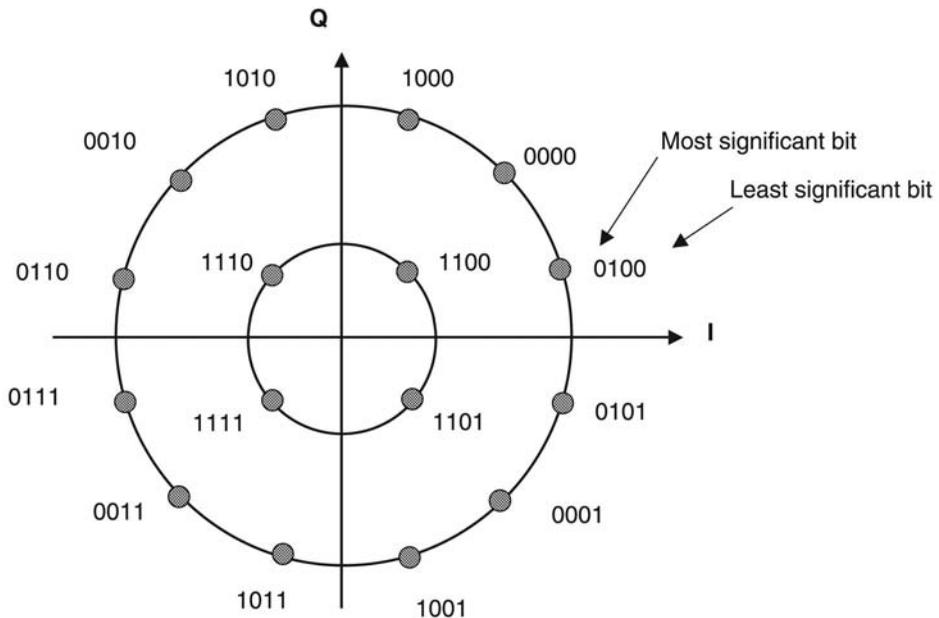
Figure 5.16 Vector constellation diagram for 16APSK, based on the DVB-S2 standard.

With the BEM constellations, there are always smaller data state regions in the phase-amplitude constellation than with BPSK or QPSK. Noise, however, corrupts both phase and amplitude, so can more easily cause wrong data state decisions at the receiver. Thus, with a higher data rate per unit of bandwidth, the C/N on the channel must be significantly higher than would be acceptable for PSK or FSK.

The ultimate payoff from BEM is in a dynamic arrangement called adaptive coding and modulation (ACM) wherein the modulation format and coding rate are automatically altered on the link to increase throughput under ideal conditions or fall back to a more resilient combination at a lower data rate under conditions of high channel error rate (e.g., fading).

Because of the higher data rate per unit of bandwidth, the C/N on the channel must be significantly higher than would be acceptable for PSK or FSK. Ideally, the RF channel should be linear to allow the complex modulation format to pass unimpaired. Continued improvements in satellite channel performance make hybrid modulation more practical. An example would be in a digital processing payload, in which the spacecraft receives a hybrid modulated carrier, recovers the bits, and retransmits the signal in a linear fashion.

5.3 Multiple Access Methods

The basic concept of multiple access is to permit Earth stations to transmit to the same satellite without interfering with one another. RF carriers can be maintained separate in frequency, time, or code, as indicated in Figure 5.17 by three perpendicular axes. Hence, the use of FDMA, TDMA, and CDMA, which are discussed in the following sections.
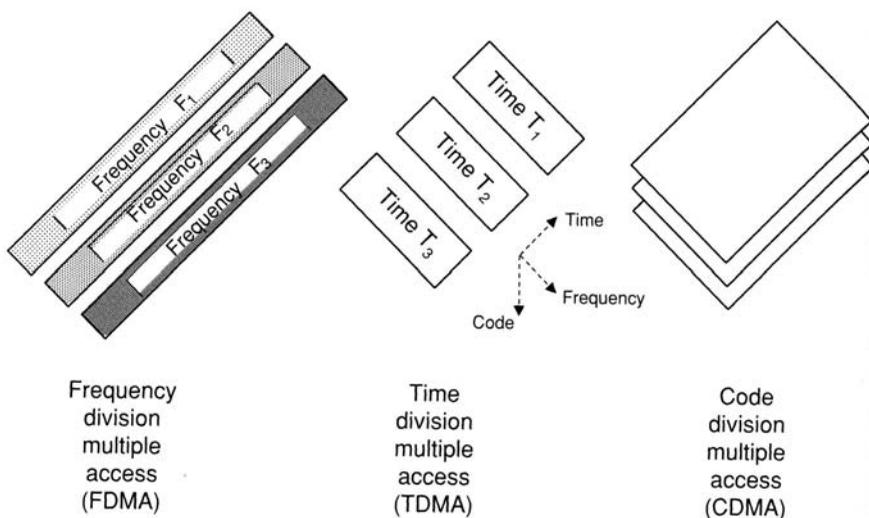
Figure 5.17 Generic multiple access methods using the three dimensions of frequency, time, and code.

5.3.1 Frequency Division Multiple Access

The FDMA technique is traditional in radio communications, since it relies on frequency separation between carriers. All that is required is that the Earth stations transmit their traffic on different microwave frequencies and that the modulation not cause the carrier bandwidths to overlap. Three such independent transmissions are indicated at the left of Figure 5.17 by long rectangles extending along the same "time" dimension but on different frequencies (indicated by different shading). A constraint in FDMA is that the sum of the bandwidths of the individual carriers cannot exceed the satellite's available bandwidth. Consequently, all three carriers in the uplink pass cleanly through the satellite repeater and are radiated toward the area of coverage on the ground.

The principle behind FDMA is that each Earth station or user terminal is assigned a separate frequency on which to transmit. That assignment can be fixed for time (permanently assigned), possibly changeable only by a manual command at either the uplink and the downlink or from a central control point. Alternatively, the assignment of the frequency can be dynamic (demand assigned, or demand assignment multiple access, DAMA), responding to user requests for service. Permanently assigned FDMA channels are useful for dedicated bandwidth services, which also are known as leased lines. Demand assigned channels are suitable for circuit-switched services, notably for telephony. Special signaling channels, also provided on an FDMA basis, are added to allow stations to request connections and to alert stations to incoming calls. In either case, the bandwidth of the channel must match the information bandwidth of the signal and allow for some guardband to prevent adjacent channel interference. The guardband may be as small as 5% or 10% of the channel signal (occupied) bandwidth, to allow for proper channel filtering and frequency errors.

An example of FDMA for a transponder with several carriers is presented in Figure 5.18. Each carrier is transmitted by a single Earth station, and thus there are a total of four such stations involved in this FDMA network. Indicated are the station location, frequency, and satellite EIRP. There are also smaller FDMA carriers used for narrowband data. This is a real spectrum representation for a C-band transponder used for distance learning, where each carrier and associated Earth station originates two classes at the same time. The lumps at the bottom are intermodulation distortion (IMD) produced by the active carriers in the transponder, which is being operated within 3 dB of multicarrier saturation.

The overall efficiency of FDMA is affected by intermodulation distortion (IMD), which results from multiple carriers in a common nonlinear amplifier like a TWTA or an SSPA. Figure 5.19 provides an example of how C/N varies, depending on the total input power of a common spacecraft TWT power amplifier. The X-axis displays the TWT input power with respect to saturation. The Y-axis represents carrier to noise ratio, broken down into uplink, downlink, combined uplink and downlink, intermodulation, and total link C/N. We see that there is an optimum operating point in terms of TWT input, for which the total C/N is maximized. As discussed at the end of this chapter, that type of analysis requires knowledge of the particular signal characteristics and a detailed link budget.

Over the years, FDMA became viewed as antiquated compared to TDMA and CDMA. However, the fact that FDMA carriers can be transmitted without
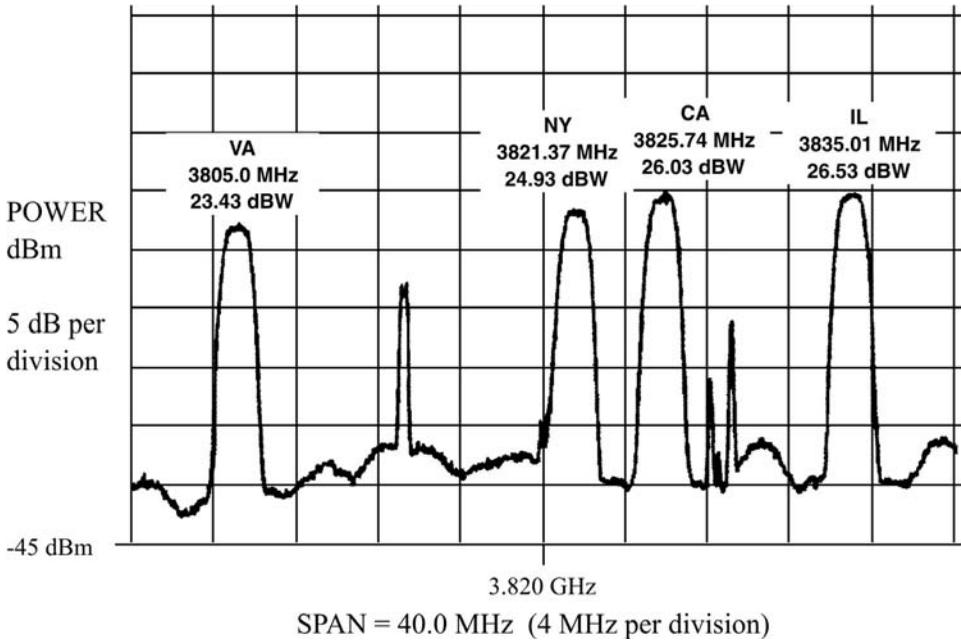
Figure 5.18   FDMA carriers within a C-band transponder with a total bandwidth of 36 MHz. Each
large carrier provides 3 Mbps of data for video conferencing and the smaller carriers
are for voice and data return channels.
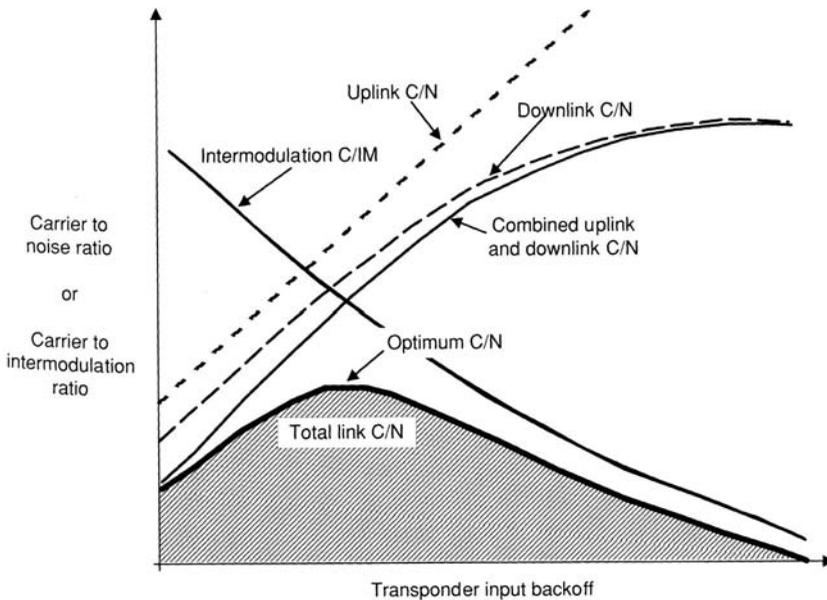
Figure 5.19   FDMA service optimization of the TWTA operating point for maximum C/N total (all
curves are representative but not accurate).

coordination or synchronization gives it longevity in telecommunications in general and satellite communications in particular. Many mesh network applications still rely on the simplicity and the efficiency of FDMA. Due to a variety of practical limitations on equipment power and cost, FDMA often is combined with the other multiple access methods to provide smaller networks that can operate more or less independently of one another. The basic principle of the channelized repeater and the transponder itself reflect the principle of FDMA.

There are some possible constraints to the effective use of FDMA, so satellite communication engineers have to do their homework before making a selection (that is true when doing the tradeoffs, because there are so many factors to consider in that type of decision). Some applications use a very low data rate per user, such as 4-kbps speech or text transfer. Using the SCPC FDMA technique, the RF bandwidth per carrier is comparable to those data rates. At only 4 kHz of bandwidth, the carrier is subject to two sources of carrier frequency error: LO drift, and Doppler frequency shift. Both could cause carriers to run over each other if not controlled. LO drift can be improved through the use of accurate and stable frequency sources, something that gradually is being introduced both in space and on the ground. Doppler shift depends on the relative motion of the satellite and the user. Therefore, systems in which both are in motion (e.g., non-GEO MSS) tend not to use narrowband FDMA.

5.3.2   Time Division Multiple Access

Earth station transmissions in a common TDMA network are all on the same frequency, and each employs the full bandwidth of the RF channel, which may consist of an entire transponder (full transponder TDMA) or a segment of bandwidth within a transponder (narrowband TDMA). In the center of Figure 5.17, the wide rectangles represent full-bandwidth transmissions from Earth stations within the same satellite coverage beam. Interference between transmissions, which are on the same frequency, is prevented by synchronizing the transmission so they do not overlap in time. That is a much more complex process than FDMA because a common system of timing and control must be employed by the Earth stations sharing the same satellite channel. Individual Earth stations, therefore, transmit their traffic in the form of bursts of information, necessitating compression of the traffic in time at the transmitting end and the complementary expansion at the receiving end. A similar technique is used in PC-based LANs, allowing several PCs to "talk" to one another on a common cable loop. In TDMA, the most appropriate modulation is digital in nature, typically QPSK, since that is compatible with the compression and timing requirements of burst transmission.

An example of a TDMA frame assignment is provided in Figure 5.20. The only inefficiencies are due to the need for guardtimes between bursts, analogous to the guardbands used in FDMA, "preamble" frame overhead bits used for synchronization and network control data. Full transponder saturated operation allows the bit rate to be run up to the theoretical maximum, with no significant loss due to IMD. (The latter decreases FDMA channel capacity by 50% or more.) There are a few degradations that apply to full transponder TDMA. The first is that the bursts themselves produce pulses of dc current demand by the transponder output
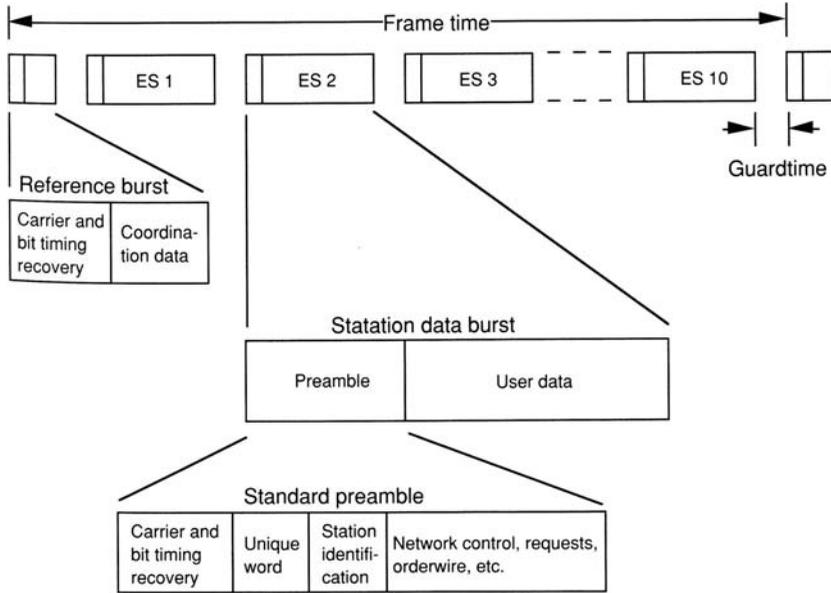
Figure 5.20   An example of a typical TDMA timeframe.

amplifier; some form of compensation typically is needed in the spacecraft power subsystem to prevent the pulses from affecting operation of other equipment on the same power line. Another concern is sidebands that can be generated in the transponder, potentially causing adjacent transponder interference. However, both factors can be dealt with effectively in the design of the modulation structure and the repeater itself, or by backing off the output amplifier.

Full-transponder TDMA networks were first introduced in the INTELSAT system and provided a maximum throughput of about 60 Mbps (based on a transponder bandwidth of 36 MHz, QPSK modulation), and no FEC. In a current system, link performance can be improved with FEC at the expense of some throughput. Moving to BEM will offset such a loss.

The alternative of narrowband TDMA has found wide application in satellite networks as well as the popular GSM mobile radio system. It is a blend of TDMA and FDMA, designed to reduce the power and bandwidth requirements for the Earth stations and user terminals. That is because the power required is directly proportional to the data rate; narrowband TDMA is much less demanding on the user terminal, which can be designed for very low power operation. That being the case, why not simply use FDMA and forget about having to provide accurate timing and synchronization? Even without considering the throughput efficiency of TDMA, there are a number of benefits to using the hybrid technique. First, the bandwidth of each carrier is proportional to the number of stations that share the same channel multiplied by the average data rate that each station is afforded. The wider bandwidth, which can still be well under 100 kHz for mobile voice applications, reduces the demand for maintaining a precise frequency; that saves guardband. Also, user terminals can frequency-hop to grab unused time slots, thereby increasing the total throughput. Frequency hopping has the additional

advantage of decreasing access delay and setup time. Thus, new applications (e.g., mobile data) and additional users can be inserted into what would have been a wasted resource. That principle is illustrated in Figure 5.21.

5.3.3   ALOHA Packet Multiple Access

A variant of TDMA called ALOHA greatly simplifies the control of a digital satellite network, although at a cost in efficiency. Developed at the University of Hawaii, it is basically the technique illustrated in Figure 3.3 for multipoint-to-point connectivity. Stations transmit packets of data only when they need to communicate; therefore, there is little need for end-to-end synchronization, and the packets arrive at the satellite at random times. The basic structure of the ALOHA timeframe is illustrated in Figure 5.22 for the specific technique called slotted ALOHA, in which packets can be transmitted only in the defined time slots. A reference station transmits the timing reference to all remote ALOHA terminals.

In the normal operation of ALOHA, the individual packets transmitted by several VSATs reach the satellite at different times and so do not overlap (jam each other). The bottom timeline in the figure shows the combined downlink. The receiving point, which in the typical VSAT network is at the hub station, acknowledges reception of every good packet. There will be instances, as indicated in the fourth time interval in Figure 5.22, when two stations happen to transmit within the same interval, resulting in what is called a collision. As a result, neither packet is received correctly at the hub. There will be no acknowledgement of successful reception, and so the originating VSATs will know that they need to retransmit the particular packets. A second collision is avoided because each station randomly inserts a short delay time before initiating the retransmission.

The throughput efficiency of ALOHA is illustrated in Figure 5.23. An unslotted Aloha network, which employs very simple Earth stations, cannot provide more than about 15% throughput efficiency due to the time delays introduced by the automatic retransmission process. We see that as that point is reached, the retransmissions increase the average end-to-end delay by many times over the basic packet
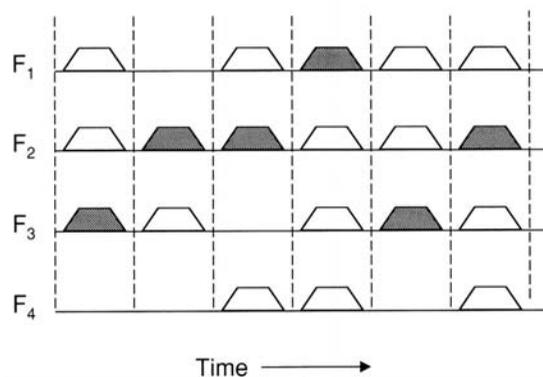
Figure 5.21   Illustration of the principle of frequency hopping applied to TDMA (F-TDMA). The shaded bursts are transmitted by the same station, utilizing empty time slots on different frequency channels.
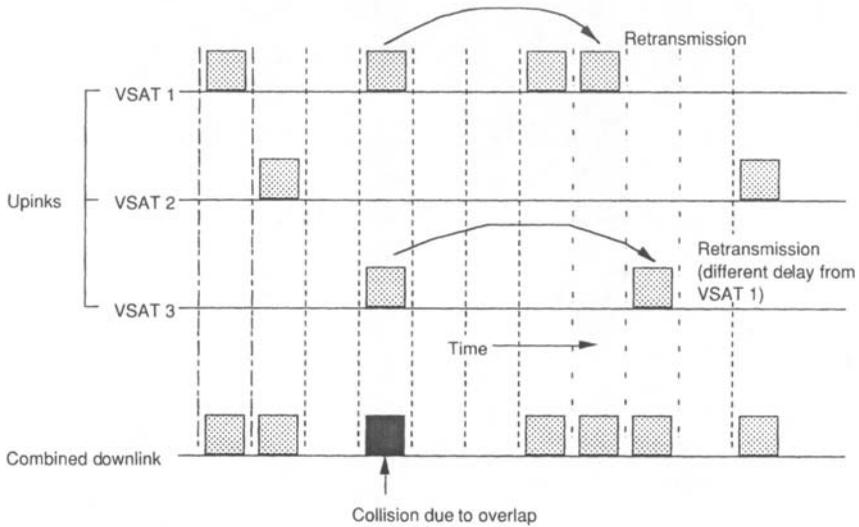
Figure 5.22   ALOHA protocol operation (slotted ALOHA).
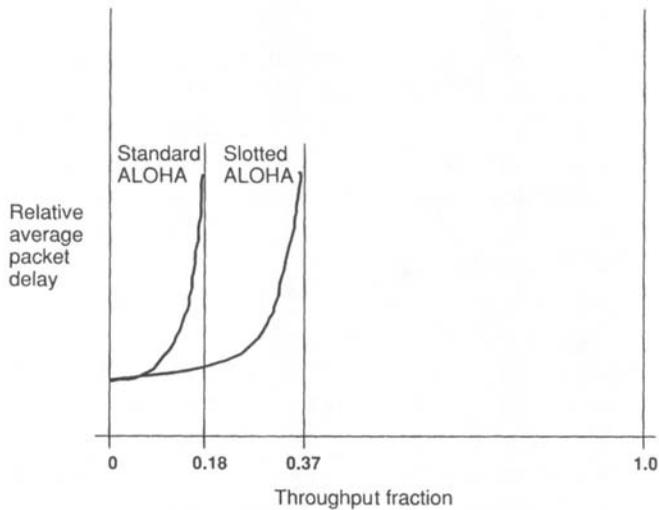
Figure 5.23   Performance of the ALOHA channel for the standard (unslotted) and slotted cases.

transmission delay over the single-hop satellite link. The slotted ALOHA technique essentially doubles throughput for the same delay.

The efficiency of ALOHA has been improved by allowing stations that have large blocks of data to pass the data in a preassigned block of time using a reservation mode. The more advanced TDMA architectures for VSAT networks incorporate features of standard TDMA, ALOHA, and the reservation mode to provide the greatest flexibility and efficiency of bandwidth utilization.

5.3.4   Code Division Multiple Access

CDMA combines modulation and multiple access to achieve a certain degree of information efficiency and protection through the technique of spread spectrum communications. First applied to guarantee security for military transmissions, it evolved into a multiple access system that promises better bandwidth utilization and service quality in an environment of spectral congestion and interference. The basic concept is to separate or filter different signals from different users, not by using frequency or time, but by the particular code that scrambles each transmission. Direct sequence spread spectrum (DSSS) uses a high-speed sequence of bits to perform spreading while frequency hopping spread spectrum (FHSS) rapidly shifts the carrier frequency to accomplish the same result. Most common is DSS, in which the code is a long sequence of bits, running at many times the bit rate of the original information. That expands the bandwidth by the ratio of bit rates; for example, if the random code sequence is 1,000 times the original bit rate, the resulting bandwidth is 1,000 times as well. On the surface, that seems very inefficient in terms of spectrum use, but it allows nearly as many signals to be transmitted one on top of the other [5, 6].

DSSS uses a pseudo-random noise (PN) binary sequence, which is nothing more than a computer-generated randomized sequence of bits designed to guarantee noncoherence with any delayed version of itself or any other PN sequence used in the same CDMA network. Figure 5.24 illustrates the basic process of sending and receiving. Each CDMA signal consists of the original data, protected by FEC, which are digitally multiplied by the PN sequence and then modulated onto a carrier using either BPSK or QPSK. At the same time the bandwidth is multiplied, the power spectral density is reduced substantially by the inverse of the expansion of bandwidth. Multiple spread spectrum signals can then be transmitted one on top on each other as long as the PN codes are either different or not synchronized. The signals can be separated in the receiver by using a correlator that accepts only signal energy from the selected PN sequence and despreads its spectrum. For that to work, the despreading must be perfectly synchronized with the incoming PN sequence, locking onto the correct start bit and maintaining tight timing during reception. At the same time the desired information sequence is reproduced in the receiver, any undesired spread spectrum carriers appear like noise at the output of the PN despreader. Capacity of the RF channel is determined by how much the noise floor from the extra spread spectrum carriers can increase and still provide adequate link margin. That depends on the bandwidth ratio and the relative power levels of the different CDMA signals sharing the same frequency bandwidth.

Processing of a digital bit stream using the spread spectrum technique can make the carrier more tolerant of RFI. Spread spectrum, which could be classified as either a coding technique or a modulation technique, provides two important benefits for commercial satellite communications. First, the high-speed PN code spreads the carrier over a relatively wide bandwidth, reducing the potential of RFI caused by the Earth station transmission. That is important in siting the Earth station and in dealing with uplink interference to closely spaced satellites. Second, the recovery process at the receiving Earth station can suppress almost all external
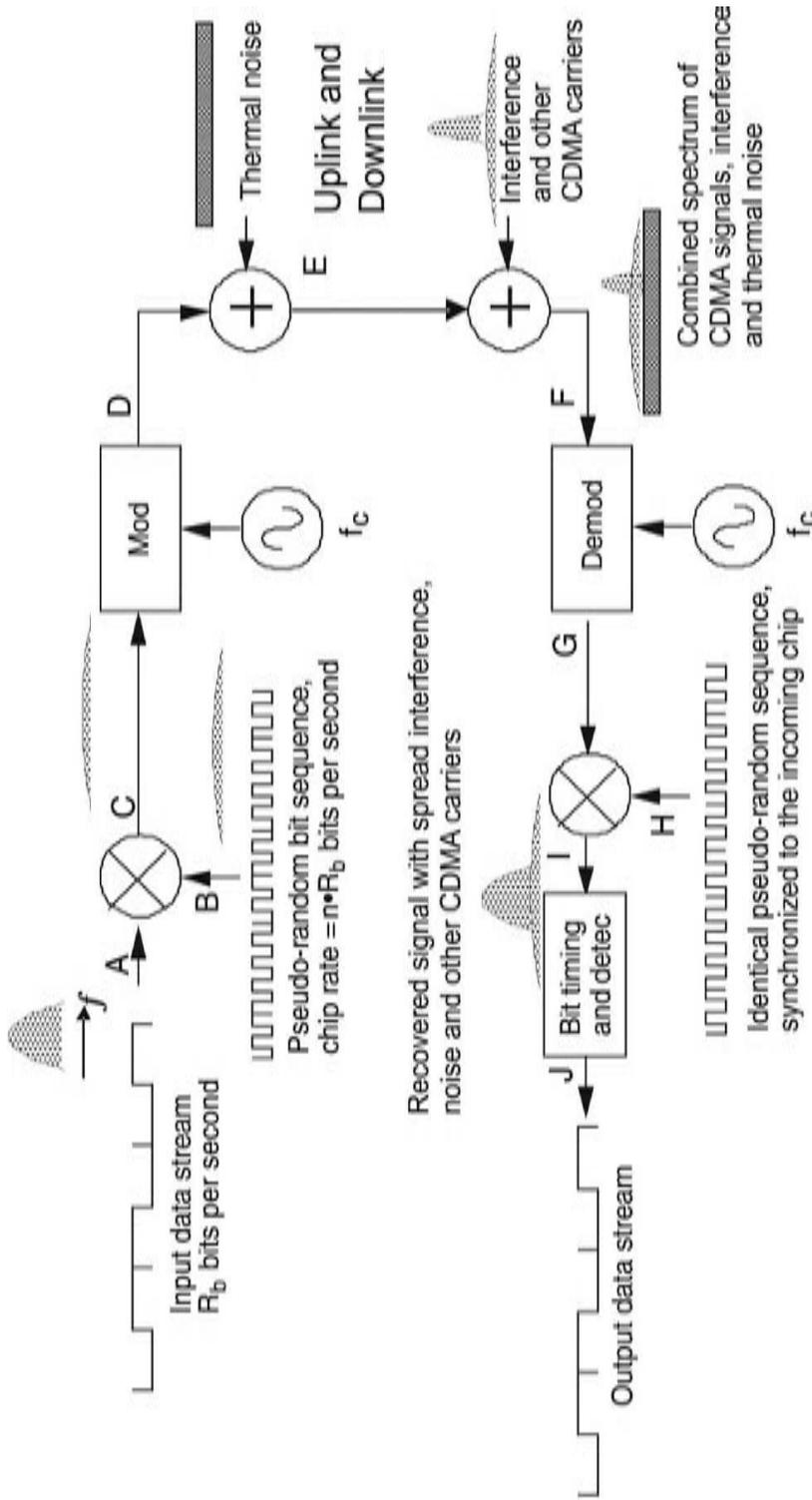
Figure 5.24 CDMA system block diagram, indicating how interference and other CDMA signals are suppressed at the receiver.

RFI located in the RF bandwidth of the spread carrier. The important data are obtained by despreading the carrier, but that causes any nonsynchronized RF signal to undergo spreading after mixing in of the PN code at the receive end. Modern CDMA receivers also include the RAKE function, so called because it can despread the direct carrier as well as other delayed copies of itself that arrive through multipath propagation.

The first commercial satellite communication application of CDMA on record was by Equatorial Communications, formerly located in Mountain View, California. Equatorial developed a one-way data broadcasting system to provide various kinds of financial market information and wire services with conventional C-band transponders transmitting to small antennas only about 40 cm in diameter. Spread spectrum modulation was chosen, not to allow multiple transmissions, but rather to overcome terrestrial interference, which is rather pronounced in the United States at that frequency. One transponder supported two to four of those carriers, each at a bandwidth of about 1 MHz. The information rate was rather low, being only 19.2 kbps for each carrier. However, the service was found to be very effective at the time since doing the same job over the analog terrestrial telephone network was much more costly and less reliable.

More recently, CDMA has gained acclaim due to the pioneering efforts of Qualcomm, another California-based company, this time in San Diego. Qualcomm successfully introduced two-way communication using CDMA in its highly acclaimed IS-95 digital cellular telephone standard. The same technique was applied to the Globalstar LEO satellite network, using a bandwidth of approximately 1.2 MHz. Unlike Equatorial's data broadcasting system, both Globalstar and IS-95 presume the existence of hundreds of users sharing the same channel at the same time. Furthermore, the technique provides improved frequency reuse, because cells that operate on the same frequency can be placed much closer to each other than is possible with either TDMA or FDMA. These benefits have motivated the ITU and others to adopt CDMA for the 3G broadband cellular systems.

5.3.5 RF Bandwidth Utilization in Multiple Access

Comparison of multiple access methods is a complex and often daunting task. To be true to this type of evaluation, one must consider a wide range of variables, including technical properties, cost, risk, and operational convenience. However, in this instance, we consider only the general issue of bandwidth utilization in the satellite transponder. The manner in which the three generic multiple access methods use that bandwidth is illustrated in Figure 5.25. Across the top are shown eight transponder channels, with the three in the center expanded to display occupancy. The channel bandwidth and spacing most likely will differ by frequency band (36 MHz at C-band, 27 MHz at Ku-band, and so forth) and also may be arranged differently for a specific satellite design. A detailed description of the design of the repeater and transponders is covered in Chapter 6. We can assume that, for the purpose of this example, the channel bandwidth and the spacing are 36 MHz and 40 MHz, respectively, typical of C-band domestic satellites operating in the North American and East Asian arcs.
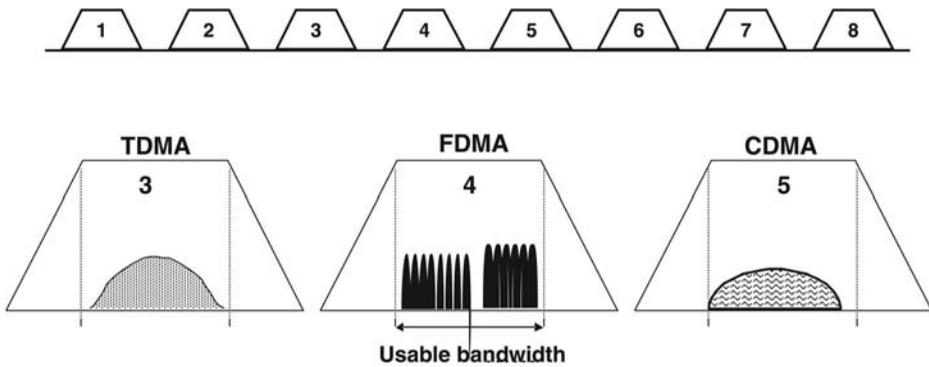
Figure 5.25   Sharing the transponder resource through TDMA, FDMA, and CDMA.

The example of a transponder used for FDMA shows two groups of RF carriers, each of equal power and bandwidth. Each carrier would come from a different Earth station and could contain multiple information channels in either analog or digital form. There is no reason to maintain constant bandwidth and power, and most FDMA transponders contain a variety of carrier bandwidths and power levels. One important constraint is that total power of these carriers cannot exceed the available transponder power, minus a backoff to account for RF IMD (discussed later in this chapter). Between each carrier is a small guardband to allow the Earth stations to separate the carriers effectively. The satellite downlink contains all the carriers, and any particular one could be selected for reception at any Earth station located in the coverage beam.

In the case of full-transponder TDMA, all stations transmit bursts in the uplink, producing equal power as received at the satellite. A single wideband carrier transmitted by one station exists in the transponder at any instant in time. There would be breif blank periods of time (not visible in a static spectrum display) between station transmissions because full utilization is nearly impossible to achieve in practice. The guardtime intervals are needed to prevent interburst interference from the overlap of transmissions. (The criteria that determine the shape of the digital carrier were covered earlier in this chapter.) It also is possible to use TDMA with a lower data rate that uses less RF bandwidth to permit the sharing of a transponder on an FDMA basis. Networks of inexpensive Earth stations with limited uplink power can utilize narrowband TDMA because the power required is proportional to the bandwidth of the carrier, all other things being equal.

The display for CDMA in Figure 5.25 indicates multiple carriers, which ride one on top of each other (which is difficult to illustrate on paper because the signals appear as one image). Individual carriers are centered on the same frequency and are separated only by their code modulation. An alternative is to divide the transponder on an FDMA basis and to have multiple CDMA channels operating at the same time. That also facilitates frequency reuse, in which the frequencies are also assigned to particular beams in a cellular-type of pattern. The same total power constraint of FDMA applies to CDMA as well.

To summarize, multiple access is the process by which several Earth stations transmit to a common satellite and thereby share its communications capacity. The

capacity and connectivity achieved depend on both the multiplexing and modulation formats of the individual transmissions as well as the multiple access method employed. In Section 5.5, we discuss how multiple access signals interact with one another and with the transmission systems used to uplink, repeat, and downlink the associated RF carriers. (Nontechnical readers may want to skip Section 5.5.)

5.4   Distortion and Impairments

Generally, impairments are the distorting effects (exclusive of noise and external interference) that a communication channel has on the information. This section reviews the impairments that involve the transmission of one carrier through the channel at a time. (Multiple carriers such as are used in FDMA are covered later.) An example of typical single-carrier impairments is presented in Table 5.4 from the standpoint of both the cause and the effect. In column 1 of the table are the impairing characteristics of the empty channel, that is, those that can be measured without the desired information present. Among those causes are variations in the amplitude versus frequency response (i.e., bandwidth limiting) and the phase versus frequency response (also called delay distortion: the rate of change of phase is equal to time delay). Also of interest are two forms of gain nonlinearity: amplitude nonlinearity (AM-to-AM distortion) and phase nonlinearity (AM-to-PM distortion). The distortion properties are illustrated for a standard C-band transponder and TWTA in Figures 5.26 and 5.27, respectively.

5.4.1   Digital Signal Impairments

Impairment of digital data modulated on a carrier with QPSK or 16 APSK typically results in an increase in the BER. While that can be compensated for by an increase in overall E_b/N_0, there can be a point where increased power is not available as a solution. Predicting those impairments and designing around them involve the use of actual measurement or computer simulation techniques. The process of using a simulation tool involves setting up a simulation model of the end-to-end system within the software. An example of such a model is illustrated in Figure 5.28 for a hypothetical three-transponder system such as would be found in a bend pipe satellite link. The desired channel is transmitted through the center transponder,

Table 5.4   Matrix of Transmission Impairments Due to the Uplink, Transponder, and Downlink, from the Standpoint of Cause and Effect on Digital Signals

Cause Within the Channel	Effect on Digital Signals
Amplitude versus frequency	I to Q coupling in N-PSK causes additional bit errors (N > 2)
Phase versus frequency	Intersymbol interference causes additional bit errors
Amplitude nonlinear distortion (AM to AM)	Spectral sideband regrowth causes adjacent transponder interference
Phase nonlinear distortion (AM to PM)	Signal constellation rotation (I to Q coupling) causes additional bit errors
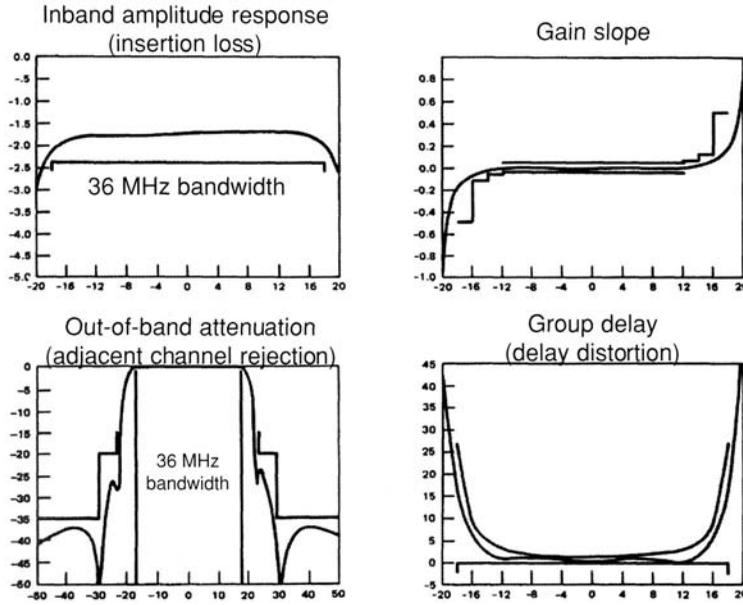
Figure 5.26   Distortion characteristics of a typical C-band transponder (IMUX).
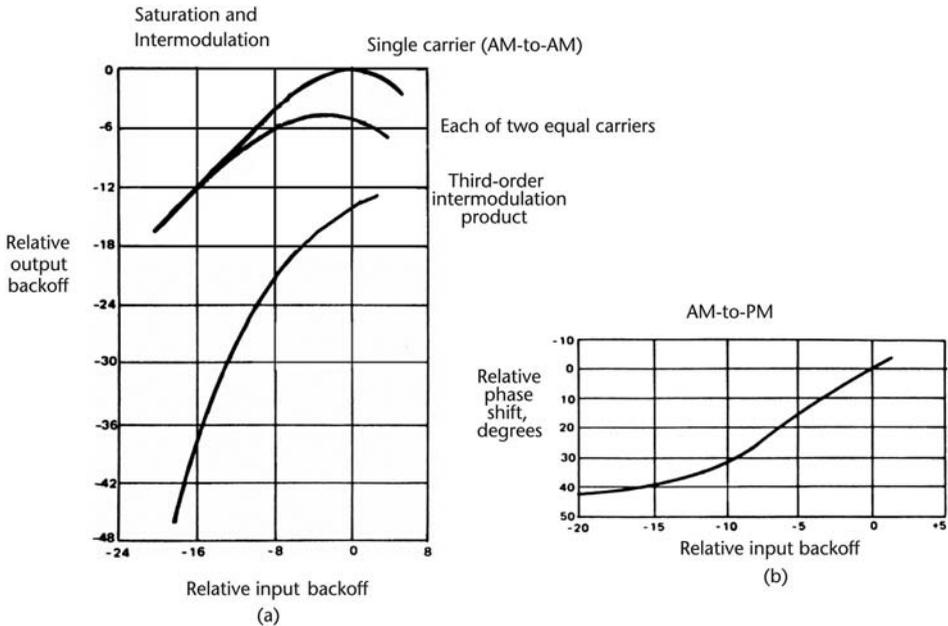
Figure 5.27   (a, b) Nonlinear characteristics of a typical space TWTA (C-band).

and potentially interfering channels exist on either side in the adjacent transponders. That allows engineers to evaluate many different types of signals, bandwidths, power levels, and transponder characteristics, all without having to build and test any real hardware. Done correctly, the simulation can get to within 0.25 dB of the overall impairment performance to be expected in a real system.
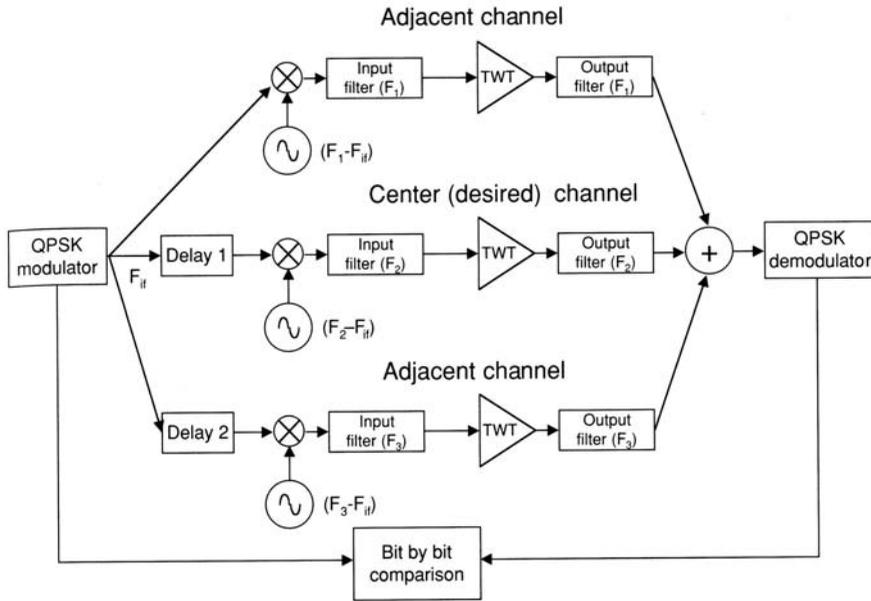
Figure 5.28   Typical computer simulation block diagram for a three-channel QPSK system. Only one modulator is needed as Delay 1 and Delay 2 produce noncoherent adjacent carriers interference.

The preceding discussion is meant to give the reader an idea of the types of factors that are taken into consideration when a channel is analyzed in detail on a single-carrier basis. Most of the impairments, while important, affect the performance to a lesser degree than power levels and antenna gains, which are taken into direct account in a link budget. However, there are many situations when detailed transponder characteristics can have a meaningful impact on the system, particularly if the designer needs to squeeze the last 0.5 dB out of the design.

The active nonlinearities (AM to AM and AM to PM) of satellite power amplifiers (shown in Figure 5.27) tend to produce unwanted interference to the output carriers by the process of RF IMD, which appears as unwanted interfering carriers. A single unmodulated carrier with a constant amplitude will not cause RF IMD, no matter how nonlinear the amplifier may be. Any harmonics, which appear at integer multiples of the carrier frequency in the output of a nonlinear amplifier, are located way outside the downlink band and are easily removed by the output filters of a satellite transponder. The following paragraphs review the considerations of multiple-carrier RF IMD in FDMA and single-carrier RF IMD for TDMA.

5.4.2   Transponder Intermodulation Impairment

The following is a brief discussion of transponder RF impairments that produce additional forms of interference noise that need to be accounted for in the link budget.

5.4.2.1 Multiple-Carrier RF IMD

Multiple-carrier RF IMD (abbreviated IMD for the remainder of this section) results from FDMA operation of a common transponder amplifier. The distortion actually looks like additional RFI in the same transponder and can overlap into adjacent transponders as well. The most basic type of IMD comes from two or more carriers that share the same nonlinear amplifier. To see how two unmodulated carriers produce the amplitude variations needed to generate IMD, examine the time waveforms in Figure 5.29. Illustrated at the top of the figure are a sinewave of amplitude 1.0 at frequency F_1 and another sinewave of one-quarter that amplitude at a slightly lower frequency, F_2. The algebraic sum of the two waveforms added point by point is shown in Figure 5.29(b). It turns out that the AM is periodic with a frequency equal to the difference $(F_2 - F_1)$.

   The gain nonlinearity of the amplifier responds both to the degree of the AM and to its frequency. If the nonlinear characteristics and input signal parameters are known, the IMD can be predicted with accuracy sufficient to design the link. The input spectrum composed of the two carriers is shown in Figure 5.29(c). The IMD products at the amplifier output occur at frequencies that are arithmetic sums and differences of the input carrier frequencies. For example, the two carriers will generate the strongest intermodulation products, located at frequencies $(2F_1 - F_2)$ and $(2F_2 - F_1)$, illustrated in Figure 5.29(d). The power level of those products will be determined by the relative input power of each carrier into the amplifier, the total output power relative to the maximum capability of the amplifier (the operating point), and the particular nonlinear characteristic of the amplifier. With two equal amplitude carriers operating at the maximum power point of a typical TWTA (the point of saturation), each product will be approximately 10 dB down from saturation. Figure 5.27(a) contains a plot of TWT IMD performance, where the level of the IMD product is shown in the lower curve. At −10 dB, the interference

(A) Time waveforms overlayed for two unequal carriers

(B) Time waveform of linear sum showing amplitude modulation

(C) Frequency spectrum of two unequal carriers

(D) Frequency spectrum of amplifier output showing intermodulation products
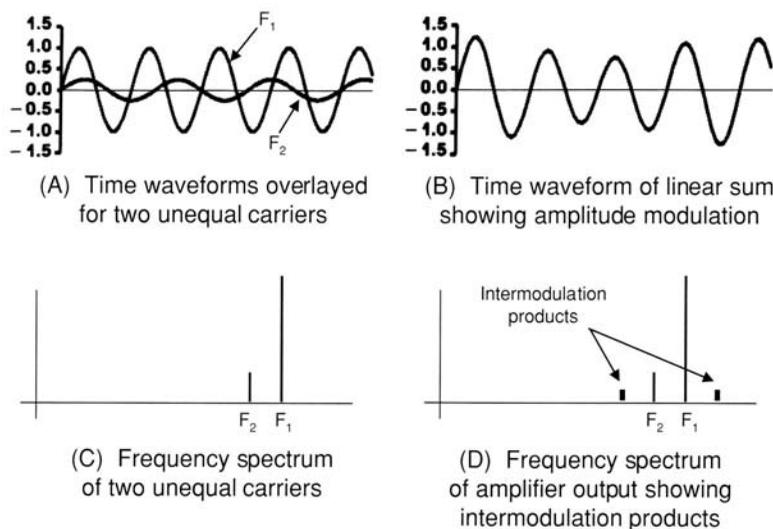
Figure 5.29   (a–d) AM from two carriers results in intermodulation products in the output of a nonlinear amplifier.

is unacceptable for most transmission modes, although the products do not fall directly on top of the carriers in that example. However, as the total input power is reduced (i.e., as the input is backed off), the level of the products also decreases but at a somewhat greater rate. As illustrated in Figure 5.27(a), a total input power backoff of 10 dB results in an IMD product that is approximately 20 dB below the carrier level. The fact that downlink carrier power drops at the same time means that an optimum point of link operation is reached where the sum of link noise and IMD is minimized (illustrated in Figure 5.19).

The principle of how nonlinearity acts on FDMA is shown at the left of Figure 5.30. The FDMA system illustrated occupies a single transponder and consists of four equal amplitude carriers with rectangular-shaped modulation each transmitted from a different Earth station. The amplitude and phase nonlinearity of the satellite power amplifier respond to the rapid amplitude fluctuations of the composite signal. The output spectrum, shown at the lower left of the figure, contains IMD products in the form of small carriers with triangular-shaped modulation. Each such product is the result of the interaction of combinations of either two or three of the input carriers, according to the following algebraic relationships:

$$F_{(2)} = 2F_a - F_b$$
$$F_{(3)} = F_a + F_b - F_c$$

where F_a, F_b, and F_c are the frequencies of sets of three different carriers in the input.

The combinations on the right side of the equations are frequencies in the same band as the carriers themselves and can overlap useful information. The degree of overlap depends on the number of carriers and their placement in the RF channel, as governed by the formulas. It is important to note that the modulation on the IMD products is a combination of the original carrier modulations; IMD products, therefore, occupy a greater bandwidth than any one carrier. With enough input carriers and wide enough modulation, the spectrum of the IMD appears as a noise
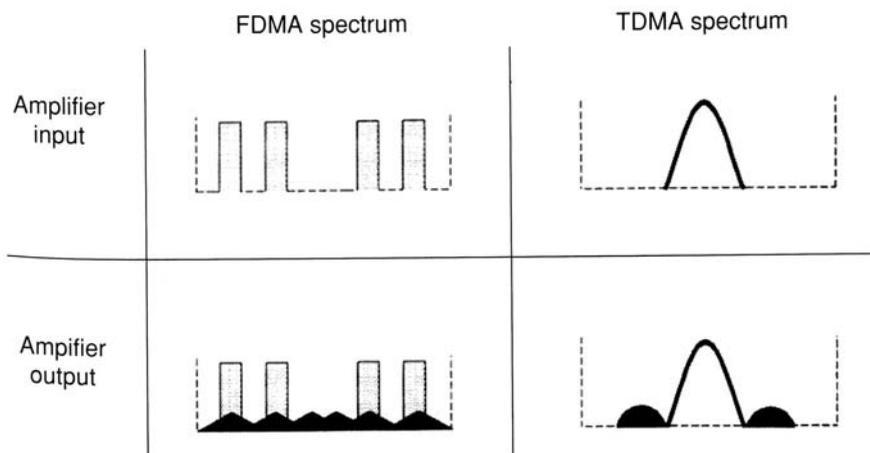
Figure 5.30 IMD products generated in a nonlinear amplifier by FDMA and TDMA.

floor almost like white noise (Figure 5.18). Most of the time, IMD can be treated like white noise over the bandwidth of the carrier in question and analyzed in that manner in the link budget calculation. However, IMD products that fall back on one of the generating carriers include the original modulation. That is only a consideration in analog FM transmission and can be ignored for BPSK and QPSK.

The particular IMD products described in the previous example are called third-order products, indicated by the sum of the absolute values of the frequency terms (i.e., $2 + |-1|$ and $1 + 1 + |-1|$, both of which are equal to 3). Fifth-order terms, such as $3F_a - 2F_b$, also fall in band and would need to be included in the analysis since they can add a significant although lesser amount of interference. Generally speaking, it is the odd-numbered orders of IMD that fall in band; as the order increases, their impact on performance decreases.

5.4.2.2   Single-Carrier IMD

A single carrier operating at or near saturation usually is not expected to experience detrimental effects from IMD. However, it will be shown that digital transmission using QPSK often is subject to amplifier nonlinearity, particularly when pulse shaping such as that described previously in this chapter is employed. A PSK carrier (or either the I or Q components of a QPSK carrier) has a reversal of phase at the transition from one binary state to the other. Rectangular pulses would cause the modulated carrier to always maintain a constant amplitude, which eliminates the possibility of either AM-to-AM or AM-to-PM conversion in a nonlinear amplifier. As explained previously, rectangular pulses have infinite bandwidth of the form shown at the bottom of Figure 5.7.

The input and output spectra for a TDMA system using PSK with raised cosine spectrum pulse shaping is shown at the right in Figure 5.30. At the top, the ideal raised cosine shaping is evident and no sidelobes are present. However, after passing through the nonlinear amplifier, the sidelobes reappear, having been regenerated. Another way of viewing this is that the saturation of the amplifier flattens the tops of the AM pulses and the resulting resquared pulse modulation has an infinite spectrum of the form $\sin(x)/x$.

The new sidelobes of the digital carrier defeat part of the purpose of shaping the pulses in the first place. Sidelobes are approximately 14 dB down from the carrier and can cause interference to carriers operating in adjacent transponders. By backing off the amplifier input by 1 dB or more, it is possible to reduce greatly the sidelobe level and make adjacent transponders more compatible with one another. Once again, there is a tradeoff in total link performance, since backing off the input reduces the downlink power. Fortunately, the appropriate amount of backoff may reduce the downlink only slightly.

Sideband regeneration can be prevented by using a modulation waveform that does not contain wide amplitude swings associated with filtered QPSK. An early approach was to stagger the Q channel by one half of a bit period with respect to the I channel. As a result, the swing is reduced because a complete 180-degree phase reversal cannot occur. This technique is called offset QPSK (OQPSK) because of the relationship between the I and Q channels. Maintaining a more stable envelope for the carrier is the goal of another method called minimum shift keying

(MSK) that also provides two bits per modulation symbol. The envelope is smoothed out using a filtering transform ahead of the QPSK modulator. One drawback is that while there is no sideband regeneration, the nominal bandwidth of an MSK signal is approximately 1.5 times that of a filtered QPSK or OPSK signal. A reduction in the MSK bandwidth is obtained by transforming the input using a filter with the shape of a Gaussian characteristic. The term used for this technique is Gaussian MSK (GMSK); it has been applied in some VSAT equipment and more broadly within the GSM standard. GMSK is preferred in systems with amplifier nonlinearity because it has both a narrow bandwidth and resistance to sideband regeneration. Its wider adoption is only limited by the complexity of the modulating waveform itself.

When several narrowband TDMA carriers are operated in the same transponder, sideband regeneration is nearly insignificant; hence, the carriers can be packed tightly together in frequency. Multicarrier IMD would have to be carefully considered in setting the operating point of the transponder, since several carriers will be transmitting at the same time. That is the case for VSAT networks using the various TDMA and FDMA modes, demanding that the operation of the transponder be carefully examined before full-scale operation is commenced.

5.4.3 Uplink and Downlink RF Interference

Electromagnetic compatibility (EMC) is often a complex and daunting topic in the design and operation of satellite communication systems. EMC problems may not even exist in some systems, while in others it is an overriding area of concern and may even prevent proper operation. Sources of EMC problems include various forms of self-interference that arise in the Earth station or the satellite on its own. Those types of problems can be prevented by proper design and testing of the applicable component, subsystem, or system before it is put into operation. What we are addressing in this section is RFI that comes from sources external to the satellite or Earth station, usually from other operating transmitters that use the same frequency band.

A variety of interference sources (called entries) can invade a satellite network. In addition, the operation of a satellite or Earth station can interfere either with another satellite network or with a terrestrial microwave network that shares the same band. We are not considering out-of-band RFI, which is produced by undesirable radiation from antennas not within the intended band of operation. We concern ourselves here with what can be called space system interference, which is RFI that is coupled between two different satellite networks. It involves both uplink and downlink interference, which are defined in Figure 5.31(a, b), respectively.

Uplink interference is the undesired radiation from an Earth station into an adjacent satellite that happens to cover the same portion of the Earth's surface on the same frequency. The undesired radiation is produced by the sidelobe pattern of the interfering Earth station's transmitting antenna. In FSS and BSS applications, the large uplink antenna typically radiates at a relatively low power level from its sidelobes; hence, uplink interference potentially is controllable through the process of frequency coordination. That might impose a limit on the transmit power applied

A. Uplink interference

B. Downlink interference
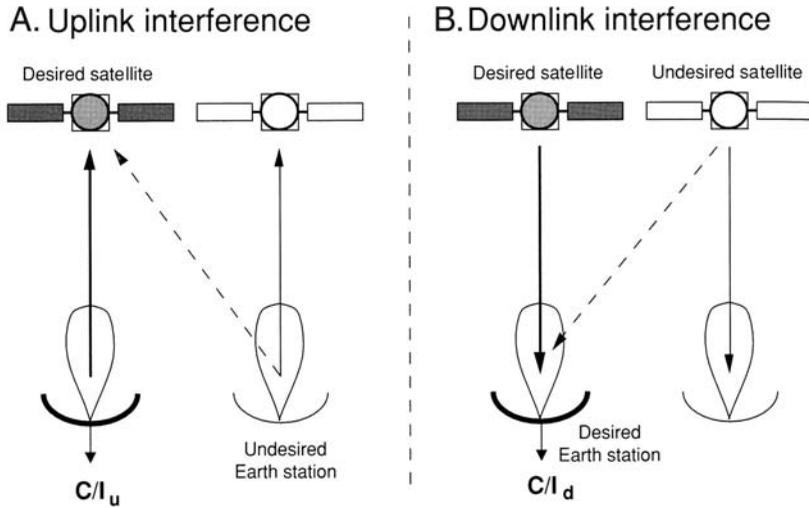
Figure 5.31 Simplified (a) uplink and (b) downlink interference geometry.

to the antenna or on the sidelobe gain itself (discussed in Chapter 4). In MSS systems, the small uplinking antenna may have nearly the same gain in both the desired and the undesired directions; hence, little can be done to reduce its effect on an adjacent satellite. The only measure that can be taken here is to ensure that the interfered-with satellite either does not receive on that particular frequency or that its spacecraft antenna has reduced gain in the direction of the interfering Earth stations.

The formulation of the uplink interference problem is based on using the sidelobe pattern of the uplink Earth station along with the footprints of the desired and interfered-with satellites. To that is added the power level of the desired carrier coming from the desired uplink. The equation for the carrier-to-interference ratio (C/I) generally has the following dB form:

$$C/I_{up} = [G(0) - G'(\Theta)] + [P_t - P'_t] + [G_s - G'_s]$$

where

$G(0)$ = the antenna gain of the desired Earth station in the direction of the desired satellite (e.g., the peak gain);

$G'(\Theta)$ = the antenna gain of the interfering Earth station in the direction of the desired satellite, which is at an offset angle, Θ, which corresponds approximately to the orbit separation between the desired and the interfering satellites; generally the form of the relationship between $G'(\Theta)$ and Θ is $G'(\Theta) = 29 - 25 \log(\Theta)$, in decibels, for Θ in degrees;

P_t = the transmit power applied to the desired antenna;

P'_t = the transmit power applied to the undesired antenna;

G_s = the receive gain of the desired satellite in the direction of the desired Earth station;

G_s' = the receive gain of the desired satellite in the direction of the interfering Earth station. (The term $[G_s - G_s']$ can also be expressed in terms of the G/T values.)

That relationship assumes that the path lengths from the satellite to both the undesired Earth station and the desired Earth station are the same. If they are different, an appropriate adjustment factor should be included.

The relationship looks complicated and confusing, but it really is nothing more than logic. Simply stated, the uplink C/I, in decibels, is the relative value of the three terms on the right side of the equation: the difference in gain of the transmitting antennas plus the difference in input power levels applied to the antennas plus the difference in gain of the satellite receive antenna. Figure 5.32 provides the uplink C/I as a function of uplink antenna size. Performance at three common microwave frequency bands is plotted, assuming a 2-degree fixed orbit spacing. If both P_t and P_t' are equal, the two networks are said to be homogeneous. That means the C/I performance is based only on the relative antenna gains. Otherwise, the power levels will have a decided impact on the compatibility of the two systems.

The case of downlink interference can be evaluated in much the same manner. This time, we formulate the problem in terms of EIRP differences:

$$C/I_{dn} = [P_t + G_s] - [P_s' + G_s'] + [G(0) - G(\Theta)]$$

where

$[P_t + G_s]$ is the EIRP of the desired satellite in the direction of the desired Earth station;

$[P_s' + G_s']$ is the EIRP of the undesired satellite in the direction of the desired Earth station;
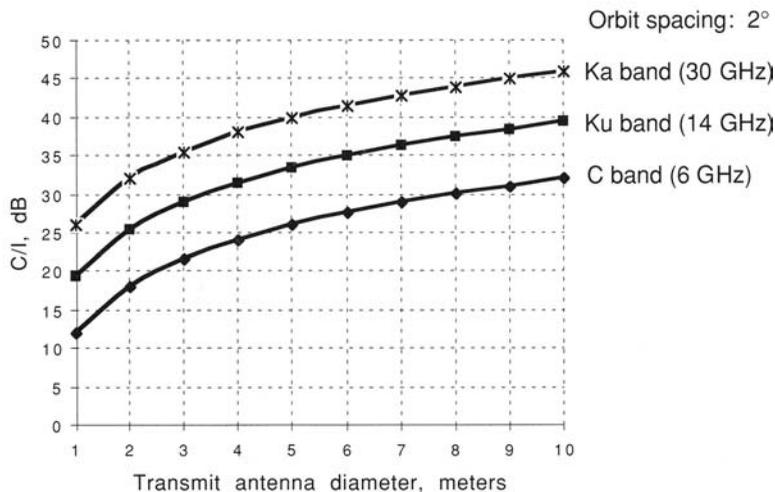
Figure 5.32   Uplink C/I from Earth stations of the same diameter, assuming standard ITU radiation characteristics and 2-degree orbit spacing.

$[G(0) - G(\Theta)]$ is isolation offered by the receiving antenna of the desired Earth station.

The downlink C/I generally is what we worry about in receive-only subscriber-based services like DTH TV. We have what should be a good signal from our satellite; however, a certain degree of interference is going to exist from other potentially interfering satellites that operate on the same frequency. An example of how the C/I is dependent on receive antenna diameter and orbit spacing is shown in Figure 5.33. The principal protection from uplink and downlink RFI is that from ground antenna sidelobe suppression. In an actual case, additional isolation could be provided by more careful consideration of the actual antenna patterns (for instance, the first null of 60-cm dishes at Ku-band is at 2°), by offsetting carrier frequencies, by using cross-polarization, or any combination of these.

There is still the question of how that interference will affect the operation of our satellite network, particularly the resulting quality of reception by the user. For many signal and interference types, the effect is like an equivalent amount of white noise. We therefore only have to include the C/I term in the link budget, combining it through the general equation of the form:

$$C/N = (N_{up}/C + N_{dn}/C + I_{up}/C + I_{dn}/C + I_{xpu}/C + I_{xpd}/C)^{-1}$$

where C/N is the total value, considering all sources of noise and interference, expressed as a ratio (and not in decibels). In the case of C/I, the interference power is measured within the bandwidth of the desired carrier power, C. That may require careful analysis if the interference is wider in bandwidth than the information signal. On the other hand, narrowband interference from either an unmodulated carrier or one with analog modulation potentially could cause greater harm than its pure power would indicate. The formula for adding in the interference on a power basis (given above) would have to be replaced with a more rigorous approach, such as computer simulation or even direct measurement in the laboratory or over an actual satellite link.
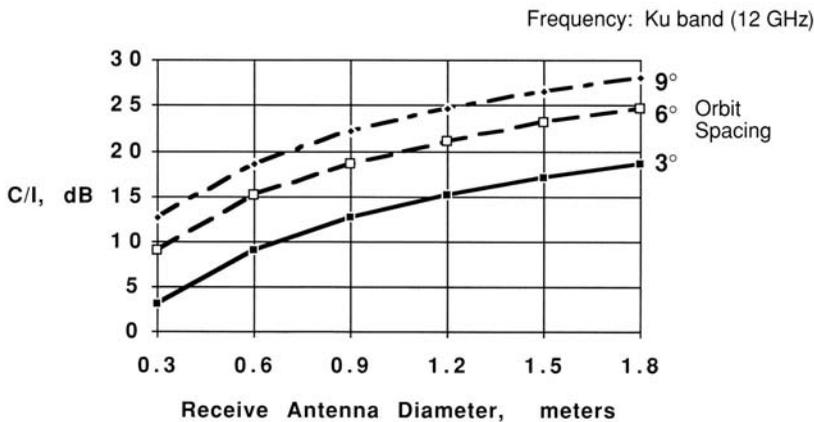
Figure 5.33 Downlink C/I versus dish diameter and orbit spacing for DTH service.

References

[1] Kondoz, A. M., *Digital Speech: Coding for Low Bit Rate Communications Systems*, New York: Wiley, 1994.

[2] Elbert, B. R., *The Satellite Communication Applications Handbook*, 2nd ed., Norwood, MA: Artech House, 2004.

[3] Michelson, A. M., and A. H. Levesque, *Error-Control Techniques for Digital Communication*, New York: Wiley, 1985.

[4] Sklar, B., *Digital Communications: Fundamentals and Applications*, 2nd ed., Upper Saddle River, NJ: Prentice-Hall, 2001.

[5] Viterbi, A. J., *CDMA: Principles of Spread Spectrum Communication*, Reading, MA: Addison-Wesley, 1995.

[6] Glasic, S., and V. Branka, *Spread Spectrum CDMA Systems for Wireless Communications*, Norwood, MA: Artech House, 1997.

Spacecraft and Repeater

The purpose of this chapter and Chapters 7 and 8 is to explain how a satellite works and to review the factors in its design. These chapters should familiarize the reader with the key concepts and terminology that are common in the satellite industry. As a first definition, a *spacecraft* is the actual piece of hardware that is launched into orbit to become an artificial *satellite* for the purpose of providing a radio repeater station.

The physical elements that a communications spacecraft comprises can be divided into two major sections: the *communications payload* (or just the payload), containing the actual radio communications equipment used for reception and transmission of radio signals, and the *spacecraft bus* (referred to as the bus), which provides the supporting vehicle to house and operate the payload. The emphasis in this chapter is on the requirements and specifications of the payload and how the general types of hardware designs can meet those requirements. Chapter 7 provides detailed information on spacecraft antennas because they probably are the most important single element of the payload regarding its communication mission. Chapter 8 covers the various approaches to the design and operation of the bus in sufficient detail to give the reader an appreciation for the subsystems that ensure safe and continued operation of the payload. We begin with a general introduction to spacecraft design.

6.1 Overview of Communications Spacecraft

The design, manufacture, and operation of a communications satellite are no simple matter, complicated by the fact that it must survive the rigors of launch and deployment in orbit, followed by many years of satisfactory operation without physical intervention by human beings. In laypersons' terms, we assume that a certified repair person cannot travel to orbit to repair or reconfigure a satellite. That is certainly the case for GEO satellites, but it is not economically feasible even for large constellations of LEO satellites. It is critical that the payload and the bus work hand in hand to establish a highly efficient radio repeater on a stable space platform. The result is shown in Figure 6.1, which illustrates a typical geostationary communications satellite serving a country or region of a continent. At an altitude of approximately 36,000 km, a beam with dimensions 3 degrees by 8 degrees would cover an area the size of the United States. From the standpoint of a typical LEO satellite, the coverage is more limited in area simply because the satellite is much closer to Earth.
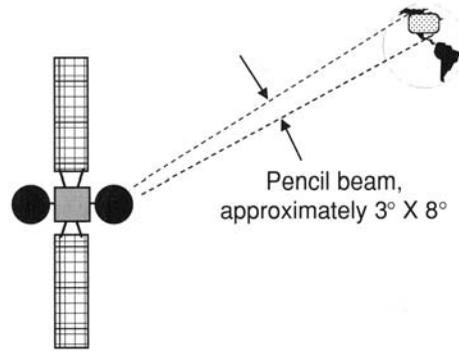
Figure 6.1   Coverage of a land area from GEO requires the generation and control of a pencil beam onboard the satellite.

Figure 6.2 indicates the general arrangement of the spinner and three-axis bus designs. The spinner employs a drum-shaped solar panel that is able to produce continuous dc power as the spacecraft rotates at approximately 60 rpm. In contrast, the three-axis spacecraft uses flat solar panes that are separately pointed at the sun while the body of the spacecraft is pointed at Earth.

All communications satellites use solar cells connected in arrays to convert sunlight into DC power. This provides prime power for all operating electrical and electronic equipment on a continuous basis. The bus section also contains one or more energy storage batteries for use during eclipse periods when the solar arrays are blocked from the sun by the earth or moon. Figure 6.2 illustrates the standard *three axis* configuration of the spacecraft that keeps the two flat-panel solar arrays
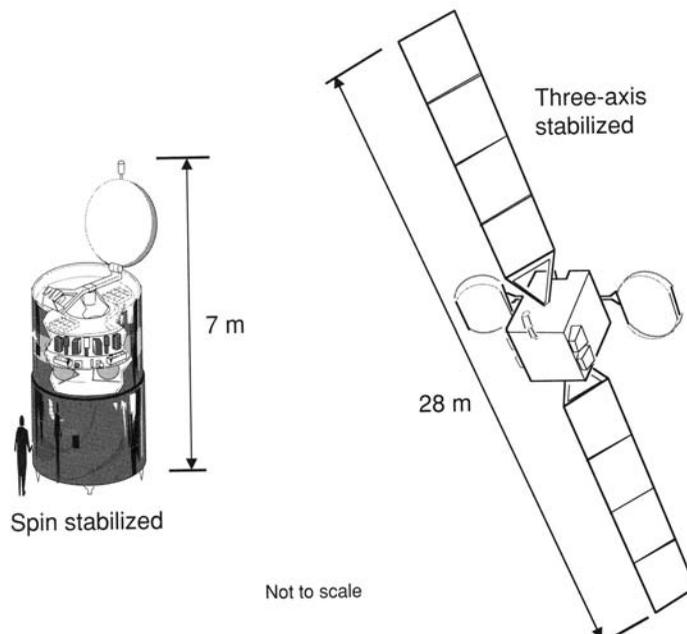
Figure 6.2   Basic satellite configurations: the spinner and the three-axis.

pointed in the general direction of the sun while the body and its antennas point at the earth. The two arrays are made from separate panels that are folded like an accordion for launch and deployed once the satellite has reached orbit.

The drum-shaped panels of the legacy spinner design are deployed like a collapsible telescope, which is less efficient in terms of area and mass. However, the spinner was elegantly simple in its attitude control and propulsion systems, but its mass did not scale up to higher powers as well as the three axis spacecraft. Consequently, spinners were limited in power to approximately 2 kW while there are three axis satellites operating with over 15 kW of power. This is the result of having more physical area on flat solar panels and the ability to continuously orient those panels toward the sun.

As will be discussed in Chapter 8, many other factors must properly be taken into account for us to have a stable and long-life communications satellite. Many of these principles are illustrated in Figure 6.3, which is the conceptual framework for overall satellite design. Obviously, the repeater will not operate on its own and needs the supporting functions illustrated around it. For a satellite in geostationary orbit, eclipse occurs during a one-month period centered on the equinoxes, lasting up to 70 minutes per day. Requirements for eclipse operation increase in frequency for MEO and LEO missions because the angle subtended by the Earth increases as orbit altitude decreases. Such eclipses occur multiple times per day, consistent with the orbit period. Without batteries, a spacecraft would shut down entirely during eclipse and might not activate properly after reentering sunlight.

A satellite or a constellation of satellites is operated from the ground through a bidirectional radio communications link dedicated to that purpose. Spacecraft control personnel can monitor the health and the status of all onboard subsystems using the *telemetry* downlink and can control active devices onboard the satellite (e.g., to configure the payload and to activate and disable bus functions) using the *command* uplink. In addition, the position and the orientation of the satellite are determined by measurements of distance and angle that are made with *tracking* signals sent over the telemetry and command (T&C) links. Following the launch
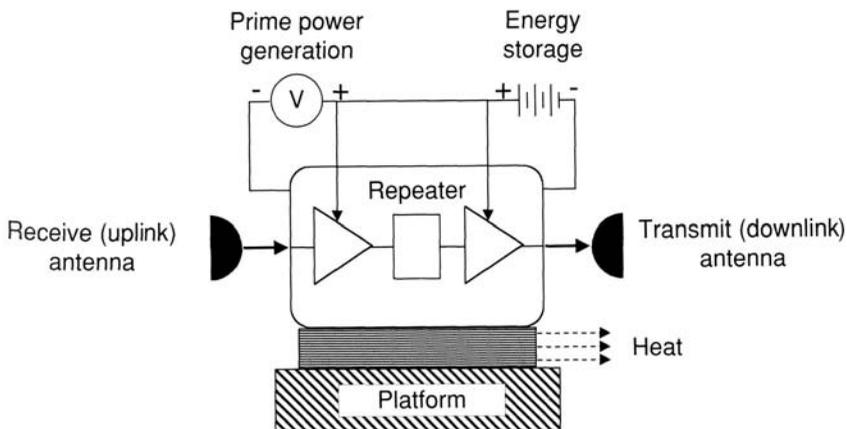
Figure 6.3 Functional diagram of an independent microwave repeater station for use in space.

sequence, the T&C link provides the only means to locate the satellite in its initial orbit, transfer it to its final operating orbit, and configure it for service.

The repeater generates heat, much like an operating TV set or microwave oven. If not removed from the electronics, the heat will cause the temperature to rise sufficiently to affect or even damage the sensitive parts of the spacecraft, particularly the repeater electronics. The visible and invisible radiation from the sun also tends to raise the temperature of the spacecraft, while undesirable low temperatures can occur during eclipse and on darkened sides of the spacecraft at other times. Therefore, the flow of heat within from and to the spacecraft must be monitored and controlled. Another function of the spacecraft is to provide a stable plateform throughout the life of the satellite, including sustaining the rigors of launch. Precise pointing of the coverage beam or beams, critical to the purpose of the satellite, depends on the stability of the attitude control system, which is integral to the spacecraft bus. The spacecraft also has small rocket engines as part of a propulsion system that is used to reach and maintain the proper orbit.

6.1.1   Overall Payload Requirements

The requirements for the payload effectively dictate the requirements for the satellite as well. For that reason, anyone who specifies the key aspects of the communication mission must be familiar with the overall performance of the satellite throughout its operating lifetime. We typically break the requirements into those for the repeater and those for the antenna system. Both depend heavily on the types of communication services that the satellite will provide. That is obvious because the payload is often the central element in the overall communication network.

We typically start with a basic description of the overall system of satellites and Earth stations, along with the type of multiple access to be used. Then we move to the subject of the link budget, where the key performance parameters of all space and ground segments must be specified. It is the responsibility of the satellite system engineer to allocate those requirements to the elements, most notably to the payload. The initial discussion centers on the simplified payload block diagram shown in Figure 6.4, which is for a classical bent-pipe repeater of up to about 16 transponders per polarization or beam. This bent-pipe repeater, which relays the uplink carriers to the downlink with only a change in frequency and power, handles several broadband channels and includes extra equipment to restore onboard equipment failures. Several important steps are taken in its design because of the need to minimize weight and power and to operate the repeater for 15 years or more without hands-on maintenance (corrective action by ground personnel is still possible over the T&C link). In Figure 6.4, all uplink signals are first amplified and then translated in frequency from the uplink band to the downlink band within a wideband receiver. Typical receivers have an operating bandwidth of 500 MHz, which is sufficient to handle the 12 channels of 36 MHz each in this example (see Figure 6.8).

Communications satellites often reuse frequencies to increase the total effective spectrum by a multiple of two or more. Satellites that offer broad coverage in a single beam may double the effective spectrum by using two independent polarizations: either vertical and horizontal linear polarization, or left-hand and right-hand
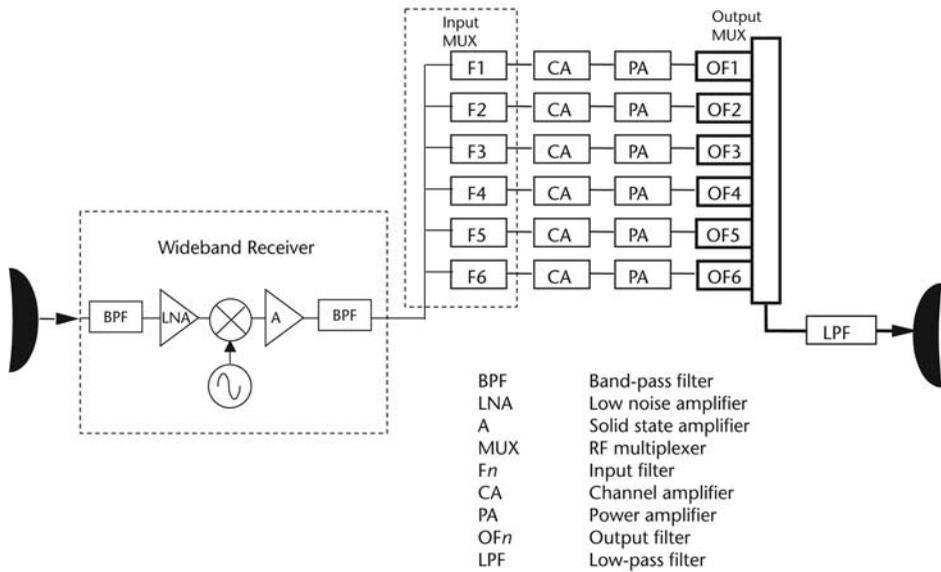
Figure 6.4 Simplified block diagram of a bent-pipe repeater using standard transponder band-widths.

circular polarization. A greater multiple reuse of frequencies is possible by breaking the coverage area into smaller cells using spot beams, a technique used effectively in terrestrial cellular telephone networks (discussed in Section 6.3.1).

Following the receiver (which provides the 500-MHz spectrum translated to the downlink band) is the input multiplexer (IMUX), individual power amplifier (PAs) per transponder, and the output multiplexer (OMUX). The complete set of amplified downlink transponder channels is then applied to the transmitting antenna of the payload. Other components and features are discussed later in this chapter.

There are many tradeoffs in satellite system design, and the requirements for the payload are a direct result. An incomplete or invalid tradeoff analysis will produce payload requirements that pose problems of one sort or another. For example, an imbalance in the allocation of antenna gain between the satellite and the Earth station might mean that an overly large antenna structure might have to be deployed in space or that the power to be transmitted could be unrealizable with practical equipment. Another trade-off in multibeam payloads is whether to perform all processing on the ground or to demodulate all uplink signals onboard the satellite. Resolving these dilemmas is well worth the time and effort of a thorough tradeoff study.

6.1.2 Transmit Effective Isotropic Radiated Power (EIRP)

At the highest level, the most important overall technical parameter of the payload is the transmit EIRP, which is the product of the antenna gain and the maximum RF power per transponder applied to it by the repeater. As illustrated in Figures 6.5 and 6.6 for typical GEO satellites, the EIRP specifies how the satellite serves
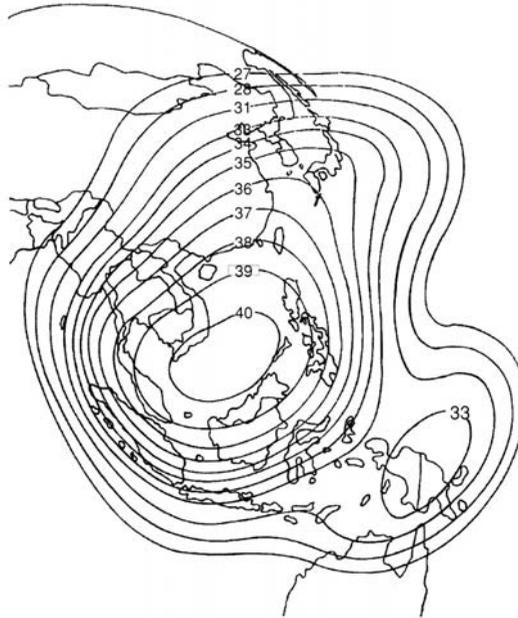
Figure 6.5 Typical C-band EIRP performance, based on the MEASAT 1 satellite. (*Source:* MEASAT Satellite Systems Sdn. Bhd.)

the coverage area on the ground. (Alternatively, EIRP also indicates the radiated power for auxiliary services like T&C and intersatellite links used to connect satellites directly to each other in space.) It is industry practice to specify the minimum acceptable value over the coverage region, which is indicated by an outermost contour or set of definition points connected by straight lines.

Power amplification, illustrated to the right in Figure 6.4 for each transponder, is important to payload operation because of the long path length to the ground station. TWTAs are the mainstay, particularly at higher frequencies (above 10 GHz) and at higher power levels (50W and higher). Microwave power transistors are used in SSPAs at the lower power levels and for applications requiring improved linearity. (Typical power levels of TWTAs and SSPAs are reviewed in Chapter 4.) The particular power amplifier must be operable for essentially the entire life of the satellite; therefore, only the most reliable devices can be selected (hence the preference for TWTAs and proven SSPAs). Also, the amplifier design must be lightweight, typically weighing under 3 kg, and must be capable of converting dc power into RF with efficiencies of 50–80%.

From electromagnetic principles, we know that the antenna performance determines the shaping of the EIRP coverage. The majority of satellites that have been launched to date have fixed antenna beam patterns, which can be altered in space only by hardware switches. That has meant the buyer of the satellite must know and adequately specify the coverage requirements with precision long before services are offered. Some satellites have had provision for changing the transmit power, through switching and paralleling of amplifier stages, which causes the overall level of the EIRP to change accordingly. Changing of beam shaping or moving entire beams can be accomplished if the needed flexibility is included in the payload.
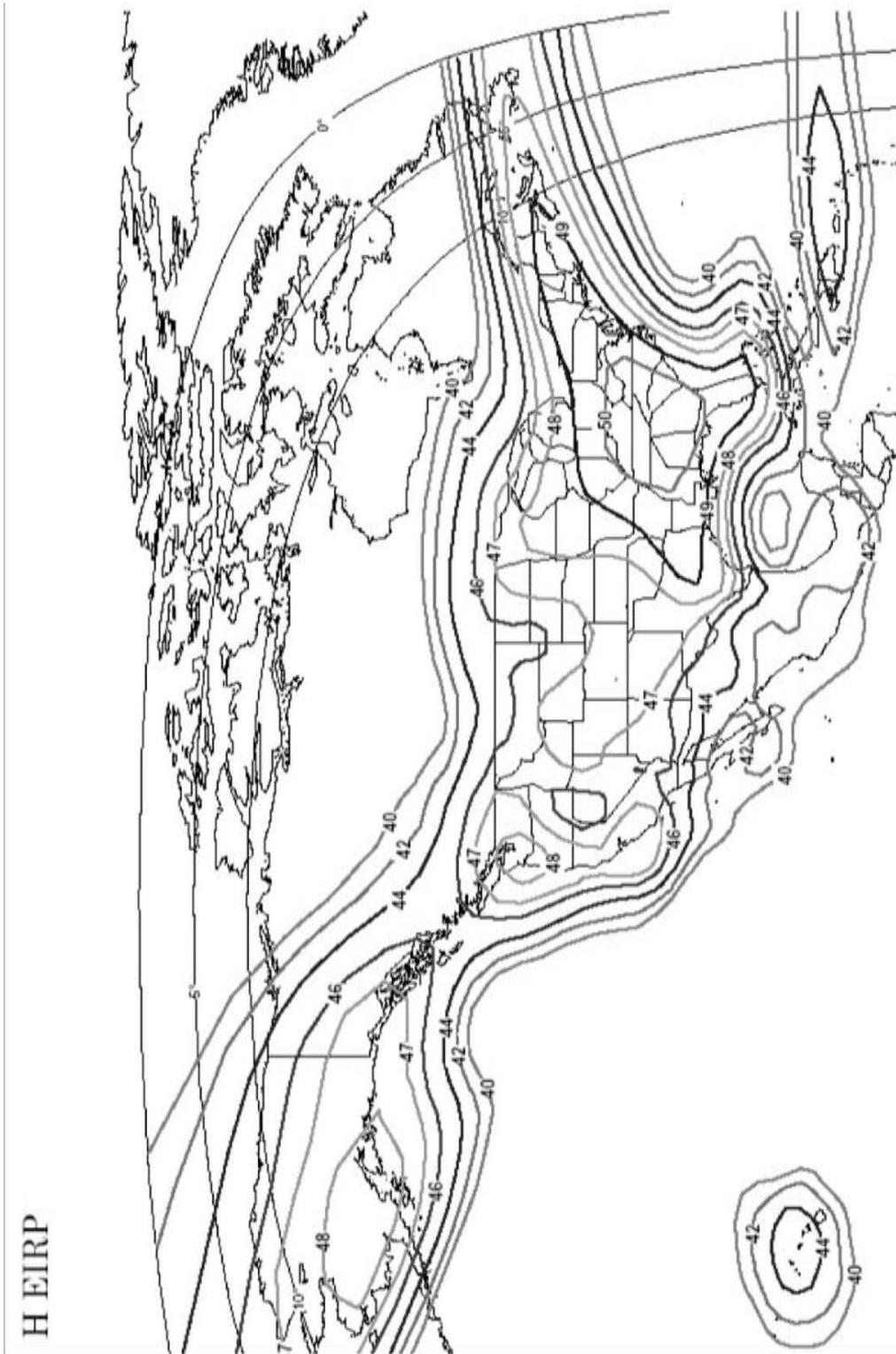
Figure 6.6 Ku-band EIRP coverage pattern for Horizons 1 at 127° WL (*Source:* Courtesy of JSAT International.)

Techniques such as movable reflectors, switched feeds, and electronically steered phased arrays are required to allow beam reconfiguration and relocation. These topics are considered later in this chapter and further in Chapter 7.

Due to the availability of larger bus designs and increasing market demand, satellites have become more complicated and sophisticated in the way they provide channels and coverage. Capabilities such as multiple frequency bands (e.g., hybrid satellites), multiple antenna systems to service several countries or regions at the same time, and channel switching are available. In Figure 6.6, we see the various EIRP patterns of the Horizons 1 satellite, which covers all 50 of the United States. The antenna system that provides that coverage is indicated in Figure 6.7, showing four reflectors that are used for transmit and receive in both C- and Ku-bands. In this particular satellite, the C-band payload is referred to as Galaxy 13.

The actual power delivered to the transmit antenna will experience some RF loss along the waveguide path between the power amplifier and the antenna feed system. The loss comes directly out of the EIRP and therefore must be minimized. It makes no sense to design a satellite for very high transmit power and then to throw a significant fraction of the power away in the waveguide components that happen to be positioned after the final amplifier stage. For example, even 0.4 dB extra loss would require 10% increase in raw dc power to provide the necessary compensation (assuming the EIRP requirement is not reduced by a like amount). On a large three-axis satellite in the 10-kW class, this could demand another 1 kW of power from the solar array and batteries. Readers should be aware, however, that reducing RF output loss by 0.4 dB may, in and of itself, be a considerable challenge, thereby making the design tradeoff more difficult on an overall basis.

Figure 6.7   General arrangement of the Galaxy 13/Horizons 1 satellite AsiaSat 3S. (*Source:* Courtesy of JSAT International, Inc.)

As a general trend, peak EIRP values for C-band and Ku-band satellites generally are around 40 dBW and 50 dBW, respectively. That implies that Ku-band payloads are more demanding in terms of antenna gain and transmit power than those for C-band. Higher EIRP also tends to increase the demand for prime dc power from the bus and makes the thermal design of the spacecraft more complex. That may explain why Ku-band satellite users tend to pay higher prices for service as well. On the other hand, the higher EIRP of Ku-band affords smaller receive antennas (which benefits the ground network) and also would be needed to overcome rain attenuation in tropical regions.

6.1.3 Receive Gain-to-Noise Temperature Ratio (G/T)

While EIRP defines downlink performance, it is the complementary parameter of G/T that is vital to the uplink operation of the satellite network. This parameter is nothing more than the result of dividing the receive antenna gain as a true ratio by the noise temperature of the satellite receiving system, T_{sys}. The latter can be calculated as follows:

$$T_{sys} = \frac{T_a}{L} + \left(1 - \frac{1}{L}\right) \cdot 270 + T_{re}$$

where

T_a is the spacecraft antenna temperature, which typically corresponds to the Earth brightness temperature of 270K;

L is the input waveguide loss factor, for example, the ratio of input ot output power (which, being greater than 1, is the inverse of the transmission line gain);

T_{re} is the equivalent noise temperature of the payload receiving system and is a specification for the payload wideband receiver itself.

The noise performance of the receiving system, like the transmit power, is under the control of the payload designer. Using the techniques of design analysis and optimization, the payload designer develops a gain and noise budget for the front end of the payload (e.g., those aspects of the payload that determine T_{sys}). From that point, subsystem and unit designers make hardware selections supporting that requirements, which in turn, produces the desired value of G/T across the coverage area. We will delve into some of those details in subsequent sections of this chapter and in following chapters.

The overall G/T characteristics of the payload after being converted to decibels per Kelvin are specified in the form of contours, similar in style to Figure 6.6 for EIRP. In the simplest area coverage payload, the EIRP and G/T requirements have precisely the same shape, differing only in the values attached to the various contours. The uplink Earth stations in these systems usually have large-diameter antennas and powerful transmitters, so G/T may not be critical to the application. However, satellite systems that support low-powered MSS hand-held user terminals and networks of low-cost VSATs are extremely challenging in terms of satellite

G/T. Another satellite receive (uplink) specification is saturation flux density (SFD), which indicates the required uplink power flux density at the satellite to produce maximum (saturation) downlink power.

6.1.4   Bent-Pipe Transponder Filtering

The first GEO satellite to be launched, Syncom, has a single wideband receiver and offered one transponder. In that way, any signal could be uplinked and subsequently received on the ground, and the satellite did not constrain the RF bandwidth in any significant way. A substantial fraction of all communication payloads (GEO and non-GEO) use a transponderized design of one form or another. The basic principle is to subdivide the total RF bandwidth into RF channels that the payload can then route and amplify on a single carrier or multicarrier basis. That can be accomplished using analog filtering at microwave or IF frequencies or through digital processing onboard the satellite (to be discussed later in this chapter). Our simplified bent-pipe repeater in Figure 6.4 uses analog filtering directly at the downlink microwave frequency. Channels of this type are several tens of megahertz in bandwidth each, to be consistent with wideband transmissions using FDMA, TDMA, and CDMA.

The term *multiplexer* in the context of the satellite repeater refers to any set of microwave filter hardware used to separate or to combine the RF channels, as distinct from the time division multiplexer (TDM) used to assemble a baseband at the transmitting Earth station. A detailed C-band repeater block diagram is provided in Figure 6.8 for reference. Draw your attention in Figure 6.8 to the series of circles (circulators) and attached boxes (input filters) composing the IMUX. The circulator is a ferrite device used to create a microwave revolving door that causes any signal
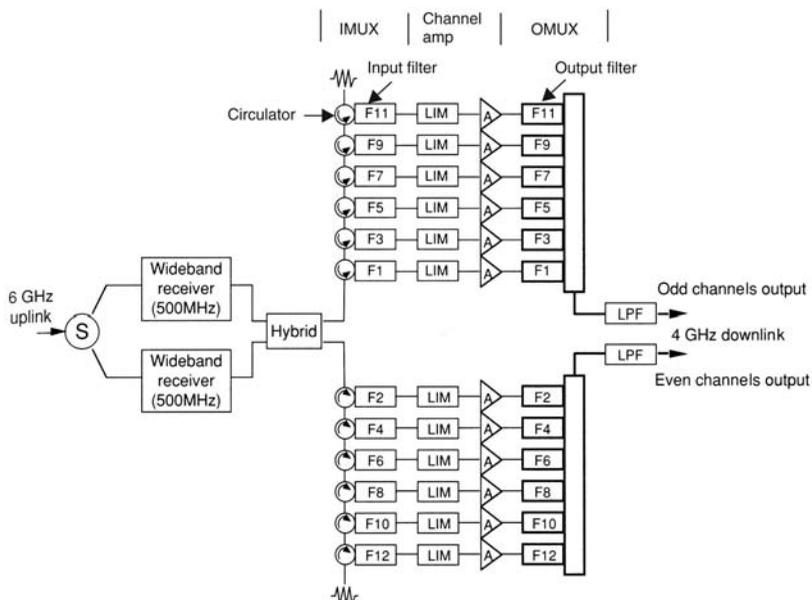
Figure 6.8   Block diagram of a typical 12-transponder C-band repeater.

power entering a port to rotate in the direction of the arrow and exit at the next immediate port. Rotation is set up by a disc of ferri-magnetic material sandwiched between a pair of magnets to create a magnetic field in a direction perpendicular to the disc in Figure 6.8. The polarity of the magnetic field determines the direction of rotation, which is fixed using permanent magnets or switchable using electromagnetism.

Because all the channels pass through the circulator together, it is the function of the input filter to select only one channel for amplification by that particular transponder. The input filter allows the specified transponder channel to pass through and reflects all others back to the same circulator. Thereafter, the five remaining channels (plus the other six channels) are routed down the circulator chain with one more channel being dropped in each subsequent bandpass filter. By the time the last reaches its bandpass filter at the end of the chain, it has experienced the reflection process five times (all other channels experience the bouncing process proportionately fewer times). That reduces the power of the last channel by a significant amount due to losses in the circulators and filter input sections. The amount of loss, however, is significantly less than 1/6 (i.e., less than 10 log 6, or 8.8 dB), which is what would have been the result with a broadband power divider.

At the final arm of the chain is a passive load (called a termination) to absorb any remaining channel energy and to prevent power from being reflected back up the chain. Circulator coupling is extremely effective for the purpose of separating channels, ensuring that only one channel enters each transponder chain. Leakage of adjacent channels produces an undesired effect called repeater multipath (which is different from propagation multipath discussed in Chapter 4). A step taken to reduce the problem is to narrow the IMUX filters as compared to those of the OMUX. (Repeater multipath is reviewed in Section 6.4.3.)

The other filtering component of the satellite repeater is the OMUX, which connects to the output side of each power amplifier. This device uses reactive combining instead of circulator coupling to minimize the loss of valuable RF power (recall the discussion of this issue earlier in the chapter). It effectively sums up the power of the individual channel power amplifiers (TWTA or SSPA, depending on the payload design), using the fact that they are on different frequencies to prevent undesired interaction and power loss. The way that the OMUX does that is discussed later in this chapter.

In addition to the functional components of the repeater, extra receivers and power amplifiers are provided to allow replacement of failed units by ground command. Another role of switching is to allow channels to be connected to different beams, either on an occasional basis (such as once a year or possibly monthly) versus requirements that might be more dynamic in nature. In regard to the latter, the digital processing type of payload is the preferred approach. Processing payloads also can support dynamic beam forming and reshaping, which are extremely useful in broadband applications.

6.1.5   Linearity

The EIRP, G/T, and transponder channelization requirements are among the most important in meeting the basic service requirements. The challenge from this point

is to ensure that the secondary performance aspects of the payload are in keeping with the services that the satellite is supposed to provide. The first of those relates to the linearity of the transponder and its impact on distortion effects produced on the carriers that traverse the payload. In Chapter 5, we presented several of those aspects and explained the basic engineering methodologies used in satellite network design. The particular types of transponder linearity concerns were identified as well. Another important aspect is how susceptible the satellite is to uplink and downlink interference, which relates to the isolation properties (if any) from the spacecraft antenna system.

A growing area of interest is in how digital payloads influence the quality of service. That is because such repeaters incorporate A/D conversion as a key step in preparing the uplink channels for processing and D/A conversion to restore the analog format of the downlink. In Chapter 5, we explained that distortion introduced by the converter is directly affected by the sampling rate and the number of bits per sample. The onboard processor likewise is constrained regarding those parameters because they determine the amount of digital data bits that must be processed and ultimately the power requirements and the mass of the processor itself. That imposes critical tradeoffs between distortion on the one hand and digital payload complexity, cost, and spacecraft bus demands on the other.

6.1.6   Frequency Translation Effects

The last payload requirement to be considered here is one that can affect the accuracy and stability of the frequency of all signals that pass through the repeater. That is the error introduced by the frequency translation process common to nearly every satellite payload design. The exception is the rare spacecraft that does not employ frequency translation but relies completely on time switching to separate transmit from receive (the Iridium spacecraft includes this feature, but nevertheless it still provides frequency translation to be able to route a high quality of narrowband channels between beams, satellites, and gateway Earth stations).

The basic concept of frequency translation is similar to the operation of the upconverter and downconverter (discussed in Chapter 4). Most of the bent-pipe designs since the 1960s have a single translation from the uplink band to the downlink band, using a fixed frequency offset that should correspond exactly to the difference in frequency between the centers of the uplink and downlink bands, for example, $F_{up} - F_{dn}$. For that arrangement, there is no inversion of the band during translation; in other words, the lowest uplink frequency will reside at the lowest downlink frequency instead of being flipped to the higher end. That translation frequency, $F_{up} - F_{dn}$, is created by an oscillator in the payload. In any real design, there will be some error in that frequency, much like the time error of a quartz clock. However, unlike a wristwatch, the satellite oscillator cannot easily be adjusted. It is common for that error to amount to about one 10-millionth of the translation frequency, which is about 2 kHz at C- and Ku-bands. That amount of error can be ignored in the majority of applications.

In addition to the absolute error, the translation oscillator adds phase modulation noise, which could interfere with services. Phase noise is an undesired variation

in oscillator frequency that produces spectral sidebands, as shown in Figure 6.9. During frequency translation, the sidebands are transferred to all carriers that pass through the mixer stage. In the receiving Earth station, phase noise is demodulated and contributes to the total baseband noise picked up along the satellite uplink and downlink. In a properly designed oscillator and mixer, phase noise is attenuated rapidly in the region of user baseband frequencies. That was important in analog FM and PM systems because they were particularly sensitive to the impairment. Digital signals are less impaired provided that the modulated data rate is greater than about 100 kbps.

Along with phase noise, an improperly designed downconverter stage on the satellite can introduce unwanted frequencies called spurious frequencies (spurs) that are radiated by the satellite along with the downlinked carriers. It is up to the receiver designer to hold spur levels to 20 dB or more below the level of the weakest signal to be transferred. There are situations in which spurs cannot be further reduced in power, and the system must find ways to accept or avoid them. It is possible that some spurs will reside at frequencies outside the specified downlink band and therefore will not interfere with the satellite network. While not a problem for our service, the spurs still may violate ITU rules for allowable radiation within the microwave spectrum. Alternatively, spurs actually may have to be coordinated just as if they were intended communication frequencies.

We have reviewed the primary specifications that are placed on the payload. Depending on the particular system design and user requirements, the specifications may be expanded to include other technical matters of importance. Examples of these include channel group delay distortion, gain slope and amplitude modulation to phase modulation distortion (AM/PM). It is up to the system designer to work closely with the engineers, who will actually design and manufacture the satellite so that all the relevant factors have been identified and properly specified. In the real world, some of those effects cannot be controlled to everyone's satisfaction, so a compromise must be developed. To do that properly, trade studies can be performed to learn what parameters can be relaxed to allow the hardware to be built within cost and time constraints.
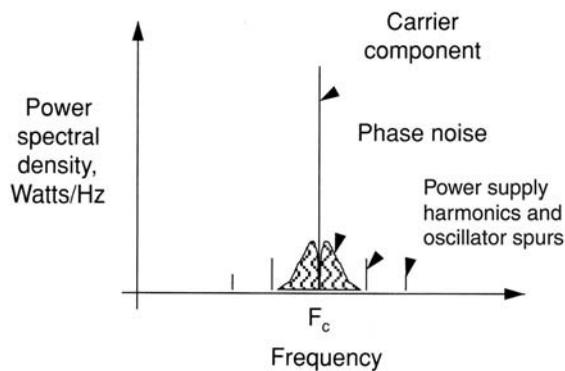
Figure 6.9 Illustration of phase noise sidebands on a carrier.

6.2   Analog Bent-Pipe Repeaters

The next step in payload design is to determine the general type of repeater that will meet the higher level system requirements. The basic bent-pipe design is intended for applications that involve wideband carriers for analog and digital transmission over relatively wide area beams, such as that illustrated for Horizons 1. This approach is the simplest to implement and provides the most flexibility as to signal types, bandwidths, and multiple access. In fact, a bent-pipe repeater can accommodate all three multiple-access methods, that is, FDMA, TDMA, and CDMA.

The simple block diagram in Figure 6.8 provides a basic example of a bent-pipe repeater capable of 12 transponders in a single wide area coverage beam. Repeaters of this type are being built, launched, and used at the time of writing of this book and therefore will be available for service well into the future. They represent the starting point for all repeaters, which retain all or part of this structure.

While the bent pipe is the most versatile approach, it provides limited options for designers to optimize performance for a particular service. For example, a bent-pipe transponder has no intelligence to permit routing of channels between Earth stations. Nor can it help improve the throughput and error rate performance by restoring signal quality prior to retransmission to the ground. As the link budget clearly shows, satellite impairments and link noise are directly additive to the service during the bent-pipe turn-around process.

This brief introduction to the bent-pipe repeater gives some of the top-level guidelines for the payload designer. In Section 6.3, we discuss options for designing and employing digital processing repeaters. Many of the techniques can be combined with those of the bent pipe to realize greater overall service capability.

6.3   Digital Processing Repeaters

The service capabilities of the bent pipe have been well established during 40 years of evolution, culminating in large three-axis satellites with up to 100 active transponders. Owing to the limitations of the bent pipe coupled with advances in DSP and solid state devices and the availability of high-speed VLSI chips, digital processing repeaters are entering the satellite marketplace in greater numbers. Satellite communications engineers generally have applied those technologies on the ground to improve the cost and performance of Earth stations and user terminals. Examples include the digital modems found in data VSATs, DVB-S and DVB-S2 encoders and receivers used in DTH TV systems, and handheld devices that support various voice and data applications over satellites.

In years past, digital processing was limited to RF or IF switching for use in wideband TDMA systems such as INTELSAT's SS/TDMA. That type of service depends on large, expensive Earth stations that can develop the quantities of information traffic to justify the investment on the satellite. Current generations of MSS spacecraft have processing payloads that introduce narrowband channelization services for digital telephony and low-speed data. An important constraint in these designs is the maximum processing speed and sampling rates allowed by

then-existing DSP technology. There are now wideband versions that offer true multimedia applications to low-cost Ka-band VSATs, where the user can access the network directly over the satellite at megabit-per-second speeds. The fundamental approach to digital processing is to actively alter the signals in the satellite before they are retransmitted to the ground. That involves such functions as narrowband channel selection and routing, switching of time division channels or packets, decoding and recoding of data streams for onboard FEC, and complete reformatting of data for optimum downlink transmission. The technologies to accomplish that include high-speed A/D conversion, fast Fourier transform (FFT) computation, digital filtering, data speed buffering, and coding. It is difficult to try to describe all the possibilities in a short chapter; in fact, the technologies continue to evolve rapidly.

6.3.1   Multiple Beam Switching and Routing

Satellites used for current and evolving interactive voice and data services are very demanding in terms of the aggregate amount of information to be transferred. In addition, the satellite antenna system must have the highest possible gain to minimize uplink transmitter power as it often originates from a low-cost device like a handheld phone or ultrasmall aperture terminal (USAT). The most direct way to address this from a satellite architecture perspective is to employ many small spot beams to cover the region of interest. Each beam has a much higher gain than that of the typical area-coverage type satellite; and it affords a high degree of frequency reuse with the multiple beam approach.

Figure 6.10 indicates a coverage arrangement for the United States and Puerto Rico with 22 spot beams that cover the major metropolitan areas. The fact that small cities and rural areas are not covered limits the service, but puts most of
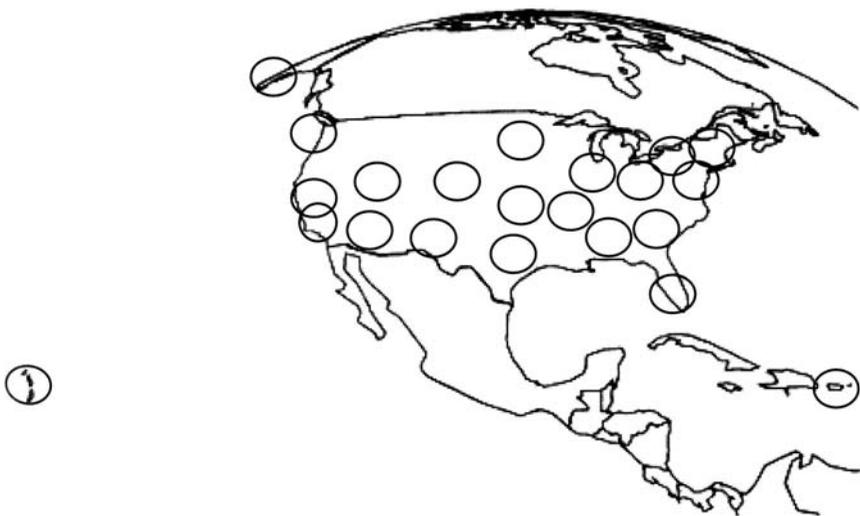
Figure 6.10   Satellite coverage with multiple spot beams employing frequency reuse and interbeam routing.

the power into the potentially more attractive markets. (This may end up being unattractive since people in metropolitan areas have more options for wireless services than those in regions uncovered by terrestrial wireless providers.) Alternatively, one could provide a complete cellular-like coverage of the entire region and thus leave nothing out. With 22 beams, it might be possible to achieve a reuse factor of 20, meaning that the effective spectrum is 20 times the raw ITU allocations. If that allocation was 500 MHz, then this satellite, in principle, could achieve a total of 10 GHz. The actual usable spectrum will be reduced by interference produced by groupings of beams like the cluster of four in the northeastern United States. Overlapping of the beams in the west that cover Los Angeles and San Francisco would be resolved through polarization discrimination like that used in area coverage systems. Likewise, the beam on Seattle and on Anchorage could be isolated just by geographic separation with use of the same polarization.

A satellite with 22 beams represents the low end of the range with as many as 200 being more typical. The Spaceway satellite is actually capable of more than 1,000 beams, each with a proportionately smaller size. The challenge in multibeam satellite design is to provide the interconnecting paths within the payload so that traffic can be transferred accurately and efficiently. This is similar to the operation of switching telephone calls in the public switched telephone network or the routing of data packets in the Internet. A generic diagram of such a satellite payload is illustrated in Figure 6.11. There are four techniques for accomplishing this beam-to-beam transfer:

1. Intermediate frequency (IF) channel routing, retaining modulated carriers;
2. Baseband switching, which requires that carriers be demodulated on board and the information treated as streams of data that can be switched by a digital matrix;
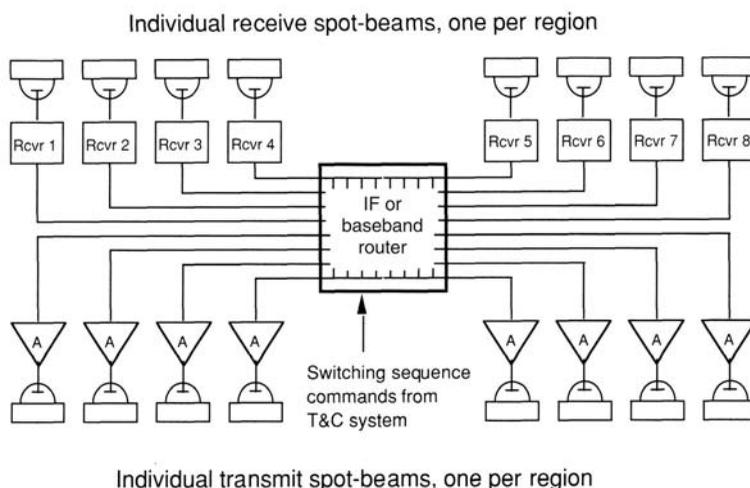
Figure 6.11   A generic multibeam routing repeater.

3. Packet switching, a refinement of the baseband routing, to treat each block of data individually and route it in the same manner as an Internet Protocol (IP) router.
4. Ground-based beam forming which relies on phase and amplitude processing in a master Earth station.

All four configurations are currently in use onboard operating multibeam satellites (see Figure 6.12). The following discussion is based on the IF channel routing approach. The traffic from uplink sites is contained in narrowband bursts of information, based on a low data rate TDMA scheme.

The actual routing pattern, which defines the traffic distribution matrix, is stored in an onboard memory. The pattern is uploaded from the ground control station using a data communication link (which might be part of the command system). Because traffic transfer requirements change over time, the routing pattern can be altered by uploading new data into the switch memory. One approach is to maintain several patterns in the onboard memory so a new configuration can be activated in rapid response to demand. The pattern also could be altered automatically at specific times of day.

Synchronization of the overall network is critical because bursts must arrive at the correct time to be transferred between beams without collisions. That, of course, is easy to do when the onboard router ties the same uplink and downlink
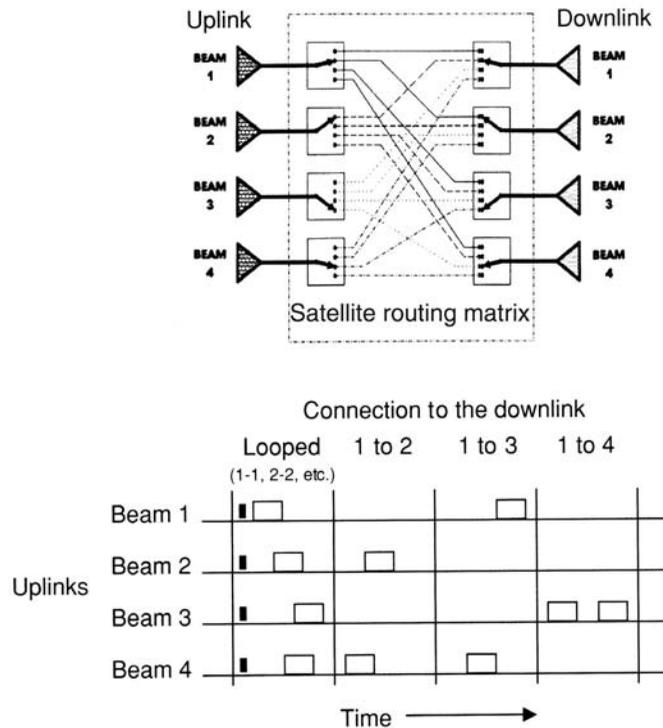
Figure 6.12   Burst timing and traffic transfer within a multibeam payload.

beams together. Stations must be given strict time slot assignments within the overall frame to ensure that bursts do not collide.

6.3.2 Digital Processor Architecture

All digital processing payloads utilize mathematical representations of the necessary functions in the repeater, including narrowband channelization, frequency translation, multiplexing, and aggregation of traffic for retransmission to the ground [1].

A simplified digital processor architecture is shown in Figure 6.13, wherein each uplink beam is provided with a separate wideband receiver. In the case of L-band, the bandwidth of the receiver is on the order of only 30 MHz, which is consistent with the available spectrum width and the overall capacity of the processor. The output of the receiver contains the spectrum of user channels at a low IF suitable for processing operations. In the processor, the bandwidth is digitized in the A/D converter to produce a combined bitstream. All functions from this point are performed with mathematical algorithms that represent analog functions like frequency translation, filtering, and routing. The frequency converters are used to move individual narrowband channels to different frequency slots in a time division stream. Channels can be routed using time division techniques similar to what is done in a conventional telephone switch; that is indicated in Figure 6.13 by vertical arrows. Next, the math is performed to convert the channels to the final frequency slot assignments and the resulting groups are transformed back to analog spectra with D/A converters. Final frequency conversion and amplification are then performed in the conventional manner.

The fact that the bandwidths involved are relatively narrow in comparison to bent-pipe satellites reflects the fact that the first generations of these systems are intended to support low-cost user terminals, particularly the hand-held variety used for mobile telephone service. The actual transmissions can use FDMA or narrowband TDMA, as discussed in Chapter 5. Current technology probably can support CDMA as well but at a reduced capacity. That is because of the added complexity needed to despread each CDMA signal onboard the satellite.
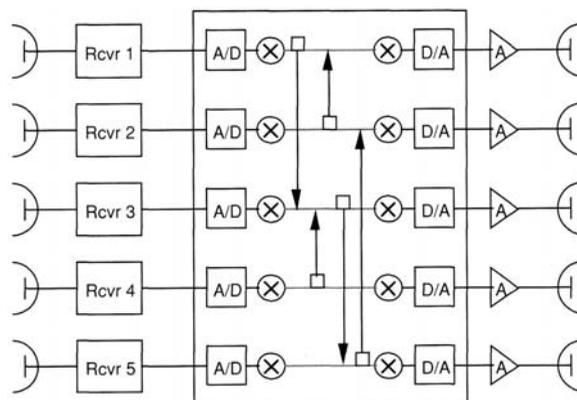
Figure 6.13 Example of a digital processing repeater providing narrowband channel routing.

More advanced digital processor architectures have been assembled with the following needs in mind:

- Narrowband channel routing, to allow users to transmit information within channels in the range of 100-kHz to 10-MHz bandwidth. Within the constraints of this channel, any modulation or multiple access method can be supported.
- On-demand reconfiguration, to provide the network operator with the ability to change the channelization assignments in response to changing demand.
- Beam forming to remove the restriction on the number of beams and their shape imposed by physical feed horns. This requires that a given channel be able to drive multiple feed horns, after the replicas have been adjusted properly in amplitude and phase to produce the correct beam shape and gain.

Whenever analog signals undergo the A/D process, we can anticipate that there will be some distortion and information loss. These can be held to acceptable levels if the following constraints are addressed correctly (a detailed discussion of these characteristics is beyond the scope of this book):

- Satisfactory signal structure (multiplex, modulation, and multiple access). They must satisfy the requirements of the link as well.
- Bandwidth, which determines the size and complexity of the processor. Ultimately, it is a question about how the required processing functions can be performed with the current state of the art.
- Processor timing, speed, and complexity, which are constraints that are difficult to assess in the abstract and that require a careful analysis by relevant experts.
- Impairments, because the ideal processor that causes no impairment to signals cannot be built with real hardware.
- Dynamic range, because the processor must deal with the full spectrum of signals that enter at its input ports. Dynamic range refers to the range of power levels of these signals. It can cover as much as a 20-dB range or more in a typical application.

The needs on the one hand and the constraints on the other bound the design of an adequate processor that is also buildable through current technology. Systems engineers have to work closely with processor engineers so that all relevant factors are properly considered and that a practical design is developed. Any reduction in capability along the way should not pose a risk to the overall project. That can be handled only by proper management of the design.

As an example of a typical digital processing repeater design, consider the block diagram in Figure 6.14. The diagram represents a European developmental project for an MSS spacecraft that provides links between mobile users and a fixed gateway Earth station [2]. In the block diagram, the uplink at C-band is fed through a wideband receiver to an A/D converter, followed by a demultiplexer in digital
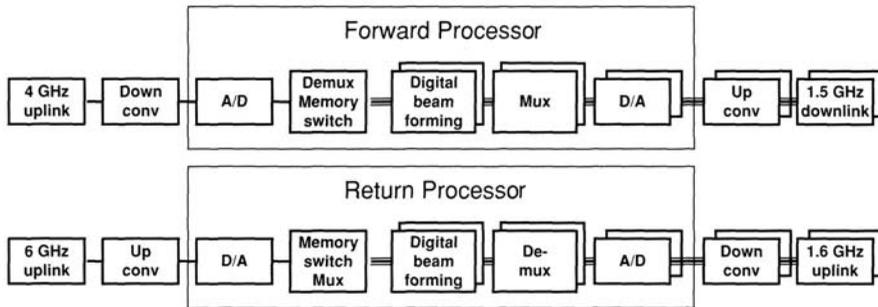
Figure 6.14 Example of a digital beam forming mobile payload.

form. That, in turn, drives a memory switch that is used to separate out the individual frequency channels for further processing. The digital beam-forming element modifies the channels through complex multiplication by specific amplitude and phase values to produce the desired downlink beam. The values are uplinked from the network control station and can be changed as required over the satellite lifetime. The remaining functions are used to adjust the power level and translate the frequency to the correct beam channel. A D/A converter restores the analog format before RF frequency conversion, power amplification, and transmission from the antenna system.

Without going into details as to the design and supporting mathematical theory, the channel selection and beam forming are accomplished through the discrete Fourier transform (DFT) technique. That allows the processor to perform its functions in the frequency domain using mathematical representations of filters and mixers without using actual hardware. In that approach, the DSP designer must consider these factors:

- The number of bits used in A/D conversion;
- The required number of A/D and D/A oprations;
- The size of the DFT used in processing, for example, the number of points, amplitude granularity, and sampling rate;
- The window overlapping factor, to smooth the data for DFT processing and to minimize the amount of computation and memory; must be optimized for the types of channels upon which the processor operates;
- The time window, which determines the amount of resulting ripple and inband interference;
- Techniques to ensure that all necessary operations are properly synchronized within the relevant stages of the processor.

Traditional payload engineers are accustomed to working with analog parameters, such as those discussed in previous sections. The factors listed in this section tend to be more difficult to assess and optimize. However, the situation may resolve itself as the tools used to simulate and synthesize the necessary circuitry are improved and enhanced. Some satellite architects have chosen to perform the

processing and beam forming on the ground to simplify the payload and allow significant modifications to be made post launch.

6.3.3   Demod-Remod Repeater

The last category of digital processing payload that we will discuss is one that places the demodulator function at the midpoint in the satellite link. This offers the opportunity to completely restore the digital format of the information and thereby improve signal quality before transmission to the ground. Furthermore, specific protocol information can be detected onboard and used to control data transfer in real time. An example of a demod-remod transponderized payload is shown in Figure 6.15. Individual channels are filtered as is done in a bent-pipe design. From that point, the uplink is demodulated on an individual carrier basis. Its output in digital data form can be switched and remultiplexed before remodulation and amplification. The last step is conventional output multiplexing to combine the powers of all the transponders prior to delivery to the transmit antenna. The demod-remod functions could be performed by discrete hardware or with a DSP repeater.

This particular approach really provides only one basic benefit of the demod-remod payload, namely, the bits are restored onboard the satellite. That prevents the uplink noise and interference from being directly transferred to the downlink. Instead, those impairments simply introduce bit errors into the data. While such errors cannot be eliminated on the satellite (except through FEC decoding), the result on the final received data has fewer errors than if demodulation were done only on the ground. The maximum benefit of almost 3-dB improvement occurs if the uplink and downlink C/N values are equal. However, if the uplink level is increased to give a substantial difference in C/N, the benefit reduces essentially to zero.

The Gen*Star payload that was developed by TRW (now Northrop Grumman) has extensive baseband processing and packet switching capabilities. As illustrated in Figure 6.16, the processor takes as input the uplinks of many beams, each coming from individual users and from gateway earth stations. The demodulators within
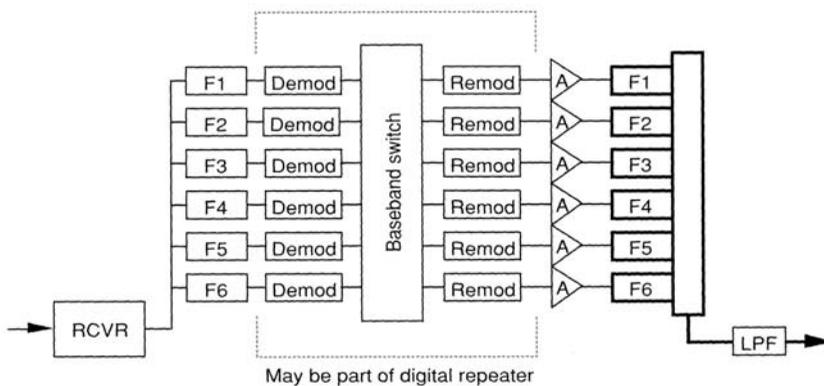
Figure 6.15   Demod-remod repeater configuration (transponderized design).
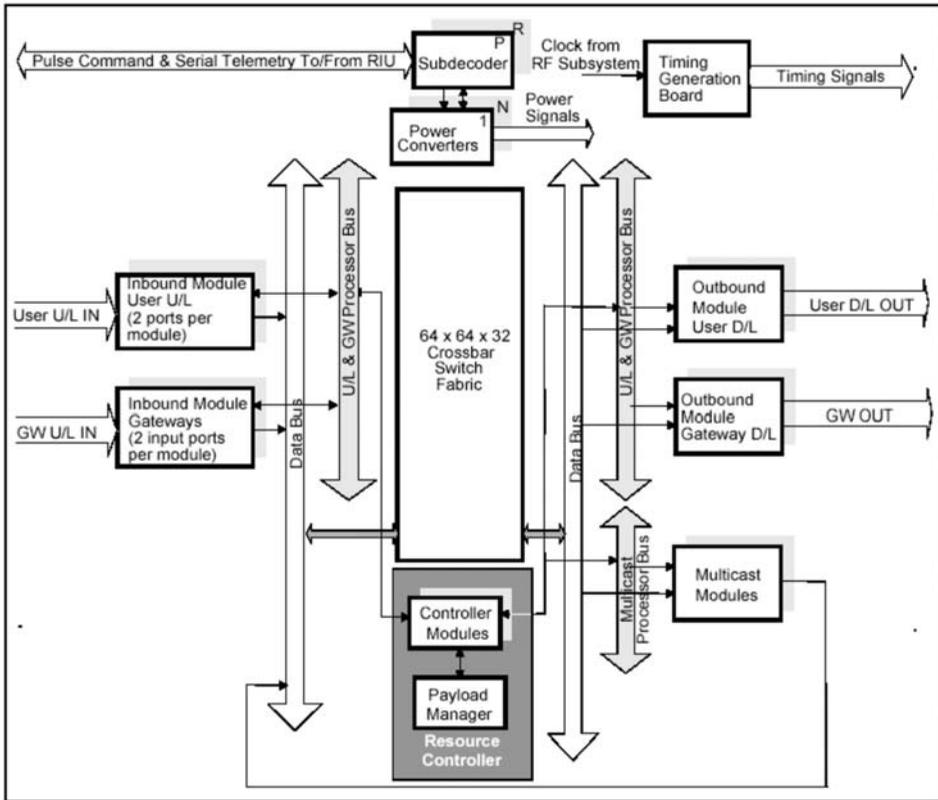
Figure 6.16   The Gen*Star payload digital processing payload concept employed by Astrolink [3].

the Inbound Module on the left convert the signals to digital form and samples are processed to select three different channel burst rates, up to a maximum of approximately 16 Mbps. Each channel is demodulated and decoded. The demodulator outputs the data formatted into ATM cells for presentation to the ATM switch, which uses an onboard computer for control. A special feature of the processor is that ATM cells can be assigned priority through a quality of service (QoS) scheme. In addition, the switch can generate statistics on the cells and routing performance to be evaluated at the network control facility. Following the output side of the processor, the analog carriers are amplified at Ka-band using high-power TWTAs. The quantity of individual inputs and outputs can be designed based on the mission, with 64 given in this example. TRW supplied this payload for the Astrolink satellite program, which actually used 48 fixed beams.

6.4   Standard Repeater Elements

Modern communication payloads have a wide variety of types of subsystems and elements, each of which must be relied on to perform an important and sometimes critical function. The previous sections of this chapter reviewed the requirements and the architecture of analog and digital payloads, relating their performance to

the needs of the overall mission and applications. In this section, we go into detail on the building blocks of the repeater so that readers will have a full appreciation for the design and interaction of these important pieces of space hardware. We follow the functional sequence of major components in the typical repeater block diagram shown in Figure 6.4. Readers not needing this level of technical detail can proceed to Chapter 7.

6.4.1   Wideband Receiver

Satellite repeaters have as their first active task the job of amplifying the desired uplink frequency band, and this is normally done in a unit called a receiver. As shown in the block diagram in Figure 6.17, the range of channels in the uplink frequency band appears at the input; the same bandwidth appears at the output, but translated to the downlink frequency band. In addition, the power level of the signals has been increased greatly, by an order of 50 to 60 dB (i.e., amplified in power by a factor of 10^5 to 10^6), depending on the particular design. This makes up about half of the total gain of the repeater (typically 120 dB). The difference, $F_u - F_d$, is a fixed value offset between uplink and downlink center frequencies. In C-band, for example, the difference frequency is 2,225 MHz, and the same difference applies for every transponder of the repeater. In the receiver, translation is produced through downconversion in the mixer. There are two inputs to the mixer: the uplink frequency range and a pure single carrier from the LO. The frequency of the local oscillator, F_{lo}, clearly is the difference frequency. It is conve-
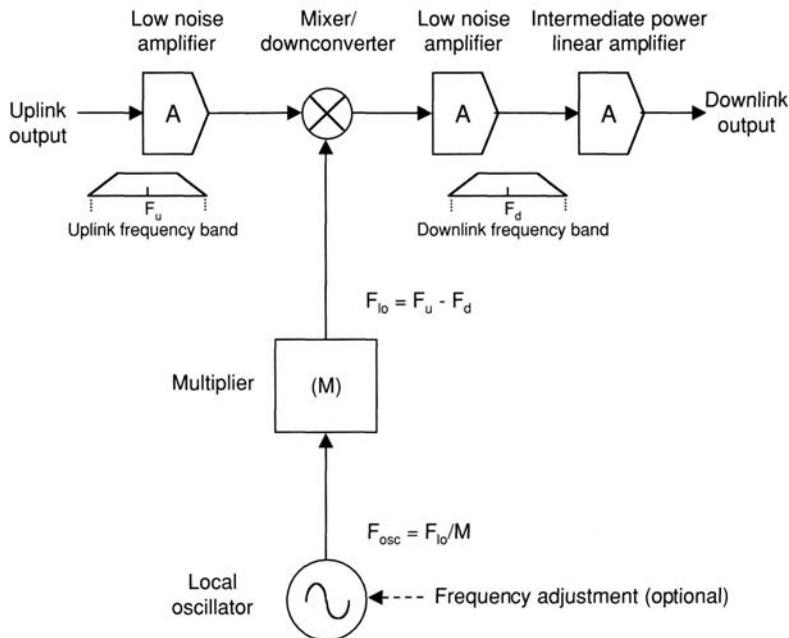
Figure 6.17   Block diagram of a typical wideband communication receiver that amplifies the 500-MHz uplink band and translates it to the downlink frequency range without modification.

nient to use a device called a frequency multiplier to generate F_{l_0} as a harmonic from a crystal-controlled oscillator operating at a much lower frequency. For our C-band example, the oscillator typically runs at 139.0625 MHz, which means that the multiplication factor has to be the integer 16.

Because multipliers and mixers are nonlinear devices, they generate a variety of harmonics, spurs, and intermodulation products in addition to the desired frequencies just described. Therefore, filters and frequency traps are inserted in the input and output legs of the mixer to preclude propagation of such products through the receiver or backward toward its input.

In addition to frequency translation, the receiver must provide a great deal of the required amplification to overcome a portion of the uplink path loss. Approximately 50 to 60 dB of gain is easily obtained with several transistor amplifier stages in tandem, as illustrated in Figure 6.18.

An important consideration in satellite receiver design is the thermal noise, which is produced in all active and passive microwave devices. The term *low noise* (discussed in Chapter 4) is used to describe an amplifier with specially designed and selected transistors that have the property of minimizing the generation of internal noise. A tandem sequence of amplifiers, such as that in Figure 6.18, provides significant noise reduction. In the following equation, we see that the total noise temperature of the sequence is calculated based on the individual noise contributions, reduced by the gain of preceding stages:

$$T_{re} = T_1 + T_2/G_1 + T_3/G_1 G_2 + \ldots$$

where

T_1 and G_1 are the noise temperature in Kelvin and gain (as a ratio), respectively, of the first amplifier stage;

T_2 and G_2 are the noise temperature and gain (as a ratio), respectively, of the second amplifier stage;

T_3 and G_3 are the noise temperature and gain (as a ratio), respectively, of the third amplifier stage, and so on.

This type of analysis shows that the total noise temperature is dominated by the noise contribution of the first amplifier stage. The contribution of all the stages that follow is reduced by a factor equal to the combined gain of the stages that precede it, expressed as a true ratio and not in decibels. If one of the stages provides
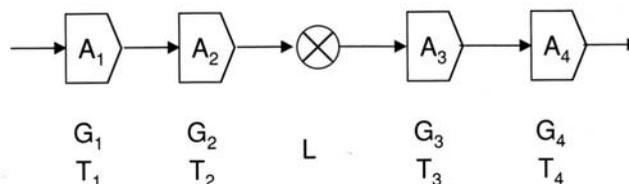
Figure 6.18 Use of amplifier stages in a typical wideband receiver.

frequency conversion in a mixer, we can anticipate a loss instead of a gain for that particular stage. That moves the factor from the denominator into the numerator of the equation. For example, inserting a downconverter after the second amplifier, as shown in Figure 6.18, results in the following modified equation for the receiver noise temperature:

$$T_{re} = T_1 + T_2/G_1 + LT_3/G_1 G_2 + LT_4/G_1 G_2 G_3 + \ldots$$

where L is the conversion loss factor ($L > 1$).

The conversion loss has the unfortunate effect of amplifying the noise power contribution of the subsequent stages. There is a tradeoff in receiver front end design, in that there is a desire to increase the gain ahead of the mixer to reduce its impact on total noise, countered by the concern that the higher signal level produced by the gain can increase the spur level coming out of the mixer. For that reason, there are at most two stages in front of the mixer stage.

Following the mixer typically are two separate transistor amplifiers, the first still being designed for low noise. The final stage increases the power level of each of the channels sufficiently to drive the input to the TWTAs or SSPAs, as appropriate. While these stages do not add significantly to system noise, the relatively high power level does place considerable demand for linearity on the last stage. The intermodulation products produced in this wideband stage, which increases as the output power demand increases, must be kept at an acceptable value or the transponders will be subject to unwanted RF IMD interference. Because all carriers pass through this stage, it must have a peak power output capability considerably in excess of actual requirements (e.g., by 20 dB or more).

The elements of the receiver along with the associated power supplies are housed in a metal box to protect the components and facilitate placement on the repeater equipment shelf of the satellite. Because frequency translation is influenced by the stability of the internal LO, the typical receiver includes a thermostatically controlled heater. That minimizes the variation of temperature directly affecting the crystal reference, particularly as the satellite passes into and out of eclipse. A properly designed and compensated receiver produces a shift of less than 1 kHz during the eclipse cycle. More accurate frequency translation may be provided with an external frequency correction or an atomic standard.

6.4.2 Redundancy Switching

The active components of the repeater are the most likely parts to fail and thereby reduce or cease service performance. Included are the receivers just described, along with transponder amplifier and gain control devices. To counter that, all repeater designs include backup in the form of redundant components and subsystems. They usually are turned off and held in reserve until needed, at which time they are individually switched on by ground command. Manual commanding operation of redundant equipment has been the practice but viable approaches for automatic restoration are available. In the following discussion, we review the RF switching devices that are used to replace a failed active unit by an appropriate spare unit.

Techniques for switching active and spare units consider three important attributes. First, the switches must pass microwave signals with the minimum amount of loss and distortion. Second, the switching scheme must permit the repeater to be reconfigured to allow access to spare units without disrupting the remaining operating equipment. Switches also must be reliable because a given switch often represents a potential single-point failure mode (i.e., failure of a switch could render a transponder or group of transponders useless). Switches that use coaxial connectors (similar to those used for interconnecting video equipment) are appropriate for RF power levels under 50W and where some measurable loss is acceptable. Waveguide switches, although heavier and bulkier than coaxial switches, are necessary for high power levels and low-loss applications, particularly at Ku- and Ka-bands.

The following paragraphs review the currently accepted switching devices and redundancy configurations, showing how they are used in the repeater. Because this subject is important but usually ignored in the literature, readers may find the discussion valuable. The illustration that is provided with each switch shows the symbol, the schematic diagram, and a typical application of the particular switch installed in a repeater.

6.4.2.1 S Switch

The S switch, shown in Figure 6.19, is the simplest and most basic form of a redundancy switch. The single coaxial input can be routed to either of two coaxial outputs; thus, it is a single pole–double throw switch. The application drawing with two amplifiers (or perhaps receivers) shows how a backup unit can be used to replace a failed or degraded oprating unit. This scheme is often called two-for-one redundancy, meaning that two identical units are carried to provide one operating channel of service (not to be confused with the designation "one for one" used in terrestrial radio systems for the same configuration). Note how the
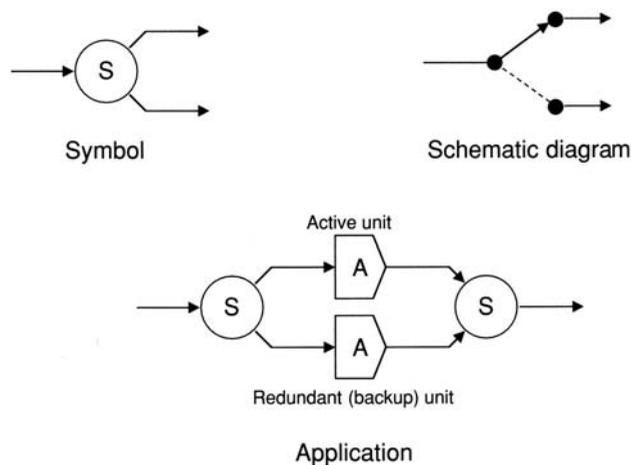
Figure 6.19 Definition and application of the S switch, which is used to switch in redundant equipment.

second S switch is reversed to connect from two paths back to a single output. Two-for-one redundancy with S switches is required whenever the devices must cover the same frequency range or when full redundancy is needed to meet reliability requirements.

The S switch can be implemented with two different technologies. The most common approach is to use coaxial input and output connections and a mechanical wiper arm, similar to the A-B switch used to connect a TV receiver to either of two cables. Mechanical motion is produced by an electronically driven actuator controlled by ground command. A type of nonmechanical S switch uses ferrimagnetism to direct microwave energy from the input to the selected output. It operates similarly to a circulator and is controlled by pulsing a magnetic coil to reverse the direction of the internal magnetic field. Ferrite switches often are used at the input to C- and Ku-band receivers because the ferrite cannot fail as an open circuit even if the cotrol circuit stops in midposition. That failure mode introduces a 3-dB loss on both switch outputs, requiring only that one of the two units not be active.

6.4.2.2 2/3 Switch

As shown in Figure 6.20, two S switches can be packaged together to provide a versatile switch with two ports on one side and three on the other. If the two-port side is the input, it is called a 2/3 switch; with the three ports, it is a 3/2 switch (it depends only on how the switch is used in the repeater). The fundamental application, shown at the bottom of Figure 6.20, is to allow three-for-two redundancy, where one spare unit can back up two operating units. An extension of this method uses a pair of 2/3s and 3/2s along with a pair of S switches to implement a five-for-four redundancy scheme. Many C-band satellites use five-for-four
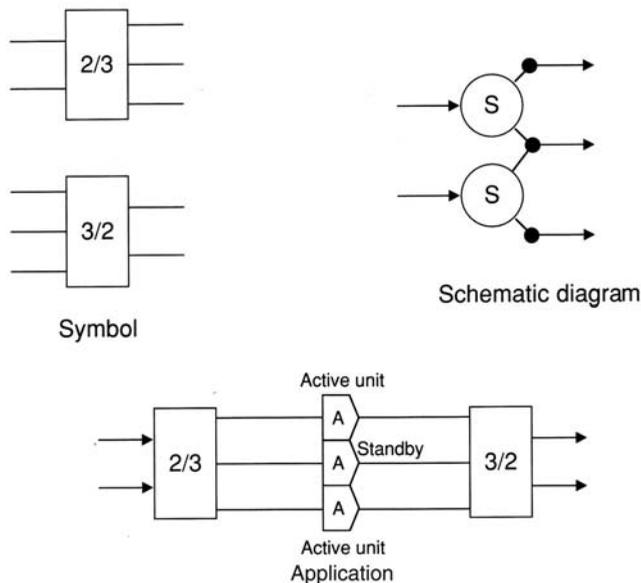
Figure 6.20 Definition and application of the 2/3 switch, which is used to share redundant equipment.

redundancy for sparing of TWTAs. The spare amplifier in such cases must be capable of operating over the bandwidth of each of the transponders that it is sparing.

6.4.2.3   C Switch

One of the more useful arrangements is the cross (C) switch, shown in Figure 6.21. The schematic diagram in the upper right corner shows how the two inputs and two outputs can be connected. In the straight-through position, the upper input goes to the upper output and likewise for the lower input. However, if the switch is commanded into the cross-over position, the upper input goes to the lower output and the lower input goes to the upper output, the only two positions in this particular design. The C switch uses coaxial input and output ports and therefore is suitable for low-level signals (i.e., under 50W). Alternatively, waveguide C switches are available that can handle hundreds of RF watts.

A typical application of the C switch, shown at the bottom of Figure 6.21, allows two operating units to be interchanged. If the upper amplifier were to fail, it would be possible to restore that particular channel of communication by commanding both C switches from straight-through into cross-over. Obviously, communications capacity is lost, but at least there is the option of maintaining the more important of the two channels. The same switch also can be used as a routing device to rearrange communication paths within a repeater.

6.4.2.4   R Switch

The last switch to be described is designed to increase the ratio of active units to spare units (i.e., to anticipate a lesser need for spare units in relation to a larger quantity of operating units). The advantages of a waveguide redundancy (R) switch are low transfer loss, nearly perfect isolation, and high power capability. At Ku-band, the dimensions and weight of an R switch are reasonable for a typical
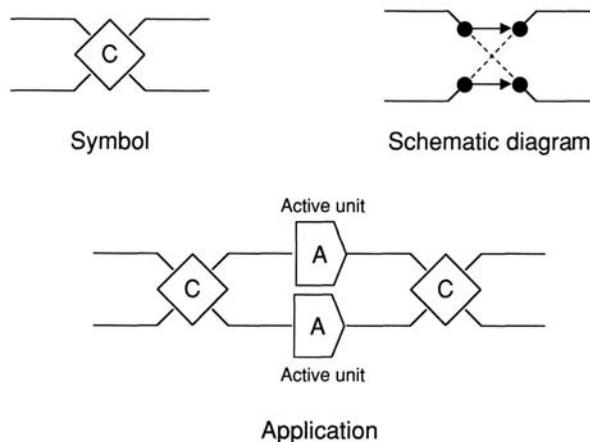
Figure 6.21   Definition and application of the C switch, which is used to interchange equipment or switch paths.

payload. The situation at C-band is different because the corresponding waveguide R switch is three times as large and up to nine times heavier. The solution was to develop a coaxial R switch, which was first introduced in the late 1980s.

The device is called an R switch because it is used to restore a failed amplifier with an available spare by providing four separate waveguide ports, as shown in Figure 6.22. The schematic diagram of the switch is the same as its cross-sectional view, showing a metal disc having straight and curved channels machined into it. The channels provide waveguide paths inside the switch for microwave energy to pass between the four ports. It has been called a baseball switch, an obvious reference to the appearance of the cross-sectional view. The schematic diagram shows three possible positions: two cross-over positions and one straight-through position. In some applications, one port is terminated with a resistive load to absorb the full power of an amplifier if the switch is commanded to disconnect the active element.

The redundancy scheme shown in Figure 6.22 is called ring redundancy. A three-channel system is depicted with the R switches in the straight-through position (position 2 in the schematic diagram), placing the backup amplifier completely out of the circuit. If the upper amplifier were to fail, then the spare amplifier could be used to restore service. To use the spare, the input and output R switches for upper and spare amplifiers are put in positions 1 and 3, respectively. That causes microwave signals to be routed from the upper channel away from the failed amplifier and into the spare. The ring is created by continuing the pattern of R switches and amplifiers upward and downward from the four sets shown in the diagram, until the chain connects back into itself.
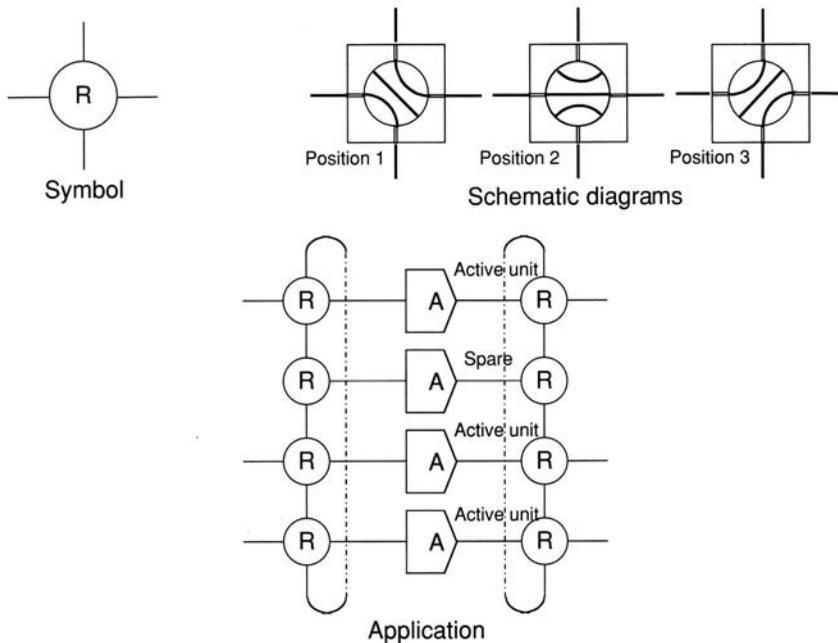
Figure 6.22   Definition and application of the R switch, or baseball switch, which is used to switch in redundant units or between waveguide connections.

The R switch designs and configurations are typical of what is used in the industry, but this discussion is not intended to be all inclusive. There are many variations to the schemes described here, and new switch technologies appear frequently. You can use this framework to understand the specifics of a given repeater and switch design.

6.4.3   Waveguide Filters and Multiplexers

Bent-pipe satellite repeaters can provide several wideband channels of communications (i.e., transponders) when the individual channels are separated from each other on the basis of frequency. As discussed in Section 6.1.4, the input multiplexer separates the channels for amplification, and the output multiplexer combines them again for transmission through the downlink antenna. The typical multiplexer contains a microwave filter for each individual channel, the filter being tuned to pass the full bandwidth of that channel and to reject (or reflect) all others. In addition to microwave filtering, repeaters that operate on narrowband channels such as applied to MSS can use alternative technologies that filter at much lower frequencies. An example is surface acoustic wave (SAW) filters, which can provide excellent channel filtering for bandwidths down to a few kilohertz. Our focus in this section, however, is on the microwave variety.

The input multiplexer is the most critical repeater element with regard to creating the transponder channel. Figure 6.23 is an example of an input multiplexer taken from a typical C-band satellite design. There is a total of 12 channels, each with 36 MHz of usable bandwidth, which are separated into even- and odd-numbered groups (the odd group is illustrated). We will explain why the channels are split that way later in this section. The hybrid in Figure 6.8 is used to split equally the total power of the downlink spectrum, transferring all 12 channels to each output leg. An input multiplexer typically uses circulators to route the entire

Example of a C-band input filter
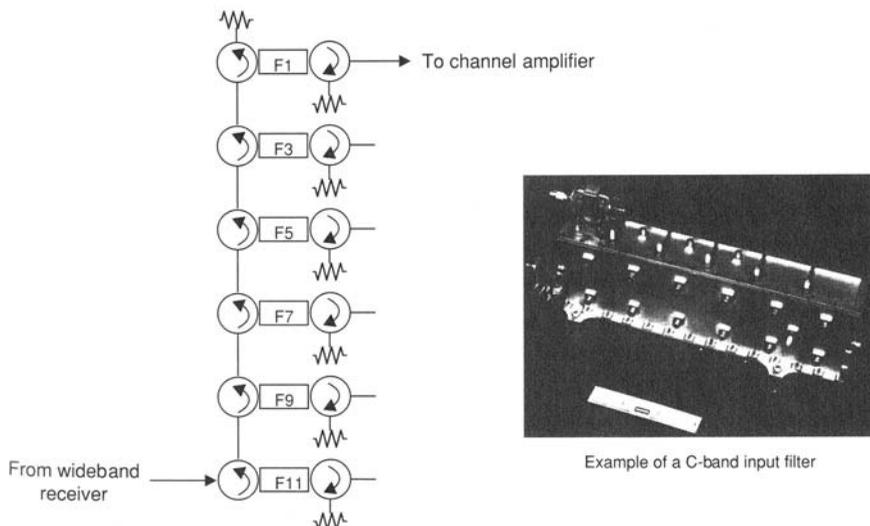
Figure 6.23   A six-channel input multiplexer composed of individual microwave filters.

spectrum of channels to the filters in a sequential manner, each filter being used to drop off one of the channels. One nice feature of using circulators is that they essentially isolate the filters from each other, which from a technical standpoint tends to simplify multiplexer design, manufacture, and final tuning of the complete assembly.

More compact, innovative waveguide filters have been introduced in the past decade. Waveguide cavity filters, made of aluminum and internally loaded with dielectrics, are found to provide superior in-band RF performance (compared to traditional filters) with a sharp roll-off and a flat passband response. An example of such a filter is shown in the input multiplexer chain of Figure 6.24. There are a total of 18 channels, separated in to even- and odd-numbered groups. This input multiplexer has nine channels and occupies about 75% smaller envelope and is 33% lighter than traditional filters.

We have been describing the action of a microwave filter without being specific about how it affects the signals that pass through it. Figure 6.25 gives a qualitative view of the frequency response of the filter by itself and of the combined action of filters in a multiplexer. The characteristic of a single filter (shown at the top of the figure) includes a passband of sufficient bandwidth to allow all the signals within the particular transponder to be transmitted with minimal distortion. A perfect filter would have a flat passband and infinite loss (a "brick wall") for all frequencies that are outside the passband (referred to as the out-of-band response). Real filters are not perfect, so there are variations from the ideal. For example, there is some residual loss, called insertion loss, within the passband. The perfectly flat insertion loss characteristic shown is impossible to achieve, and we must be satisfied with variations in loss of typically ±0.5 to ±1 dB across the passband.

Figure 6.24 A nine-channel dielectric Ku-band input multiplexer. (Courtesy of COMDEV International.)

(A) Single input bandpass filter frequency response

(B) Combined response of a two-channel input multiplexer
(odd channels on the same multiplexer)
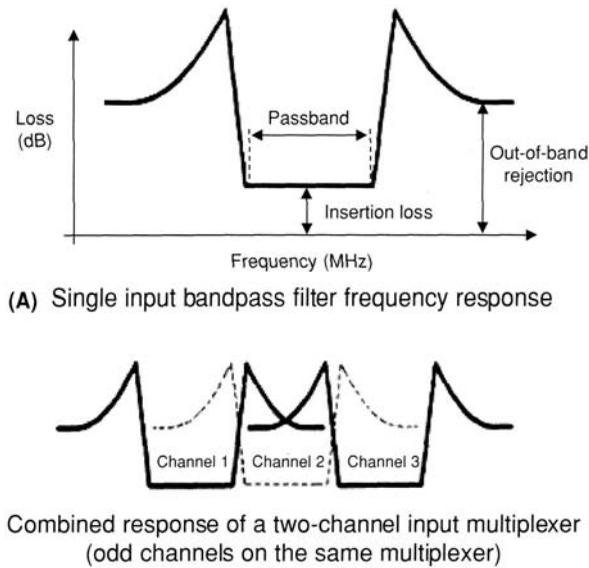
Figure 6.25   (a, b) Requirements for input multiplexer.

That variation appears as a rounding of the edges of the passband, a ripple within the passband, or a combination of both. The out-of-band portion of the characteristic shown in Figure 6.25 is more representative of real filters, where the loss increases rapidly starting at the edge of the passband until a point of maximum loss is reached. Beyond that point, the out-of-band loss can decrease but generally stays above some minimum level, which is called the out-of-band rejection.

A well-designed filter has low insertion loss, a reasonably flat passband, rapidly increasing attenuation at the edge of the passband, and high out-of-band rejection. Another important property is delay distortion (also called group delay), which was mentioned as an impairment in Chapter 5 and earlier in this chapter. Taken together, those characteristics represent tradeoffs in filter design and construction. For example, rapidly increasing attenuation at the edge of the passband requires that many filter sections (which are resonant cavities) be used, which increases both the insertion loss and the delay distortion. (See Figure 5.25 for examples of both characteristics for a real input filter.) A delay equalizer can be included to reduce delay distortion.

Combining several filters into a multiplexer implies the overlaying of their individual frequency responses, as shown at the bottom of Figure 6.25. The two filters in that ideal multiplexer are each tuned to different center frequencies corresponding to two transponders. The slight gap between the passband (the guardband) is necessary to prevent repeater multipath, which occurs when a signal on the edge of one channel finds a second path through the adjacent channel filter. If not adequately attenuated, repeater multipath can cause problems in the operation of the communications payload, such as tilted transponder passbands and frequency-selective valleys in the antenna coverage pattern.

Another problem is multiplexer spurious response, which is the leakage of power through a multiplexer on undesired frequencies. Again referring to Figure

6.25, the amount of loss at the point where the filter skirts cross over is important to the design of the multiplexer. The greater the loss at that point, the easier it is to match the actual performance of the multiplexer to theory. The loss at the cross-over can be significantly increased by splitting the channels into two groups, with each group on one multiplexer having every other channel. That is why there are odd-numbered frequency channels in the actual multiplexer design shown in Figure 6.23. One group would consist of the odd-numbered frequency channels (the darkened curves at the bottom of Figure 6.25), while the other would consist of the even-numbered channels (the dotted curve). That greatly improves the cross-over performance of the multiplexer because an entire empty channel now lies in the gap.

The actual filters used in the input multiplexer have evolved over the past 30 years. In the first generation, waveguide cavity filters were made of rectangular waveguide sections, which tended to be heavy and somewhat less stable over a typical spacecraft temperature range. Second-generation input filters were con-structed from coaxial structures, which are much more compact and can include temperature compensation. An example of such a filter is shown in Figure 6.25. Materials used in both cases evolved from Invar, a very stable metal alloy, to aluminum with temperature compensation, to even graphite-plastic composite in some cases. Another improvement is to reduce the group delay through delay equalization built into the filter itself.

The output multiplexer is constructed a little differently and without the use of a lossy circulator chain. Each filter is attached to a section of waveguide called a manifold. Without circulators to provide isolation, the filters must be positioned in accordance with wavelength along the manifold and manually adjusted (tuned) as an integrated system. A drawing of a circular waveguide cavity filter is shown in Figure 6.26. The corresponding block diagram of the multiplexer is shown in Figure 6.27. This approach is less effective against multipath and provides very little out-of-band filtering of intermodulation products or other undesired radiation. This situation is driven by the need for low insertion loss, which demands fewer filter sections and a more gentle slope of the out-of-band rejection.

Improvements in filter design have produced the contiguous multiplexer, which contains filters on adjacent channels (e.g., without the odd-even split). It has the
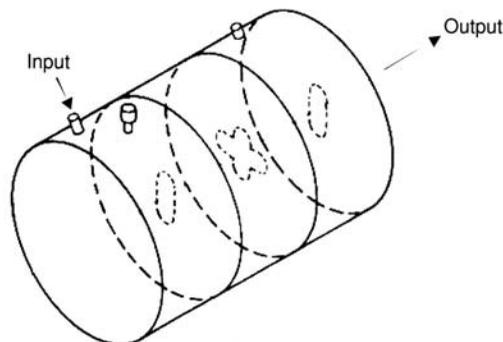
Figure 6.26   Schematic diagram of a circular-waveguide output filter.
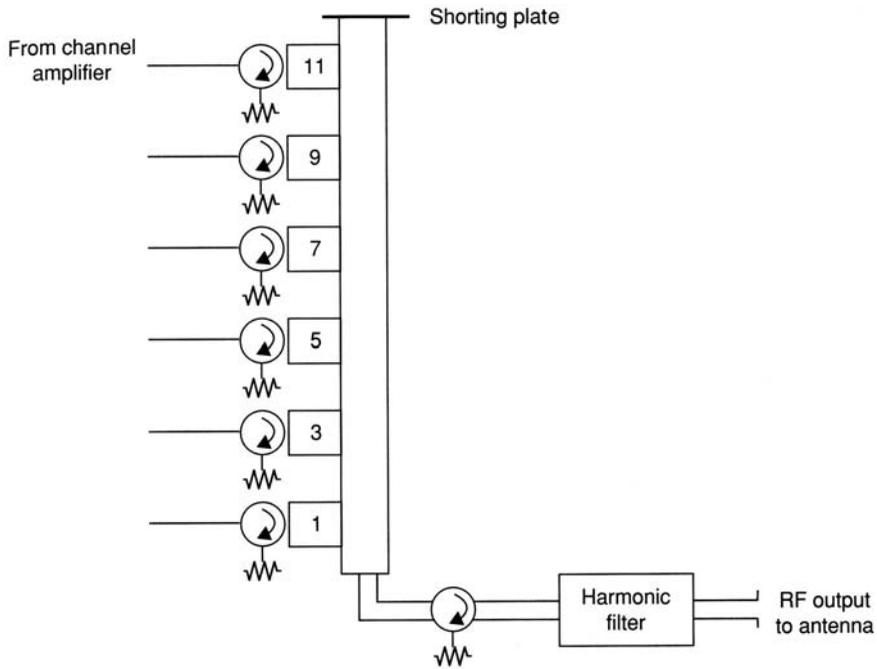
Figure 6.27 A six-channel output multiplexer configuration.

advantage that there is a single output port to connect to the antenna, as opposed to the two output schemes found with the odd-even split. On the other hand, the contiguous multiplexer has the disadvantage of limited control of the cross-over level and frequency response and so has a somewhat poorer overall response than the noncontiguous multiplexer. The performance advantages of the single connection to the antenna are covered in Chapter 7.

Accurate computer-aided design (CAD) tools and innovations in material selection have enabled satellite filter manufacturers to develop multichannel, compact, and thermally stable filters for output multiplexers. Conventional output multiplexers have Invar cylindrical cavity channel filters, mounted on a rectangular waveguide aluminum manifold. An example of a Ku-band 17-channel contiguous multiplexer with traditional design is shown in Figure 6.28. Another 17-channel contiguous multiplexer with similar performance specifications, but with channel filters made of temperature-compensated aluminum cavities, is shown in Figure 6.29. The modified temperature-compensated multiplexer is 40% lighter, occupies about 30% less footprint, and exhibits very stable thermal performance.

6.4.4 Traveling Wave Tube Amplifiers

While solid state electronics dominate the wireless telecommunications industry, there remains one application for vacuum tubes in space. The TWT continues to demonstrate its qualifications as a reliable, efficient, high-power device for use as a transmitter at microwave frequencies. The TWT is the tube itself, while the TWTA is the complete amplifier with its associated power supply. The two are
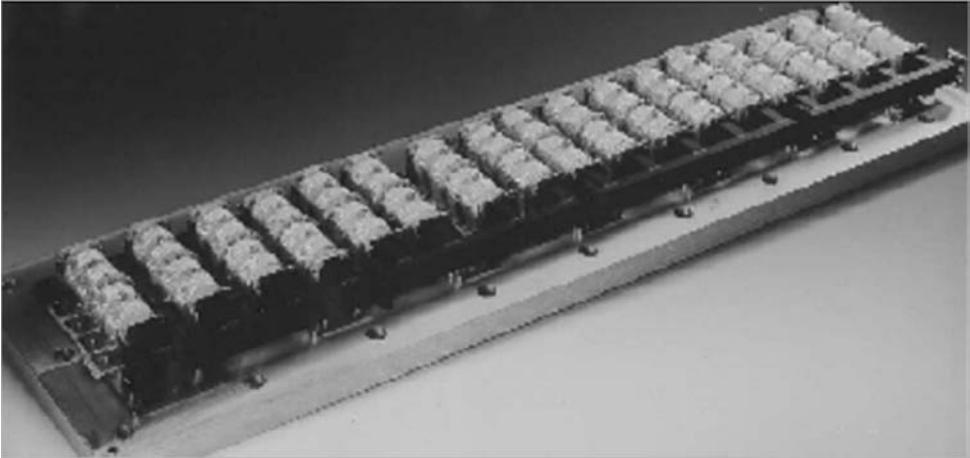
Figure 6.28   Conventional Invar Ku-band output multiplexer. (Courtesy of COMDEV International.)

Figure 6.29   Temperature compensated Ku-band output multiplexer. (Courtesy of COMDEV International.)

provided as a set because the necessary dc high voltages must be applied on an individual select basis. In some repeaters, two TWTs are paralleled and only one power supply is provided for the matched set. An alternative term for the power supply is the electrical power conditioner (EPC). As discussed later, the EPC must be built for longlife operation and efficiency of conversion from low-voltage to high-voltage dc.

A TWT is a physically linear vacuum tube that uses a narrow electron beam guided by a shaped magnetic field, shown in the schematic in Figure 6.30. Amplification is achieved by transferring energy from the beam to a microwave signal as it enters at the left end and exits at the right. The electron beam is generated by the gun (shown at the left), which consists of the heater, the cathode, and the anode. The heater raises the temperature of the cathode to cause it to emit electrons that are drawn off by the more positively charged anode. Like in any electron tube, the
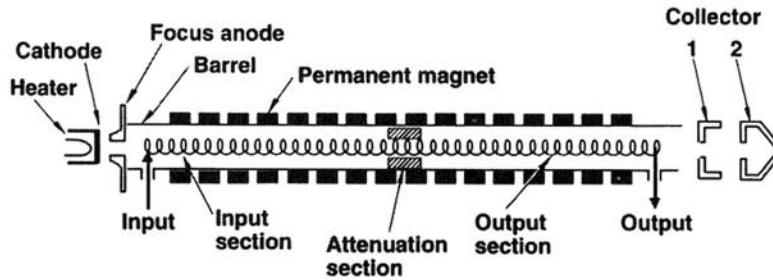
Figure 6.30 Schematic diagram of a helix TWT.

cathode is the principal life-determining component, and great pains are taken in its design and manufacture. TWTs use "dispenser" cathodes, impregnated with low-emissivity materials that slowly diffuse to the surface, where they enhance emission. Their temperature must be selected and maintained by the heater so the cathode emits electrons at the proper rate without exhausting the materials during the operating lifetime. A focus electrode surrounding the cathode aims and compresses the beam, much as a lens would. That reduces the stress and strain on the cathode itself yet produces the high beam current density needed for efficient TWT operation.

Immediately following the gun is the slow wave structure, which allows the beam to interact with the microwave signal under amplification. The low-level input signal enters through the port at the left and encounters a long, tightly wound coil called the helix. Microwave energy in free space travels at the speed of light, which is considerably in excess of the speed of the electrons in the beam. The microwaves, which are going around the spiral helix, have their axial velocity slowed down to a little less than that of the electron beam. The beam is kept tightly formed within the helix by a string of permanent ring magnets. Electromagnetic interaction of the microwave signal, which is a wave phenomenon, with the beam causes bunching of the electrons as they move along. Ultimately, that bunching transfers energy back to the microwave signal, adding considerable power to it. When the signal reaches the end of the tube, the electron beam has lost the majority of its energy of motion, while the microwave signal has become greatly amplified in power. The attenuation section at the center of the helix is an added feature to prevent self-oscillation and thereby stabilize TWT operation. The amplification (gain) of the TWT is roughly proportional to its length, which is easy to see because the interaction and transfer of energy increase as the beam and signal travel along the slow wave structure. Efficiency is enhanced by using a tapered helix, which means that the pitch of the helix coil is altered along its length. An output port at the end of the helix couples the microwave power through an isolator to the output filter.

The collector, located at the right end of the TWT in Figure 6.30, is critical to efficient operation. It is the objective of tube operation that energy be transferred from beam to signal, which implies that the velocity of the beam gradually reduces as it moves toward the output end. By the time the electrons reach the collector, they still possess considerable energy and are moving at less velocity than at the

gun. Efficient operation is achieved if the electrons are collected in such a way that their velocity is reduced to near zero before they actually hit the collector. That is accomplished by placing a negative voltage on certain collector stages to repel the electrons and thereby slow them down. The use of negative voltage for this purpose is called collector depression and is done with a multiple staged collector, as shown in Figure 6.30.

A typical high-power TWT can deliver over 70% of the dc input to RF output, which means nearly 30% of the dc power must be given up as heat. That can pose a challenge to the TWT and spacecraft designer, so proper measures must be taken. An option is to use a direct-radiating collector that is exposed directly to cold space. It turns out that about half the total heat dissipation of a TWT is at the collector end, so that is an excellent way of dealing with the problem. The rest of the heat must be conducted away through the base of the TWT, either through the equipment shelf or by radiating it away using a thermal doubler. Those aspects of thermal design are discussed in Chapter 8.

This discussion is by no means all inclusive, and an examination of the detailed design and operation of space TWTs is beyond the scope of this book. However, it should give the reader some idea of how a TWT works and generate an appreciation for the sophistication of the technology. Areas of critical importance include the materials and the manufacturing processing techniques used. Those aspects determine the physical properties of the TWT (weight and ruggedness) as well as the lifetime and reliability of the amplifier when operating in orbit.

It is important to say a few words about the power supply that must accompany the TWT for the purpose of providing the various voltages and currents. Examine the typical TWT voltages in Figure 6.31 to get an appreciation for the complexity of the design requirements. The circuitry to obtain the required voltages usually consists of an ac oscillator, high-voltage transformers, rectifiers and capacitor-type voltage multipliers. To provide stable operation of the TWTA, it is vital that the voltages be held constant under all conditions of TWT signal loading and unit temperature and throughout the lifetime of the repeater. Also, transient conditions
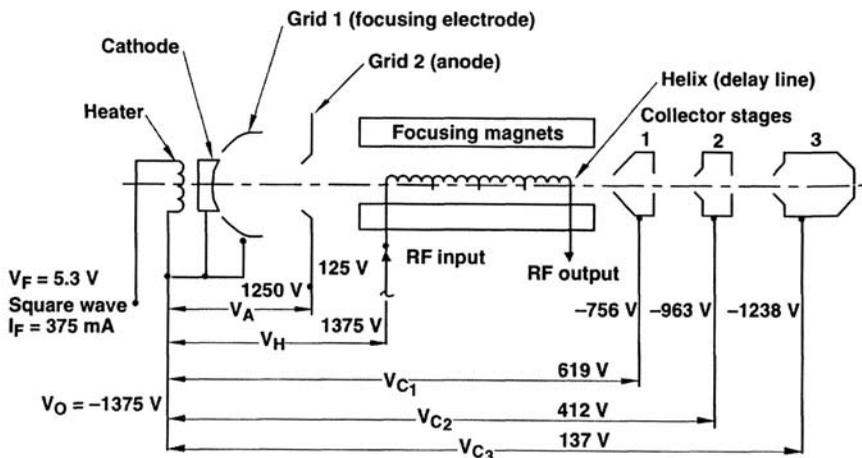
Figure 6.31   The dc voltage requirements for a C-band with a three-stage collector TWT.

on the power line coming into the power supply should not be transmitted to the TWT. A well-designed TWT power supply will meet stringent requirements and also have an efficiency (i.e., the ratio of power actually applied to the TWT to the dc input power from the spacecraft power system) in the range of 95–98%.

6.4.5 Solid-State Power Amplifiers

Historically, the TWT was the only active device on commercial satellites that was not solid state. That was because while transistors have always been found in the digital circuitry used to control the various bus and payload elements, the TWT was the only practical device that could efficiently amplify microwave signals to power levels in excess of 1W. Providing a lightweight power amplifier for the original Syncom with up to 30% efficiency was, in 1960, quite an accomplishment. Keep in mind that the context here is the operation of a microwave repeater in the space environment for many years.

The quest for an all-solid state amplifier led to the introduction of the first devices on the Telstar 3 and Satcom 4 series of satellites. Both of those satellites were intended to break through a nine-year lifetime barrier that seemed to exist due to the wearout properties of the TWT. (Of course, innovations in TWT cathodes and other aspects of tube design have overcome that impression.) SSPAs were introduced in the early 1980s at a relatively low power of 5W to 10W. Linearity of the amplifiers proved to be significantly better than that of a comparable TWT, so SSPAs were targeted to FDMA applications, where they could provide somewhat greater channel capacity. To make the SSPA practical, the key was the GaAsFET, which quickly established itself as a reliable and efficient power amplifier device and which is in use in a number of C-band satellites. SSPAs are also found on L- and S-band satellites used in MSS and DAB services. Ku- and Ka-bands SSPAs were developed for medium-power applications principally within phased arrays.

An SSPA is not a single amplifying device like a TWTA; rather, it is built up from individual amplifier modules. The basic building block, shown at the top of Figure 6.32, is the single-stage microwave transistor power amplifier. The triangle symbol represents a GaAsFET and its associated bias circuitry. Circulators at the input and the output provide considerable isolation (hence, the term isolator) from the rest of the SSPA and repeater. Note how the third port of the circulator is connected to a resistive dummy load, which catches any microwave energy reflected back into the circulator by the amplifier output. Without the isolators, the GaAsFET would tend to produce unpredictable results and may even damage itself.

The basic module can have from 6 to 10 dB of gain and a maximum power output of from 3–10W. Higher levels of gain or power out are obtained by combining stages, as illustrated in the rest of Figure 6.32. Gain is increased by adding stages in tandem, where the individual gains (in decibels) are simply added algebraically. That does not increase the maximum output power capability because all the signal still must come from the last stage; however, the input level can be reduced to a much lower level, corresponding to that delivered by the wideband receiver or digital processor. Achieving a total gain comparable to that of a TWT (roughly 50 dB) takes up to 10 stages in tandem. They usually are split into two

(A)   Single stage microwave transistor power amplifier

(B)   Two transistor stages in tandem

(C)   Two hybrid-coupled stages in parallel
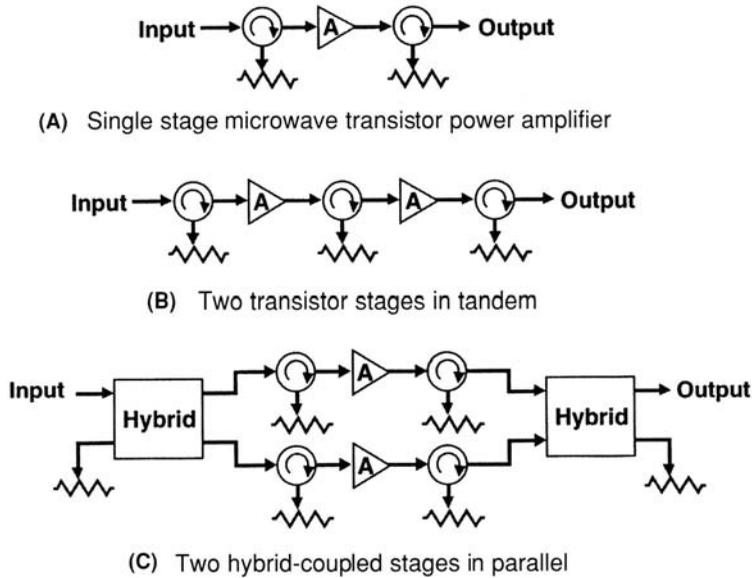
Figure 6.32   (a–c) Three versions of microwave transistor solid state power amplifiers.

different units for ease of integration into the overall repeater (discussed in the next section).

The method of paralleling stages, shown at the bottom of Figure 6.32, increases the maximum output capability in direct proportion to the number of stages in parallel. At the input, the signal power is split into two equal parts with a four-port hybrid. Note that the fourth port of the hybrid is not required and is therefore terminated, that is, connected to a passive load. Amplification is accomplished by a matched set of single-stage microwave transistor power amplifiers, whose outputs are combined by a second hybrid. The unique structure allows the power to be split and recombined efficiently, without more than about 0.25 dB of total loss at the output. The gain of the parallel combination is equal to that of a single stage, but it can output approximately twice the power since two amplifiers are working together. The key to successful operation is that the hybrids split and combine the signals with the proper phasing and that the individual GaAsFET stages be nearly identical in their characteristics. Not shown in the figure are the circuit components that allow a technician to adjust the balance and phasing of the signals to achieve the desired performance. With everything properly aligned, virtually all the output power will appear at the upper port of the hybrid on the right and none will appear at the terminated port.

The approach to further increase power is to treat the parallel stage as an amplifier module that can be further paralleled. For example, paralleling two modules gives nearly four times the power of the single transistor. However, it is easy to see that the achievement of high power and high gain comes at a price in SSPAs.

There are some unique aspects in using SSPAs instead of TWTAs. GaAsFET devices, being solid state in nature, do not contain a cathode and do not require high voltages to produce the electron beam. Unlike a TWT, which operates in a

hot condition for its entire lifetime, the delicate internal structure of the GaAsFET is sensitive to temperature and should be kept rather cool. The high field and power densitites inside a GaAsFET can limit life. Another consideration is that an SSPA is somewhat complex in terms of the number of internal microwave connections, which can increase the chance of failure of a random nature. For example, the solder joints or welds where the GaAsFETs are connected into their associated circuits can deteriorate and open up. On the performance side, the linearity of an SSPA generally tends to be superior to a TWT by itself. SSPAs can deliver 20–50% more FDMA channel capacity for the same output power level. TWTAs can be improved in terms of linearity at the expense of using a device called a linearizer (discussed in Section 6.4.6).

The areas where the TWTA continue to shine are in power output and efficiency. The current maximum practical output for a C-band SSPA is approximately 50W, while TWTs with powers in excess of 250W are being built. The efficiency of an SSPA tends to decrease as power is increased (because of the need to parallel stages and suffer losses through the hybrids that are required), while just the opposite is true for a TWTA. In some applications, TWTAs also have been paralleled because that provides greater output power for DTH services to a large footprint.

6.4.6 Transponder Gain Control and Linearization

While earlier communications satellites generally employed fixed gain repeaters, the trend has been toward providing each transponder with commandable gain control and linearization. The object of gain control is so the satellite operator can adjust the uplink segment for the desired overall effect on service quality. For example, if the uplink is coming from a large TV broadcast facility, the gain is set to minimum (e.g., maximum attenuation) to ensure both a high uplink C/N and a margin against uplink fade. In some instances, interference can be rejected by reducing transponder gain and raising uplink EIRP accordingly.

Alternatively, if the uplinks are coming from low-power VSATs, it is more desirable to increase gain (e.g., remove the attenuation) to allow the lowest possible transmit power from the ground. Networks of small VSATs need to allow ground SSPA power of 1W or less, which can place a considerable burden on the overall link. That is because there is a penalty in uplink C/N equal to the increase in gain provided at the input to the repeater power amplifier. There are three possible ways to compensate for that reduction: (1) directly compensate by increasing downlink C/N through the use of a large hub antenna to receive the signals; (2) provide more uplink EIRP through a larger Earth station antenna and/or HPA; or (3) employ a demod-remod style of repeater.

First-generation commandable attenuators consisted of one or more fixed steps of insertion loss, each providing a few decibels of attenuation. Those are placed on the input side of the TWTA or SSPA. An example of such a device is a ferrite switch with the third port connected to a passive attenuator with half the desired amount of loss. The opposite end of the attenuator is connected to a metal reflector element that bounces the microwave energy back through the loss element. Subsequently, the function has been incorporated into a channel driver amplifier (Figure 6.33) that provides adjustable positive gain to the input to the power amplifier.
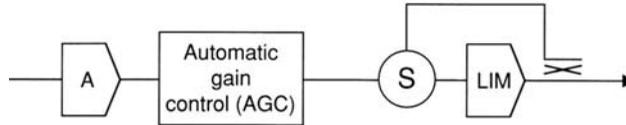
Figure 6.33   Driver-limiter-amplifier block diagram.

The device can be switched to an automatic gain control (AGC) mode, which provides a constant drive as the uplink varies over a range of 10 dB or so. A constant drive level is only appropriate for single carrier operation, however. AGC also can be provided with a limiter that effectively clips the input signal and holds it constant. A consideration with a limiter is that the input must not have significant amplitude variations. That is because the limiter is a highly nonlinear device that would produce undesirable intermodulation products or sidebands. The AGC can be substituted with a commandable fixed gain step for use in multicarrier transponder mode.

The service quality of a bent-pipe transponder can be improved through the introduction of a linearizer. By designing the circuit of the driver amplifier to compensate for the nonlinear property of the SSPA or TWTA, it is possible to provide a more linear overall service for applications involving digital transmission or FDMA. An example of the improvement in IMD from a typical linearizer is shown in Figure 6.34. The linearizer must be properly adjusted to cancel a portion of the nonlinearity and, at the same time, to provide good overall performance for any type of signal. Because that may not be feasible under all situations, the operator would have the option of switching the linearizer out of the path.
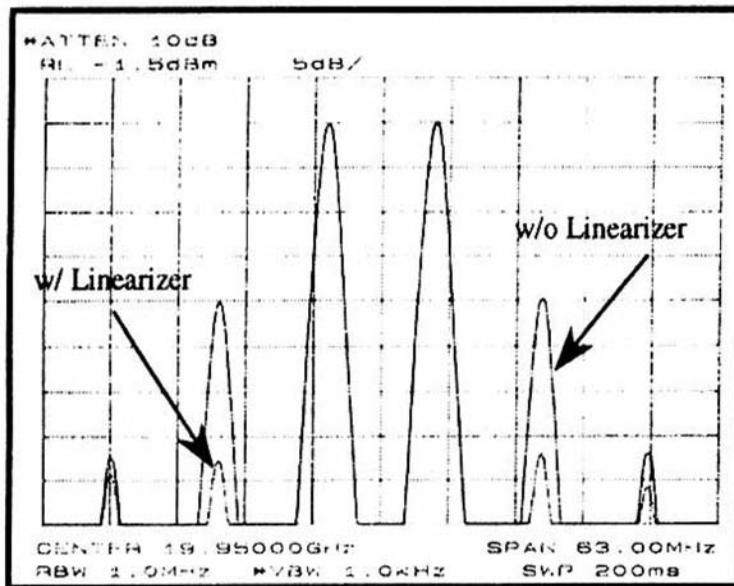
Figure 6.34   Ka-band transponder linearizer IM. (*From:* [4]. Reprinted by permission of the American Institute of Aeronautics and Astronautics, Inc.)

We have described the elements of the repeater section of typical communication payloads. These devices support analog and digital services and can be used with or without a digital processor. In Chapter 7, we review the design and performance of the remaining payload element, the spacecraft antenna system.

References

[1] Lyons, R. G., *Understanding Digital Signal Processing*, Reading, MA: Addison-Wesley Longman, 1997.

[2] Craig, A. D., and F. A. Petz, "Payload Digital Processor Hardware Demonstration for Future Mobile and Personal Communication Systems," in E. Biglieri and M. Luise, (eds.), *Signal Processing in Telecommunications*, London: Springer-Verlag, 1996.

[3] Bever, M., et al., "Broadband Payloads for the Emerging Ka-Band Market," TRW Space and Electronics.

[4] Katz, A., et al., "Wideband TWTA Linearized Driver Amplifier for Ka-Band Satellites," AIAA 2000 ICSSC.

CHAPTER 7

Spacecraft Antennas

The spacecraft antenna system, which is often viewed as part of the communications payload, has many special and important characteristics and therefore is deserving of its own chapter. Modern antenna systems provide a variety of communication coverage patterns such as domestic, regional, and spot beams, allowing the repeater to be utilized in an efficient and flexible manner. In the classical bent-pipe repeater design, the antenna coverage typically is fixed prior to launch. Changes on orbit are limited to what can be physically switched within the repeater. Onboard digital processing provides more flexibility because beams can be reshaped and even created dynamically in accordance with demand. The use of such processing is revolutionizing antenna system design because of the much wider range of performance options that can be obtained. Still, microwave antennas have not changed fundamentally since the bulk of the initial R&D was performed nearly 70 years ago.

This chapter explains how typical antennas are physically constructed and how they create the appropriate coverage patterns. An exact description of antenna operation would require use of electromagnetic theory, which (while mentioned in Chapter 4) is beyond the scope of this book. However, the somewhat simplified explanations that follow should provide a practical understanding of spacecraft antennas. Systems engineers need to develop requirements for payload antennas in cooperation with antenna experts who understand the details of the theory, the appropriate software tools, and the means to implement and test the physical hardware. Those are the stock and trade of the antenna system designer and are beyond the scope of this book. After the antenna electrical design is complete, there are still the critical matters of the choice of lightweight materials that provide sufficient strength, conductivity, and thermal control properties. Finally, the antenna must be physically attached to the spacecraft and adequate means provided for stowage and deployment in orbit. Pointing of the antenna beams, discussed briefly in Chapter 6, is another important aspect of the overall RF performance of the payload.

7.1 Horn Antennas

The basic microwave properties of waveguide were introduced in Chapter 6. Waveguides function in satellite communications as either a component of an antenna or as an antenna in its own right. If we were to take a fixed-length piece of

waveguide and use a quarter-wave monopole to drive microwave energy at the proper frequency into one end, the opposite end would radiate some percentage of that power into space. The efficiency of radiation is simply the ratio of the power actually radiated into space divided by the total input power. Power not radiated is reflected back to the transmitter or dissipated as heat within the conducting metal of the waveguide.

What this open-ended piece of waveguide has become is the simplest type of nonreflector antenna horn. Figure 7.1 shows a pyramidal horn in which the end of the waveguide is flared outward in the shape of a pyramid or a cone. The dimensions of the opening are in direct proportion to the wavelength and are dictated by the shaping of the far field antenna pattern. Also, the ratio of opening to waveguide dimensions directly affects the match between the somewhat variable and greater impedance of the waveguide to that of free space, which is approximately 377Ω. That match is needed to reduce the power reflected back toward the transmitter. For a receiving antenna, a good match ensures that the received signal energy is delivered to the repeater input port. It is that match, measured in terms of the voltage standing wave ratio (VSWR), that defines the fundamental electrical property of the antenna when connected to either the input or the output of the repeater. The ideal condition of zero reflection occurs at VSWR = 1:1 (one to one), which cannot be achieved in practice across the entire operating band. However, it is not uncommon to obtain a maximum-value VSWR of about 1.25:1 to 1.50:1, with 1.00:1 being measurable at certain frequencies. Of course, that does not define the actual radiation pattern, which is discussed below in more detail.

Examples of different types of horns are provided in Figure 7.2. Selection of the pyramidal shape over the conical one is based on the particular application. The pyramidal horn is well suited to linear polarized systems because of the natural straightness of extending rectangular waveguide. The standard waveguide mode of propagation is maintained, resulting in a symmetrical main-lobe pattern and little or no cross-polarized component. In contrast, the conical horn flares from either circular or square waveguide (in the latter case, there must be a square-to-circular transition) into a circular aperture. This type of horn is best suited to dual LP or CP, while the LP performance of the rectangular horn is superior.

Standard pyramidal and conical horns have smooth interior walls, as is the case with waveguides. That minimizes losses and reduces the generation of higher order modes that affect both efficiency and polarization performance. (We will see
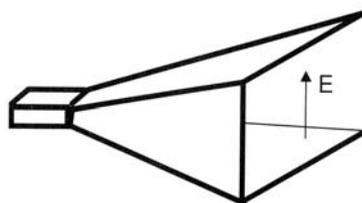
Figure 7.1   A common type of horn antenna called the pyramidal horn. The opening is wider than the waveguide to increase gain and to match the impedance of the waveguide (~400Ω) to free space (377Ω).
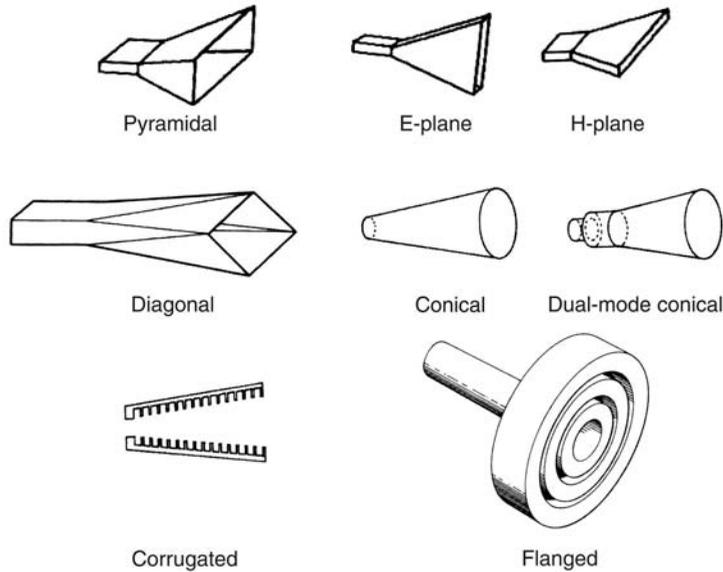
Figure 7.2   A selection of various microwave horn geometries.

that another type of horn, called the corrugated horn, intentionally generates higher order modes to improve pattern and isolation performance.) Either type of horn can be machined ("hogged") out of a solid block of copper or aluminum or assembled from flat stock using welding techniques. The latter method is somewhat prone to higher loss and the possible generation of passive intermodulation (PIM) products, discussed later in this chapter. Another concern relates to high-power transmit service, as on a DBS satellite. Imperfections and sharp edges inside a high-power OMUX or transmit horn can generate multipaction, a form of high-temperature arcing of waveguide internal structure that occurs in vacuum.

   More precise horn geometry and quality can be obtained through the process of electroforming, which, by its nature, is considerably more costly and time-consuming than the previously described methods. In electroforming, the horn is electroplated onto a reverse mold, called a mandrel. The mandrel metal material has a lower melting point than the electroplating. After the horn is electroplated and the exterior machined to the desired shape, the interior mandrel simply is melted out. The resulting horn is as physically perfect as one can produce, based on the tolerances used in machining the mandrel itself.

   Horn antennas are used extensively on commercial satellites. An early example is the Earth (global) coverage horn, requiring a beamwidth of approximately 17 degrees. The horn produces maximum power toward the subsatellite point and tapers downward to become about 3 dB below the peak at the edge of the Earth. An example of such a coverage pattern is shown in Figure 7.3, in which the upper (wide) curve applies for ±10 degrees (the upper degree scale on the X-axis) and the lower (sharp) curve applies for ±60 degrees (the lower degree scale). Antennas of this type are limited because the peak gain is only about 20 dB, resulting in relatively low G/T and EIRP performance.

   The most common use of the horn is as a radiating element for a reflector type of antenna. (Section 7.2 reviews some of the design and performance aspects of

Close in angle, degrees

Gain, dBi

Wide angle, degrees
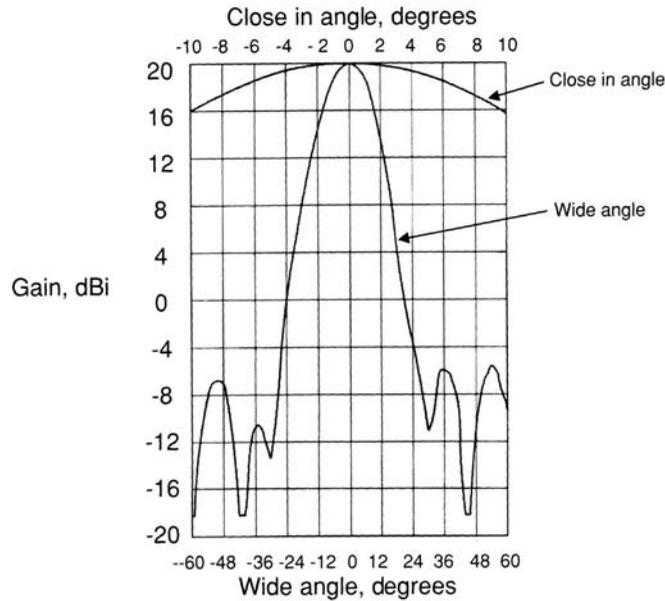
Figure 7.3   Antenna patterns of global coverage horn on a GEO satellite.

current reflector antenna systems.) The horn is placed at or near the focus of a parabolic reflector to illuminate the surface, in either a transmit or receive sense, depending on the application. As illustrated in Figure 7.4, the electromagnetic radiation on the reflector surface produces electrical currents on the surface. From those currents, another electromagnetic field is produced that eventually becomes the far field radiation pattern of the entire antenna system.

In this example the feed horn is transmitting vertical polarization and yet the reradiation from the reflector has both vertical and horizontal components. The degree of horizontal polarization depends on where the induced currents happen to occur on the surface. Essentially pure vertical polarization lies along the dish's
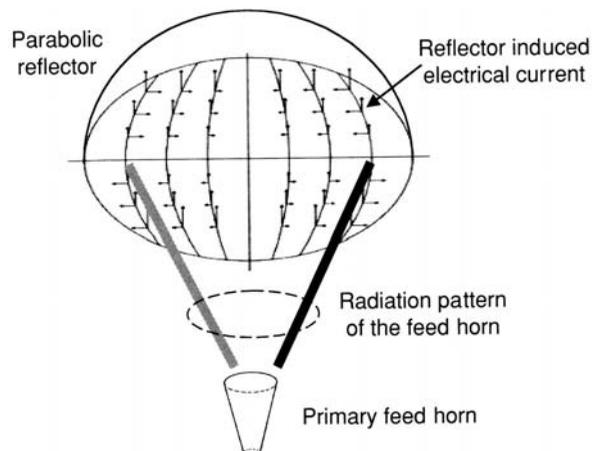
Figure 7.4   Principles of operation of a parabolic reflector microwave antenna.

centerline; for any horizontal slice through the reflector, the horizontal components on either side cancel each other at some forward distance. A horizontally polarized feedhorn would achieve the same result with respect to that sense of polarization.

The other way to view reflector operation is in terms of an optical model, in which the horn produces rays of energy that are reflected off the reflector surface to form a collimated beam. As discussed in Chapter 4, the beam does not actually materialize until the far field of the antenna is reached. The horn becomes the primary radiator of the reflector and therefore must be designed to introduce the proper radiation of microwave energy across the reflector. Also, the impedance match between horn and waveguide, which controls the amount of reflected power, is corrected by adjusting horn dimensions and spacing from the reflector. For those reasons, open-ended waveguide seldom, if ever, is used as a primary radiating element. More popular are pyramidal horns and conical horns, which are applied to LP and CP antennas, respectively. (There are exceptions when a square cross-section can support dual polarized CP and, conversely, a conical horn is used in dual LP systems.)

Returning to the horn itself, its ability to produce the proper primary radiation pattern is determined by its size, shape, and internal structure (which interacts directly with the microwave energy in the waveguide). Other important factors in horn design are cross-polarization purity (and hence isolation) and sidelobe performance. To raise performance in those categories, the corrugated horn (Figure 7.5) was introduced. The desirable radiation properties of the corrugated horn include good pattern symmetry (along the vertical and horizontal dimensions, across the pattern itself), low sidelobe levels, and high beam efficiency (approaching 95%). However, the design and manufacture of the corrugated horn are more complicated than for smooth horns, which is expected due to the intricacies of the internal structure. Electroforming is the norm, rather than the exception, in this case.

In basic terms, the indentations allow two simultaneous modes (transverse electronic, or TE and transverse magnetic, or TM) to propagate up the horn to its opening. At the same time, the depth and the spacing of the corrugations cancel energy in the other propagation modes that could produce undesired properties
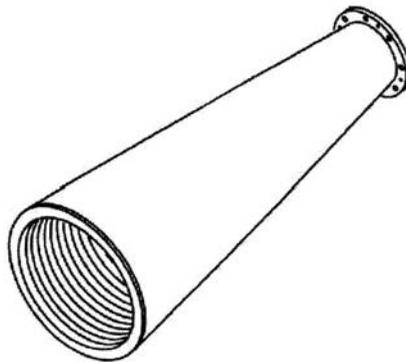
Figure 7.5   A C- or Ku-band corrugated horn.

like cross-polarization and high sidelobes. Furthermore, the pattern that the horn produces is symmetrical, which means that it is as nearly circular as can be achieved in practice. That lends itself to applications using CP with low sidelobes.

Although today the prominent technical solution is single feed horns illuminating shaped reflectors, horns often are combined into a feed array and used with a reflector. Figure 7.6 provides an example of a multiple-feedhorn array for frequency reuse through Ka-band spot beams. Antenna arrays of this type are used for narrowband L-band MSS cells that reuse spectrum and provide increased receive and transmit gain. Another application is for Ku-band wideband spot beams to transmit high data rate digital carriers. There, the feedhorns are not connected to a power divider but are used separately. Important considerations in both of those horn arrays are the physical location and the proximity of the individual horns. The radiation pattern of one of the horns is determined by the size of its opening. However, that size still must allow for the required quantity of feeds to be placed within the prime focus of the reflector antenna. If that is not physically possible, then smaller, less effective horns may have to be used. Another solution that is available only in multiple spot beam antennas is to split the horns into separate groups that illuminate different reflectors. This is illustrated in Figure 7.7 for the 48-beam antenna "farm" on the Astrolink satellite.

The last application of horns is a part of a direct radiating array antenna system, such as that shown in Figure 2.17 for WGS. As in the feedhorn array for a reflector antenna, the direct radiating array uses multiple radiating elements, such as horns or crossed dipoles to produce a combined pattern or set of patterns. A group of elements are fed the same signal but are individually adjusted in amplitude and phase. The result can be a far field pattern, with remarkably high efficiency and gain, and low sidelobe levels.

A final application is as a microwave lens antenna, in which the input end of each waveguide literally is left in the air to collect energy from a single primary radiating horn. As illustrated in Figure 7.8, the horn array refocuses the energy into a precisely defined beam, aligned to the direction of propagation. The lens also can support multiple beams at the same time, provided there is a proper primary radiating antenna system.
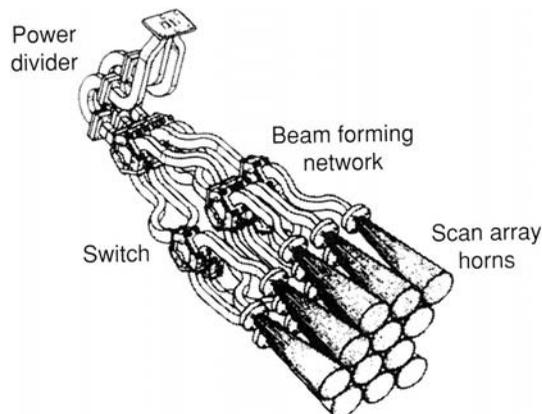
Figure 7.6   Multibeam feed array for NASA ACTS.

Figure 7.7 The Astrolink satellite design incorporated eight reflector systems, each with up to eight individual feedhorns to produce a like number of spot beams [1].
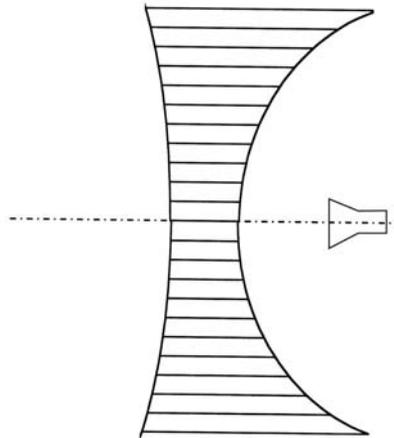
Figure 7.8 The basic configuration of a waveguide lens antenna.

7.2 Reflector Antennas

The simplest and most effective antenna system for a conventional FSS and BSS satellite employs reflectors and feed systems. That is evident from looking at the typical GEO satellite, such as that shown in Figure 7.9, with its reflectors being among the most obvious physical appendages. The feed system normally consists of one or more horn-type radiators. In the transmit mode, microwave energy enters the feedhorn from a waveguide that carries the output of the repeater. The feedhorn
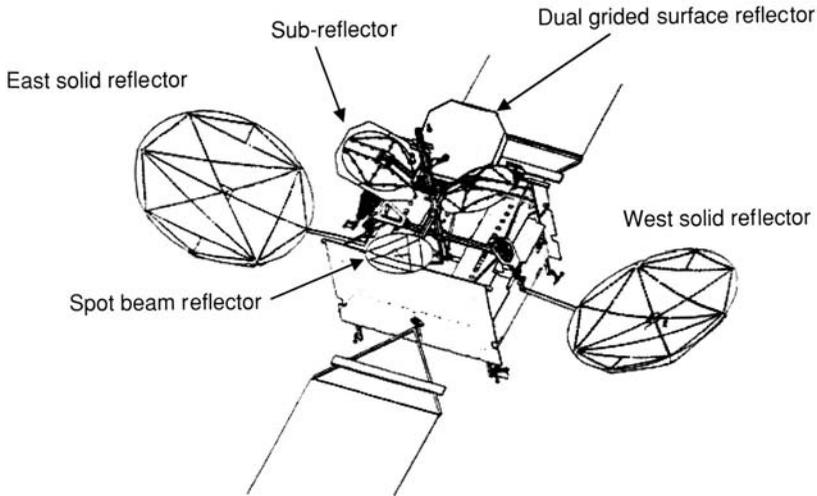
Figure 7.9   The general arrangement of a three-axis spacecraft with two offset parabolic reflector antennas. A subreflector attached to the front of the body extends the focal length (referred to as Gregorian optics).

radiates microwave energy from the focus of the parabolic reflector into space in the direction of the reflector (or subreflector, as the case might be).

A nice feature of such passive antenna designs is that a single reflector with feed can function for transmitting, receiving, or both simultaneously. This property, called reciprocity, was mentioned in Chapter 4. It is an interesting curiosity of history that this property was discovered by Hertz in 1888, well ahead of any knowledge of microwave propagation and only when the famous genius was experimenting with spark-gap transmitters and diode detectors. Much of the real work on microwave antennas, particularly that using reflectors, was carried out in the west during World War II. The results of that vital research are available in a classic text [2], which remains a good reference on microwave antennas.

Two typical antenna geometries using parabolic reflectors are shown in Figure 7.10. The center-fed parabola is circularly symmetric with the feed located at the focus, while the offset reflector parabola allows the feed to be placed below the line of transmission. The reflector surface is formed by taking a parabola and rotating it about a line drawn from the focus to the center of the parabola (forming a segment of a parabaloid). For the offset case, the required surface is cut from one side of the parabaloid (see Figure 7.17 later in this chapter for an example.) This type of antenna continues to be popular for spacecraft and ground applications because it is straightforward to design and build. Also, extremely cheap versions, which are still effective, have been produced for DTH reception around the world. The spacecraft version is not cheap because of the need to use lightweight materials for the reflector and support structure. They must be rigid to hold the beam properly and must not undergo significant changes in dimensions under the varying temperatures of space. In many cases, the reflectors must be stowed for launch and then reliably deployed in space prior to start of operation.

The size of the reflector is dictated by the required gain, as discussed in Chapter 4. Beamwidth, which is equally important in payload performance, varies inversely

(A)   Prime focus fed parabolic reflector antenna

(B)   Offset focus fed parabolic reflector antenna
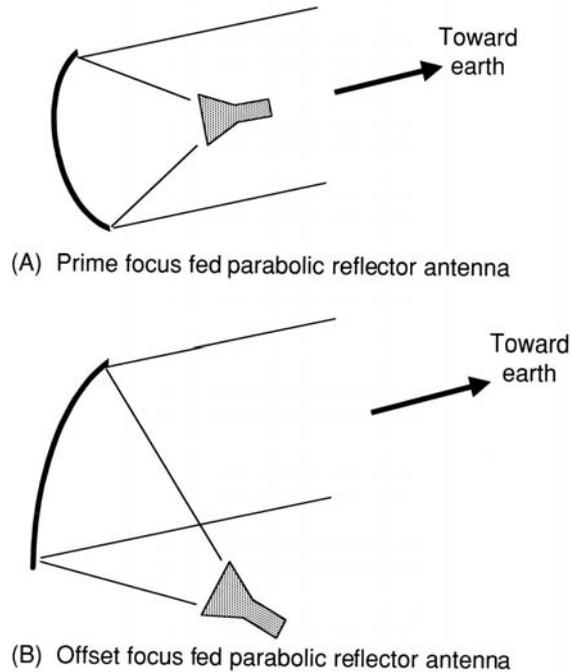
Figure 7.10   (a) The prime focus-fed parabolic reflector antenna is subject to feed blockage, (b) while the offset feed design provides an unobstructed beam.

with diameter. That provides the primary basis for determining the physical size of the reflector. Another factor is the actual frequency of operation, which must be sufficiently high to permit the reflector to perform its function and to allow the use of a reasonably sized feedhorn. The horn opening, which determines the radiation pattern across the reflector, also is dictated by the frequency of operation. Low frequencies mean a large opening. If the horn opening is too large, it could block the energy after it leaves the reflector. That is why there is a preference for the offset geometry, since blockage is avoided.

Reflector antennas may be designed with two reflecting surfaces, as illustrated in Figure 7.11. There are technical and performance advantages to using two reflectors, the main one being that it allows the feed to be located behind the main reflector instead of ahead of it (and potentially in the line of propagation). In ground antennas, the Cassegrain approach, popular in optical telescopes, allows the feed, LNA, and potentially the RF electronics as well to be placed behind the reflector in a protective housing or shelter. A spacecraft version of the Gregorian type has been used on FSS and BSS spacecraft (Figure 7.9) to increase the focal length and permit more convenient placement of the feed in close proximity to the output multiplexer.

Another technical approach to reflector antenna design can provide a shaped beam with a single feedhorn. It is the shaped reflector, a technology from Japan that became an industry standard in the mid-1990s for C- and Ku-band bent-pipe satellites. As shown in Figure 7.12, the reflector is contoured in such a way as to induce the appropriate current density and phasing on the reflector surface. That
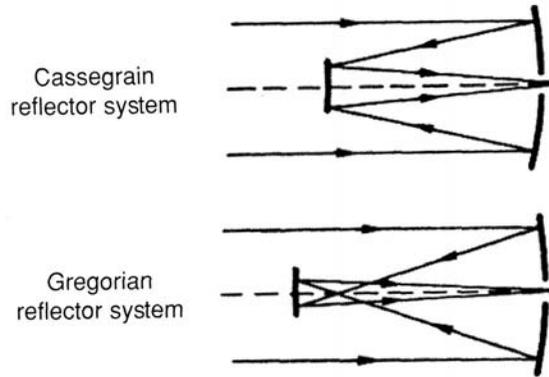
Figure 7.11   Dual reflector geometries: the Cassegrain and the Gregorian.
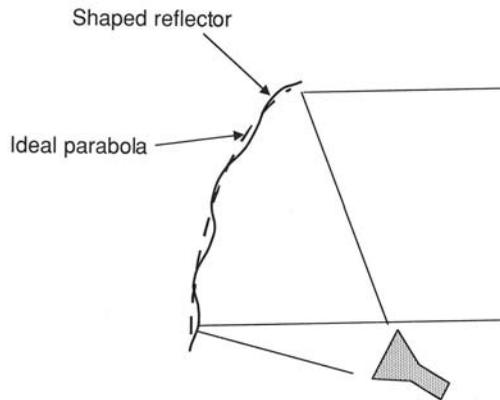
Figure 7.12   Conceptual drawing, in cross-section, of a shaped offset reflector.

creates a shaped beam in the secondary pattern. To build the shaped reflector, precise dimensions are machined in a mandrel, which is the metal mold onto which resin-impregnated carbon fiber is laid out. After curing, the carbon fiber surface is lifted off the mandrel and fabricated into the reflector assembly.

We begin our discussion of the two configurations with a review of the center-fed parabola, which is the basis for all classes of reflector antennas and the simplest to design and construct. Following this, we consider the offset-reflector antenna, now the most common configuration.

7.2.1   Center-Fed Parabolic Reflectors

The prime focus approach, illustrated in Figure 7.13, is efficient and effective because of the ideal geometry that it possesses; however, it has the drawback that the feedhorn (or system of feedhorns) can block and deflect a portion of the collimated beam. That is particularly troublesome in frequency reuse spacecraft antennas in which feedhorn blockages cause high sidelobes, scatter energy outside the desired coverage, and degrade cross-polarization isolation. The effect of
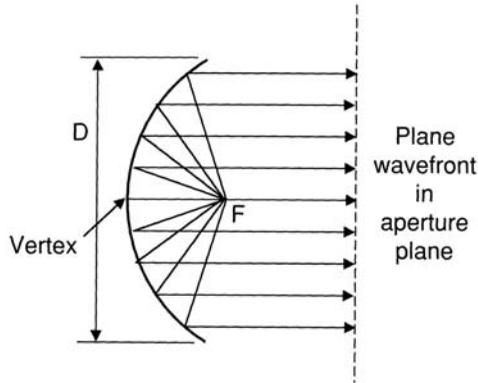
Figure 7.13 Geometry of prime-focus fed paraboloidal reflector.

blockage on sidelobe level for a hypothetical center-fed parabola is graphed in Figure 7.14 (it is approximate and is shown for illustrative purposes only).

The typical parabolic reflector is round to provide symmetry for a curricular beam. Other shapes are possible and permit the beam to have a noncircular cross-section as well. For example, an oval-shaped dish produces an oval-shaped beam, except that the major (longer) axes of the two are at right angles. The reason is that the wider dimension produces the narrower beamwidth (due to better concentration of the beam) and vice versa.

The reflector surface is smooth to provide the cleanest output wavefront. There is a relationship between the surface accuracy, called the surface tolerance, and the resulting gain, as provided by the following [3]:
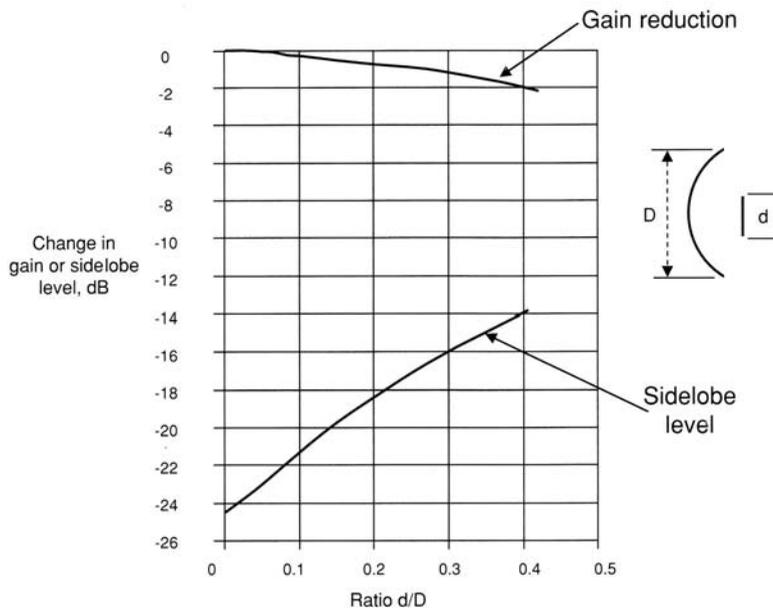
Figure 7.14 Aperture blockage effect on sidelobe level and gain in a center-fed parabola.

$$G/G_0 = \exp[-(4\pi\epsilon/\lambda)^2]$$

where

G = antenna gain with error;

G_0 = antenna gain without error (determined by the diameter, frequency, and so forth);

ϵ = surface (root-mean-square, rms) tolerance in the same units as λ;

λ = the wavelength.

The equation is an approximation that represents the average gain loss over a large number of antennas having the same physical properties. An individual antenna would have its own unique error profile and a corresponding gain loss, which would be larger or smaller than the formula would predict. For example, at an average gain loss of 1 dB, the rms surface tolerance would have to be only 1/30th the wavelength. That allows large L-band reflectors to be somewhat imperfect in absolute terms, since the required surface accuracy is relatively easy to achieve. At Ka-band, the problem is worsened by the ratio of wavelengths of more than 12 to 1, where $\lambda/30$ is one half of a millimeter.

One interesting approach for a center-fed parabolic antenna, shown in Figure 7.15, is to package a very large deployable center-fed reflector constructed like an umbrella. That allows stowage as a long post and deployment by extending the umbrella substructure. A subreflector is shown in this configuration to create a Cassegrain system with the feeds attached to the spacecraft body. The main surface normally would be made of a flexible metallic mesh or metal-covered fabric. Issues still remain regarding that design because of the potential for PIM products that are generated from high-power transmission. PIM, in turn, can radiate and conduct back into the receiving side of the spacecraft, interfering with proper operation of the payload. Control of PIM requires careful attention to the design of the mesh
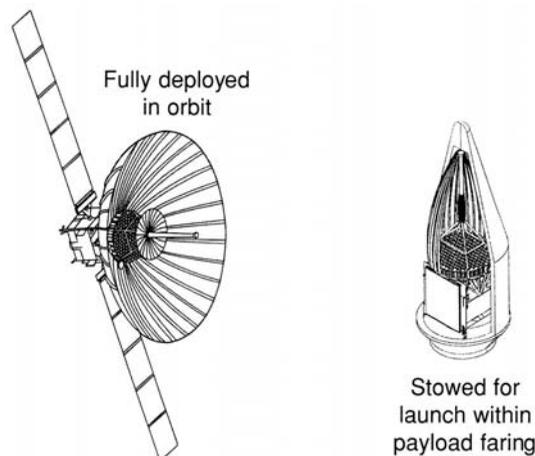
Figure 7.15   An example of a large deployable center-fed parabola for use at L-band.

or fabric and good consideration of the overall payload, the frequency plan and system design to avoid PIM from causing interference in the first place.

7.2.2   Offset-Fed Parabolic Reflectors

The offset-fed parabola, illustrated at the bottom of Figure 7.10, eliminates feed blockage and substantially reduces reflection of energy from the reflector back into the feed. While the physical characteristics have less symmetry and require more careful design and manufacture, experience has produced very efficient and high-performing spacecraft antennas of this class. The offset reflector permits placement of feeds much closer to the repeater electronics and suppresses sidelobe levels that would result from feed blockage. Subsequently, nearly all bent-pipe type satellites adopted the offset reflector type of design, taking advantage of the fact that the reflector can be stowed for launch by simply folding it directly on top of the spacecraft, as illustrated in Figure 7.16.

The specific dimensional characteristics of the offset reflector are shown in Figure 7.17. There are a few potential technical drawbacks of the offset approach, the majority of which can be overcome through certain design optimizations and device selections. The first is that the offset parabola experiences some efficiency
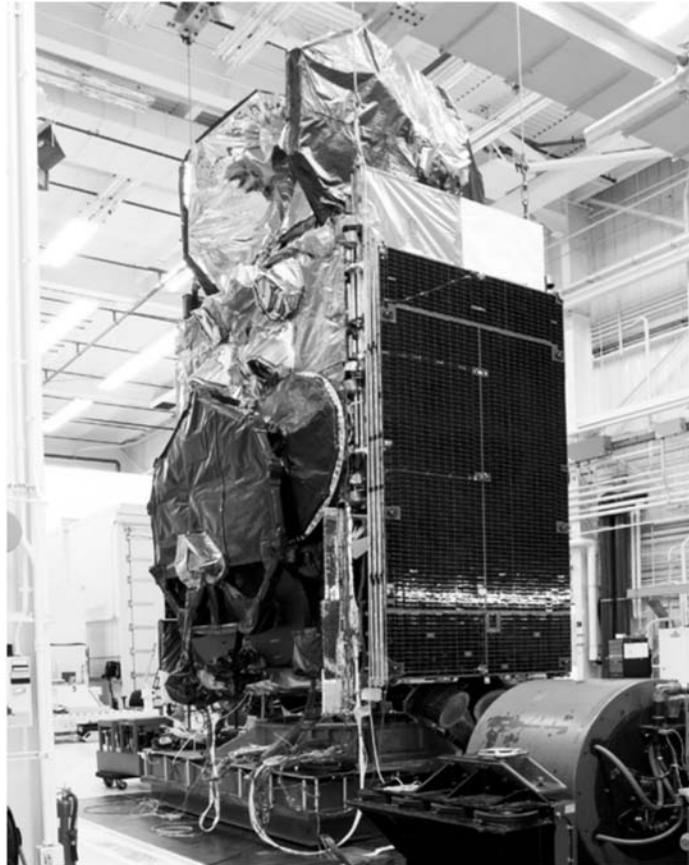
Figure 7.16   Stowed offset reflector antennas for Anik F2. (*Source:* Boeing.)
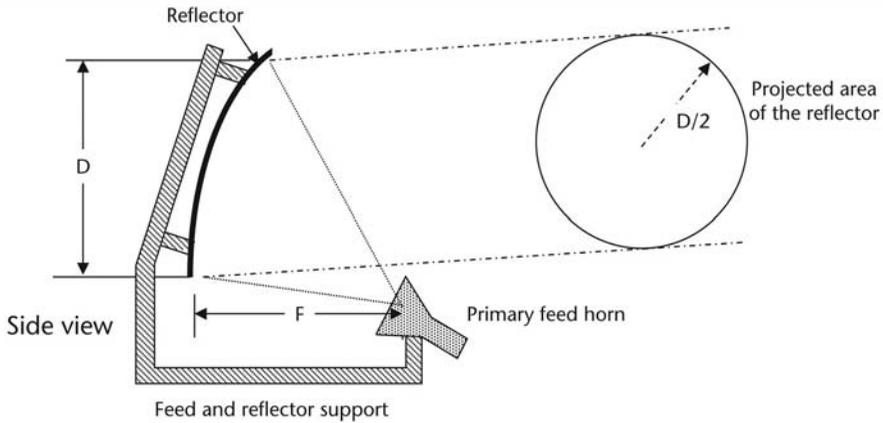
Figure 7.17 Offset-fed parabolic reflector with feed and feed support located so as not to produce blockage.

loss because the beam is tilted slightly upward from optimum. An illustrative example of how gain performance would degrade as the beam is offset from boresite is plotted in Figure 7.18. Gain reduction, which increases more rapidly as one moves away from boresite by multiples of the beamwidth, is the result of beam defocusing. For example, moving a 1-degree beam by 4 degrees causes about a 25% loss of gain for F/D = 0.35. This is equivalent to 1 dB of gain reduction. Any such gain loss usually is countered, to some degree, by the complete elimination of feed blockage. The second effect is on sidelobe levels and cross-polarization, which are different depending on which direction one looks from boresite. Some of that can be controlled through proper selection of feed type and dimensions, including the use of more polarization-pure designs like the corrugated horn. Another concern is cross polarization isolation, which could be degraded by the
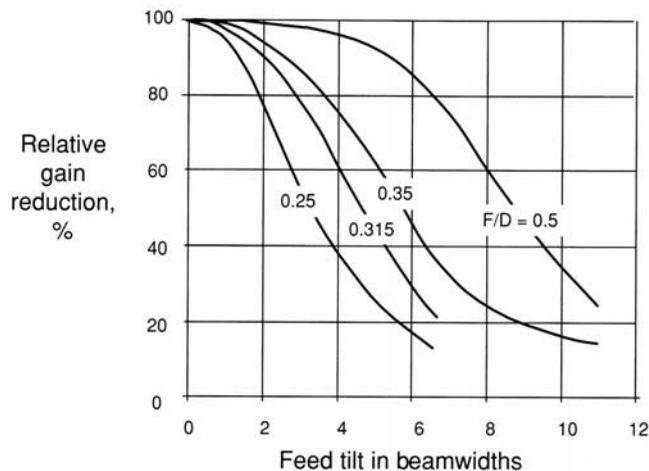
Figure 7.18 Gain degradation due to beam offset. F/D is the ratio of focal length to diameter, a value of 0.35 being typical.

lack of circular symmetry of the antenna system. That can be controlled by proper selection of the feedhorn and its placement relative to the reflector.

Finally, in CP offset reflector antennas, there is a slight effect on "squint" of the beam to one side or the other. The direction depends on which sense of CP is used: left-hand or right-hand. This has an impact on the deployment of transponders on particular satellites and could result in using twice the number of reflectors.

Another important consideration is the deformation of the reflector surface and feed support from differential solar heating and the resulting thermal expansion. That misshapes the reflector, causing the pattern to distort. Deformation of the feed support relative to the reflector causes the beam to mispoint. Two solutions are available: make the antenna less sensitive to thermal effects (e.g., prevent temperature variation) or provide a mechanism to continuously realign the reflector on orbit.

7.3   Antenna Patterns

For most spacecraft antenna systems, the feedhorn works with a reflector of some sort to produce the desired coverage pattern. At the top of Figure 7.19, a reflector is illuminated with microwave energy by a single feedhorn located at the focus. The feedhorn has a radiation pattern (indicated as the shaded area in front of the reflector), which has a maximum along the centerline and decreases in intensity toward the edge of the reflector. The specific manner in which the intensity is adjusted is called the taper. As you will recall, the primary illumination produces a corresponding distribution of electrical currents across the reflector surface. The currents reradiate their energy toward the front of the reflector, resulting in the desired secondary pattern in the far field of the antenna. Energy not radiated

Figure 7.19   Illumination of a parabolic reflector by a waveguide horn producing a secondary pattern, which contains a main beam and sidelobes.

is dissipated through the electrical resistance of the reflector metal (or metallic) surface.

The far-field pattern of a reflector antenna system is a two-dimensional Fourier transform of the horn illumination. That statement has meaning to electrical engineers and points out that reflector antennas share a mathematical property with pulse spectra. An example of a secondary pattern is shown at the bottom of Figure 7.19. The main lobe in the center provides the desired coverage, a concept introduced in Chapter 1. Visualize that the peak of the lobe is pointed at the center of a country to be covered. The main lobe extends across an angular dimension, Θ_{3dB}, which roughly corresponds to the width of that country. The parameter Θ_{3dB} is the 3-dB beamwidth, defining the point where the gain of the antenna is 3 dB down (i.e., one-half the power) from the peak. That natural characteristic of an antenna beam explains why reception from domestic satellites generally is better in the center of the country than at the borders. Real antenna patterns are shaped to match the desired service area.

The secondary pattern shows sidelobes located on either side of the main lobe. The sidelobes represent signal energy taken from the main lobe that radiates into adjoining regions on the Earth. An example of sidelobes used to cover the Canary Islands of Africa is provided in Figure 7.20. Sidelobes, of course, are potential sources of radio interference to and from Earth stations in the geographic regions where they fall. In satellites that use multiple beams to reuse frequency, it becomes particularly important to control sidelobes. That can be accomplished through the selection and location of feedhorns as well as the use of the proper taper of the primary pattern.

On digital processor repeater systems, it also is possible to electronically alter the sidelobe levels through digital beam forming. That is accomplished inside the processor, where the carriers to be downlinked are altered in amplitude and phase so they combine in the far field in an optimum manner.

Figure 7.20   Spacecraft coverage pattern, including main lobe and sidelobes.

In rough terms, the 3-dB beamwidth (in degrees) of the far field pattern is related to the diameter of the reflector through the following approximation:

$$\Theta_{3dB} = 70 \frac{\lambda}{D}$$

where D is the antenna diameter measured in the same dimensions as λ. The formula is plotted in Figure 7.21 for the microwave bands of interest. The data show that the beamwidth decreases as either the diameter or the frequency increases. The gain of the antenna, on the other hand, increases as the square of either the diameter or the frequency. Also, the relationship between Θ_{3dB} (i.e., the coverage area) and the gain of the antenna is independent of diameter and frequency, for example,

$$G_0 \cong 27{,}000/(\Theta_{3dB})^2$$

As mentioned previously, it is practical to generate a beam that follows the contour of a country or region by combining the action of several individual feedhorns or by using a shaped reflector. Figure 7.22 shows how three feedhorns can be combined to produce a shaped beam. Horn B, located at the focus of the reflector, produces the expected main lobe labeled B. Now, examine the horn labeled A and its associated far field main lobe. Shifting the horn to the left causes the main lobe to be shifted to the right, much like what happens with a glass mirror. Similarly, shifting horn C to the right causes the far field pattern to shift to the left. The sum of the three patterns is illustrated by the wavy line labeled "combined pattern in far field." The pattern is produced by a feed network that divides the signal three ways, reducing the power that reaches each feed. For example, an equal three-way split reduces the signal level by 10 log(3), which is approximately 5 dB. The power will recombine in the far field due to the overlapping
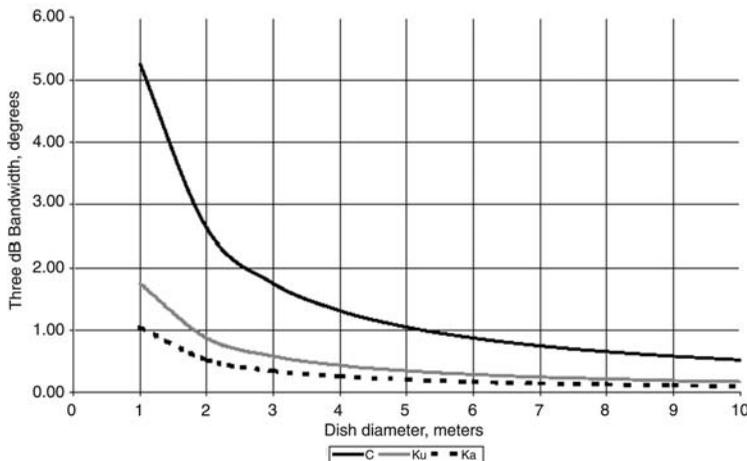
Figure 7.21   Relationship between diameter and beamwidth at C-, Ku-, and Ka-bands (downlink frequency assumed).
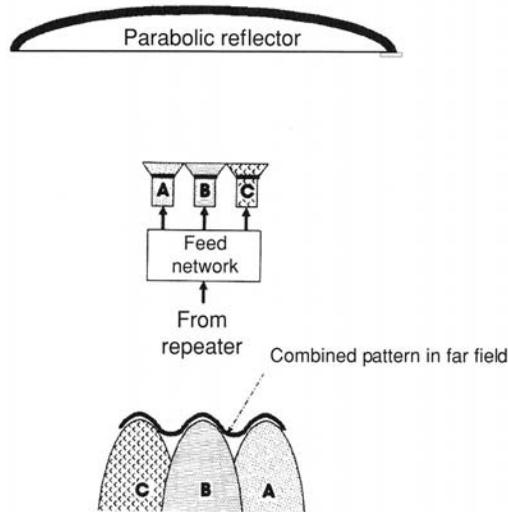
Figure 7.22   Generation of a shaped beam antenna pattern using multiple feedhorns and an associated feed network.

of the individual secondary patterns. In the case of reception, the same signal coming from the three horns is summed up in the feed network.

Beam shaping is a powerful technique for increasing the effectiveness of a communications satellite. In Figure 7.23, the coverage of Mexico provided by a single elliptical beam is compared with a shaped beam coverage from combining four nearly circular beams (indicated as the shaded area). Both the single and the shaped beams cover the entire Mexican land mass. However, the single beam radiates a substantial fraction outside Mexico, which represents a loss of about half the available signal power. In comparison, the shaped beam provides a relatively tight fit of Mexico, giving as much as 3 dB more gain at the edge of coverage, compared to the single elliptical beam.
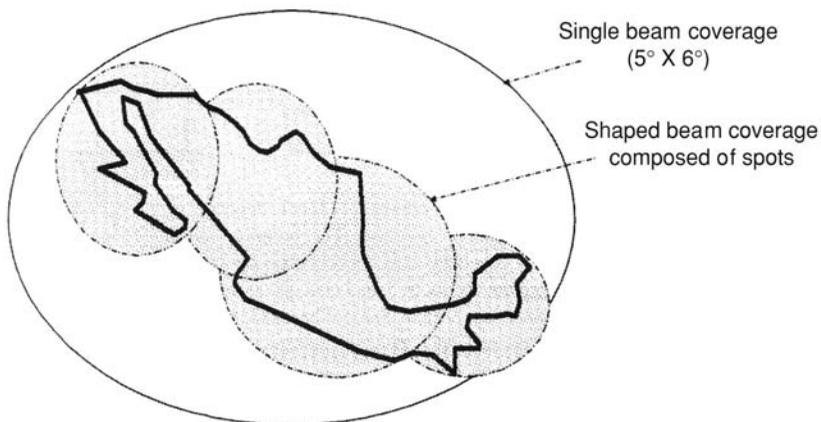
Figure 7.23   Spacecraft antenna coverage of the Mexican land mass demonstrating that a shaped beam is more efficient than an elliptical beam.

The overall communications performance in terms of EIRP and G/T depends on the spacecraft bus as well as the repeater and the antenna system. Some additional margin must be provided to allow for movement of the antenna caused by small changes in spacecraft attitude. As a result, there are periods of time when the gain at the edge of coverage is reduced relative to the nominal (properly pointed) value and times when it exceeds nominal as well. Most communications links are conservatively designed for the worst case, when the pointing error goes in the wrong direction. To take that into account, the points that represent the specified coverage area are extended outward to represent the worst cast. As a result, the minimum required EIRP and G/T automatically include the loss due to pointing error. An example is presented in Figure 8.8. Alternatively, the pattern could be shrunk to match the worst-case error.

7.4   Direct Radiating Array Antennas

There have been a few instances in the past where direct radiating antenna structures were used instead of reflector-based structures. In a direct radiating array, there is typically a large quantity of individual antenna elements that transmit and/or receive in the direction of the earth. There is nothing to block their view, other than possibly part of the rest of the spacecraft structure (something which is almost always avoided in the layout design). Like the feed horn array of a shaped beam reflector antenna, the power and phase of the same signal is applied to multiple elements.

The elements in a direct radiating array can take the form of a feed horn; combining several such feed horns would allow their individual gains to constructively add. Much lower gain elements like slots in sheet of metal or dipoles have been used in the past.

Boeing took an innovative approach in the Spaceway satellite, first launched in 2005. In Figure 7.24, we see one string of horn elements being assembled in the factory. The actual array is in the form of a circle, much like a reflector system. In this case, each element has its own SSPA and the far-field pattern is developed from appropriate combinations of amplitudes and phases of the elements.
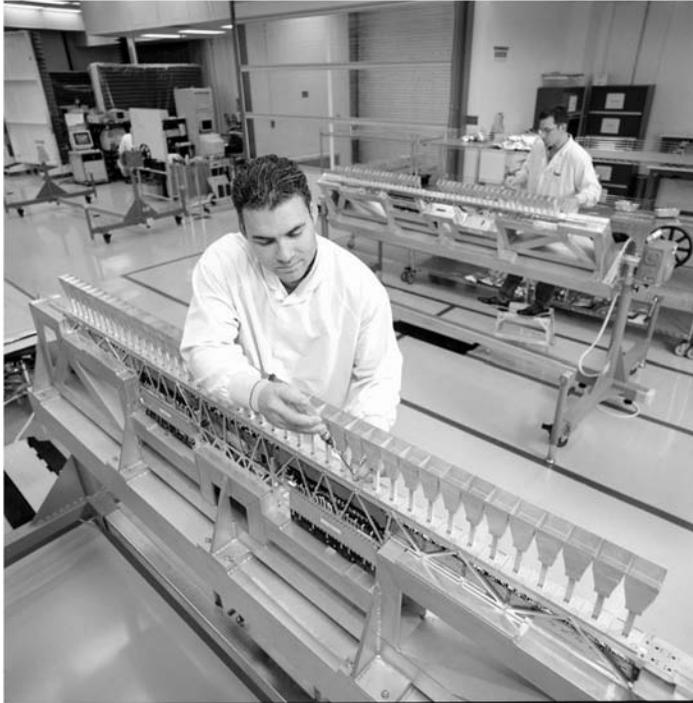
Figure 7.24   Boeing Space and Intelligence Systems technician builds part of the Spaceway phased-array downlink antenna, which will be integrated into a Boeing 702 satellite. (*Source:* Boeing.)

References

[1]   Bever, M., et al., "Broadband Payloads for the Emerging Ka-Band Market," TRW Space and Electronics.

[2]   Silver, S., (ed.), *Microwave Antenna Theory and Design*, London: Peregrinus, 1984.

[3]   Ruze, J., "Antenna Tolerance Theory—A Review, *IEEE Proc.*, Vol. 54, pp. 633–640, April 1966.

Spacecraft Mission and Bus Subsystems

The spacecraft bus is the platform that supports the communications payload during all phases of the mission, from launch, through injection into orbit, and throughout its operating lifetime. That is as challenging a task as processing and relaying the communications signals themselves and demands a thorough understanding of both the internal workings of the vehicle as well as its interaction with the various environments that are encountered both inside and outside the atmosphere. Across that spectrum of orbits and applications, the basic principles of bus design are fairly consistent. It is now a matter of adjusting the bus requirements to fit the mission.

8.1 Mission Summary

The steps necessary to provide a working communications satellite or constellation are generally termed the mission. Just as in the conventional military counterpart, each step requires careful planning and proper execution. System engineers and mission planners spend considerable time and effort to anticipate all possible situations and contingencies that the spacecraft will or might experience and to provide the hardware, software, and procedures necessary for a satisfactory outcome. A detailed discussion of the processes and technologies is beyond the scope of this book; however, many of the more important aspects are reviewed in this chapter. Some additional methodologies can be found in [1].

8.1.1 GEO Mission Profile

As an aid in understanding the functions of the bus, we use the standard get-to-orbit mission as a prime example. The bus design must deliver the payload to orbit and support it throughout the operating lifetime. The typical GEO mission, illustrated in Figure 8.1, goes through most phases that an orbiting vehicle could experience. Table 8.1 identifies the primary steps in the process, indicating that several, while brief, are critical to mission success. This is a somewhat lengthy mission in terms of time, due to the need to rest and verify every aspect of performance along the way. It may be possible to abbreviate several of the steps, something that has been done in the past when the satellite was urgently needed. The design of the bus must be consistent with the plan and vice versa. Regarding step 4, apogee injection boost, this is often done in successive stages using a restartable liquid fuel engine rather than a single-shot solid fuel rocket motor.
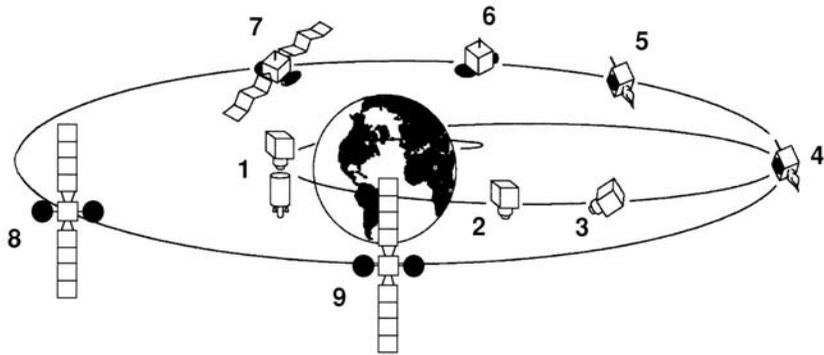
1   Separation from Launch Vehicle   6   Deploy reflectors and spin down
2   Reorient to sun-normal   7   Deploy solar panels
3   Reorient for apogee injection attitude 8   Sun acquisition and bus testing
4   Apogee injection boost   9   Earth acquisition and start of operations
5   Reorient for deployments

Figure 8.1   A typical GEO mission sequence for a three-axis spacecraft.

Table 8.1   Steps in a Typical GEO Mission

Step	Time From Liftoff	Event
1	0–30 min	Main boost phase
2	30 min	LEO achieved (GEO or non-GEO mission)
3	45 min	Transfer orbit injection (GEO mission)
4	45 min	Separation from launch vehicle
5	1 hr	TT&C link established with ground tracking station
6	Day 2	Geosynchronous orbit injection
7	Day 2 to 5	Orbital adjustments
8	Day 6 to 8	Initial bus configuration by ground command
9	Day 9 to 11	Drift to assigned orbit position
10	Day 12 to 13	Orbit and attitude adjustments
11	Day 14 to 18	Reconfiguration of spacecraft for operation
12	Day 19 to 23	Bus testing
13	Day 24 to 30	Payload testing
14	Day 31	Ready for service

8.1.2   On-Station Operation Requirements

The technical requirements on the bus after the spacecraft is ready for service are somewhat less severe that those that preceded. The space environment exposes the satellite to extreme temperatures, radiation and vacuum, and so places numerous demands on the bus. To properly take that into account, the bus is divided into functional subsystems, each of which is the subject of critical scrutiny by the relevant engineering and scientific experts. Designing and manufacturing the bus is the domain of rocket scientists, as well as some of the best minds in space physics and astronomy, thermodynamics, power electronics, structural engineering, aeronautics, and computer science. It is a multidisciplinary field in which the experts must cooperate and communicate effectively. The final critical element is the qualification

and testing of the hardware to guarantee that all mission requirements can be met even before the spacecraft is launched.

8.1.3 Non-GEO Requirements

Non-GEO systems intended for Earth observation, positioning, and science have been with us for many decades. However, they generally do not have to provide continuous service to all or a substantial part of the globe. Therefore, several innovations in that area were needed to ensure that commercially viable telecommunications services can be rendered to users.

The mission design for Globalstar is exemplary, employing 48 operating satellites. The constellation, illustrated in Figure 8.2, is designed to provide the best service to the middle latitudes, which corresponds to where the bulk of the population live [2]. Each operating orbit is at an altitude of 1,414 km and is inclined by 52 degrees.

Deployment of Globalstar satellites was in batches of four using Delta or Long March (a Zenit rocket failed to deliver 12 satellites to orbit) [2]. They were released at the lower orbit of 900 km and a stored command sequence used to deploy the solar panels, align them with the sun, and turn on the electronics. Once under ground command, a series of burns is used to move the satellites to the higher orbit.
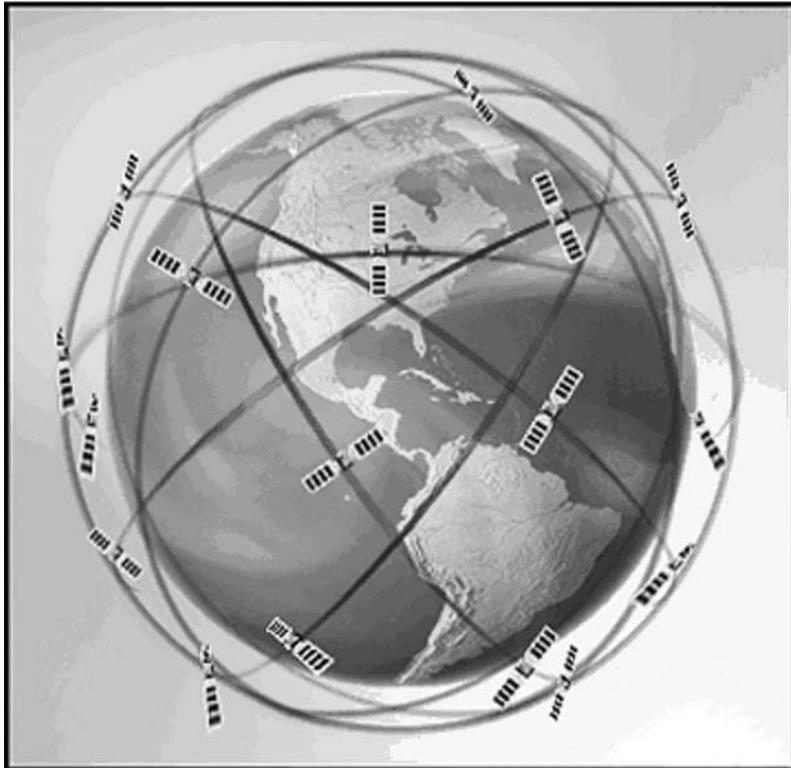
Figure 8.2 The orbital arrangement of the Globalstar LEO mobile satellite constellation.

All the Globalstar satellites are controlled from a single operation center, where data from the global network are available for analysis and storage. Useful data per satellite can be obtained for about 10 minutes per 114-minute orbit. Data from noncovered periods of time are stored onboard the satellite and downloaded during a monitored pass. The operational philosophy of Globalstar is to track differences from the norm and not to rely on 100% data availability and analysis, an approach common in GEO satellite operations. That is because of the large number of satellites to be monitored, the limited observation time, and the similarities among the satellites in operation. As discussed in Chapter 11, the Iridium system has a particularly interesting feature in that calls can be relayed from satellite to satellite, as in a daisy chain. All in all, the requirements for non-GEO spacecraft appear to vary greatly.

Spacecraft complexity, power, and mass differ greatly among the various GEO and non-GEO missions. GEO spacecraft generally are the most powerful, to deal with the added path loss and to provide adequate capacity to exploit a particular orbit position and footprint. As we move lower in orbit, the power and capacity demands on the spacecraft decrease. That is a particularly good thing because it compensates, to some extent, for the need for more spacecraft for continuous coverage. When we examine the bus designs used for a large LEO constellation of 60 or more satellites, we perhaps find that the individual spacecraft are small compared to the typical GEO design.

8.2   Spacecraft Configuration

Commercial aircraft all take essentially the same shape to fly at high speed and carry sufficient passengers and cargo to meet market demands. However, commercial communications spacecraft can and do take on a variety of shapes and configurations because the external forces in space are much weaker. There is, of course, the need to package the spacecraft for delivery to orbit by an Earth-launched rocket. Once in space, a spacecraft is unfurled into a solar-paneled vehicle that is stabilized along three axes.

This discussion is meant to give the reader some appreciation for the major tradeoffs in communications spacecraft design. Additional common features include power generation and storage, attitude sensing and control, propulsion for orbit control, telemetry and command, thermal control, structural design, and integration with the launch vehicle. Those criteria are addressed in relevant sections of this chapter on spacecraft bus subsystems.

As shown in Figure 8.3, the three-axis spacecraft (also called body stabilized) takes on the most convenient form for providing and supporting the communication function. Typically, the body is box-shaped with one side pointing toward the Earth at all times. Antennas are mounted on the Earth-facing side and the lateral sides adjacent to it. The solar panels are flat and deployed above and below the body, with orientation mechanisms provided to keep the panels pointed toward the sun as the satellite revolves around the Earth.

The three-axis design was adopted to deal with the apparent power limitation of the earlier spinning satellites and to make additional mounting area available

Height stowed
4 m (17.9 ft)

Width stowed
3.1 m (11.4 ft)
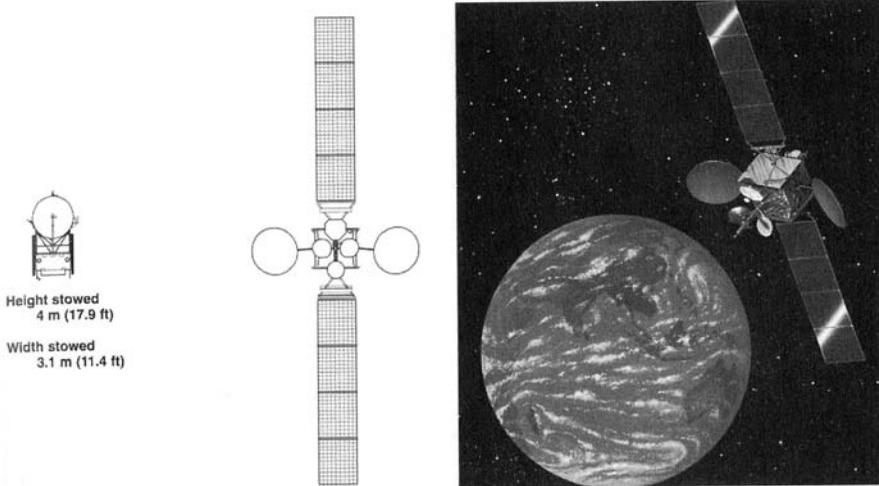
Figure 8.3 General arrangement of a three-axis GEO spacecraft.

for complex antenna structures. It took many years for the design to mature, particularly because of the complexity of stowage and on-orbit control. The power ranges available are in the 2,000–4,000W level on the low end up to the 15,000–20,000W level at the high end (at the time of this writing). As power is increased by enlargement of the solar panels, the overall design of the spacecraft becomes complex, particularly because of the added payload equipment and the requirement to manage the extra heat generated onboard.

Stability of the three-axis is ensured through an active control system that applies small forces to the body to correct for any undesired changes in spacecraft attitude. The most common three-axis attitude-control system ("biased momentum") uses a high-speed gyro called a momentum wheel to provide some stiffness and to act as an inertial reference. A basic control technique is to either speed up or slow down the wheel to cause the body of the spacecraft to rotate in the direction opposite to the speed change, which produces precise east-west adjustment of the platform (depending on the type of mission and the orbit design). The momentum wheel can sense transverse angular motion of the body (nutation), which causes a counterforce on its support. The momentum wheel can be twisted by a gimbal (a motor-controlled pivot), placing a precise dynamic force on the spacecraft. There should be at least two momentum wheels to ensure reliability in case of a single failure and to provide additional attitude control options.

An alternative approach, called zero momentum bias, provides separate reaction wheels in each of the three axes. Rather than spinning at a high rate in one direction, the reaction wheel can spin up and down and can reverse its spin and go in the opposite direction. The technique is precise because the axes are independently controlled and adjusted in orbit.

To provide large solar panels and reflectors, it is necessary to package the spacecraft for launch and then deploy all the appendages once orbit is reached. Furthermore, precise control of antenna beam pointing will rely on a fairly complex system of electromechanical devices and an onboard computer. In current genera-

tions of three-axis spacecraft, the bus section is designed as a separate module to be attached to the payload during final integration and test.

8.3 Spacecraft Bus Subsystems

The spacecraft bus encompasses those elements of the satellite that support the communications payload. As depicted in Figure 6.3, the support is multifaceted, involving the following functions: providing prime power during sunlight and eclipse; pointing the antenna beam(s) precisely at all times; controlling temperature to maintain proper opration of electronics and mechanical devices; providing physical support during launch and preoperational maneuvers; operational ground control of the satellite through TT&C functionality; and maintaining orbital position or paramters during the spacecraft's operating lifetime. Each function involves one or more of the bus subsystems, as described in the following sections.

8.3.1 Attitude-Control Subsystem

Several forces act on the spacecraft that cause the platform and the antenna to move away from optimum pointing. The attitude-control subsystem (ACS) is that combination of elements that senses any change in pointing and then causes the spacecraft to realign itself. This type of function can be found on virtually every space vehicle, including guided rockets, scientific deep space probes, and, of course, GEO and non-GEO communications satellites. Commercial satellites usually have some built-in (autonomous) ACS capability, but manual intervention from the ground often is necessary during certain phases of orbital operations and failure modes.

8.3.1.1 Attitude-Control Loop

Shown in Figure 8.4 is a conceptual block diagram of an ACS control loop intended for autonomous operation. The pointing sensor (located at the upper left) measures the pointing angle with respect to some reference. Having the ability to detect changes along two axes (north-south and east-west), the sensor is a critical element because its measurement accuracy has a large impact on overall ACS pointing accuracy. (Typical sensing techniques are reviewed in a subsequent section.) The ACS loop includes a sensor electronic unit to process the sensor's output signals and produce a properly calibrated error message. Attitude processing typically is carried out in the onboard spacecraft control computer. It accepts the error message for use in determining corrective action. Other functions with the processor tie attitude control to orbit maintenance since thruster firings introduce attitude perturbations. Ground control of the processor can be exercised during spacecraft maneuvers and to update its data and programming.

The processor control signals are used to activate the electrical windings of the mechanical devices (drives) or other transducers that cause the spacecraft attitude to actually change. Three-axis spacecraft typically use several mechanical drives and devices. For example, the speed and the gimbal angle of the momentum wheel are
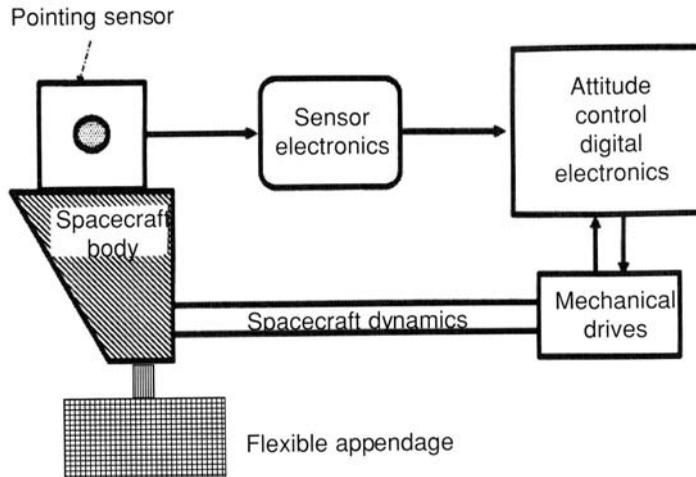
Figure 8.4   Conceptual block diagram of the ACS showing the arrangement of the control loop.

adjusted to cause small changes in spacecraft body orientation. To maintain a stable three-axis platform, those forcing functions must be applied softly because of the small degree of gyroscopic stiffness produced by a momentum wheel. However, digital processors are programmed to use adaptive control laws to minimize the extent and duration of transients during orientation changes. The control laws and procedures can be modified after launch by uploading new computer code and calibration data over the command link.

The spacecraft is not a rigid body but actually has several flexibly attached parts, particularly the solar panels and antenna reflectors. Therefore, as shown at the bottom of Figure 8.4, the force for change from the mechanical drives is applied to the body through spacecraft flexible dynamics. That simply means some parts of the body will bend and vibrate when a force is applied to the spacecraft. That, combined with the inertia of spacecraft mass, produces transients in the pointing performance of the antenna beams, which can last several seconds or minutes. It is important that the ACS prevent the peak value of the transient from exceeding the acceptable beam error and that transients are properly damped out without excessive overshoot and duration.

The control loop in this description is now complete because the sensor is mounted on the body and also will move when the attitude of the spacecraft changes. The sensor transmits a new indication of orientation, which in turn is fed into the attitude control digital processor. The active and adaptive nature of the control loop allows it to continuously adjust the orientation of the spacecraft and to dampen any transients. Failure to do that properly can result in unstable antenna pointing or, in the worst case, physical damage to the appendages. Another critical worry is spacecraft safety in case of a major mispointing event, such as due to failure of a sensor, anomalous thruster firing, or operator error. The situaion can develop when a failure causes the spacecraft to move off proper pointing and even Earth alignment. From there, solar pointing and hence prime power generation could cease. That might combine an alignment of the satellite with the sun illuminat-

ing an area of the spacecraft that cannot tolerate excessive heat (or not illuminating an area that requires solar heat input to maintain adequate temperature). As a result, communication performance is lost, and the spacecraft, if not restored to a safe attitude, could become unserviceable.

The answer to such relatively infrequent occurrences is to preprogram the onboard computer with contingency procedures to put the satellite into a safe condition. From there, ground commands are sent to restore normal operation and to determine the source of the problem.

8.3.1.2   Sensors

The two principal types of attitude sensors are the infrared Earth sensor and RF tracking. Another very precise tracking technique is called star tracking. At certain times in the mission, it is not possible to use the Earth as a reference due to interference from the sun or moon. That usually is overcome with a sun sensor made up of a silicon solar cell and optical barrel.

Depicted in Figure 8.5 is the Earth sensor concept for a three-axis satellite, using a scanning optical device consisting of two small telescopes with infrared detectors at the viewing end. An alternative design employs a fixed matrix of infrared detectors. The star tracking technique mentioned previously also uses an optical telescope to find and track a bright star. This is an extremely precise but more complicated tracking scheme.

The RF tracking technique, depicted in Figure 8.6, is an accurate system for sensing spacecraft attitude. Some of the same devices used for reception of the uplink microwave signals are applied to tracking. Each axis requires two receive horns and an associated hybrid. Out of the hybrid come two ports, one containing the sum (A + B) and the other the difference (A − B). The sum signal represents

(A) Arrangement of north-south and east-west earth sensors
on the body of a three-axis spacecraft
(double-headed arrows indicated scanning motion)

(B) Examples of earth sensor pulse timing waveforms
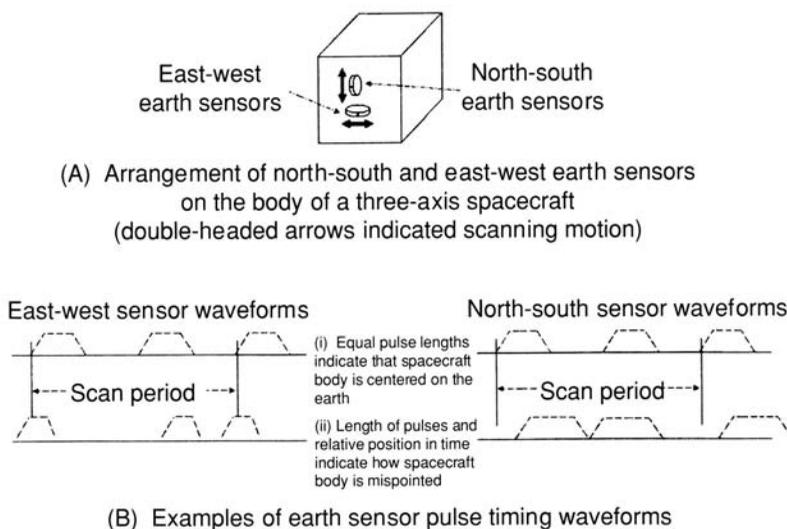
Figure 8.5   (a, b) Arrangement and operation of dual scanning Earth sensors on a three-axis spacecraft.
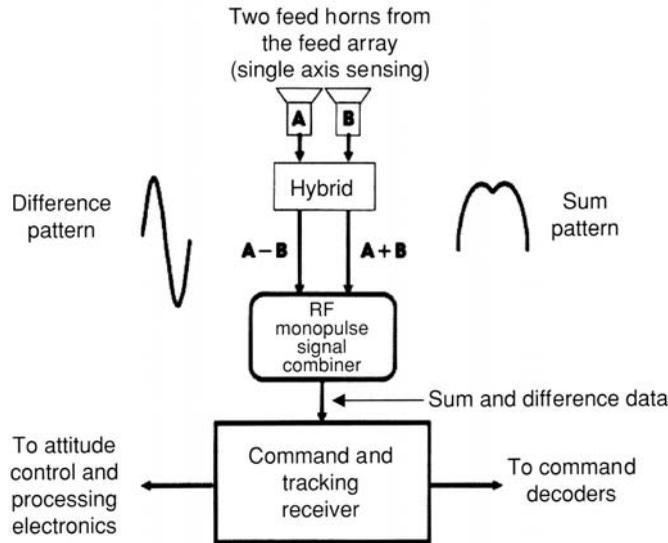
Figure 8.6   Attitude sensing by the spacecraft antenna and command receiver when tracking an RF pilot signal transmitted from the TT&C Earth station.

the combination of the two beams, having a combined pattern similar to the one shown on the right. However, the difference signal effectively passes through zero in its response precisely where the individual horn patterns overlap. The difference pattern, which has the appearance of a sinewave, is used to detect a null in signal energy corresponding to proper alignment of the antenna beam. That requires that a reference Earth station be located where the null should be placed and that the station transmit a continuous signal called a pilot carrier.

There remains the problem that the difference signal by itself does not provide a continuous error signal when the beacon enters the null. That is overcome by modulating the A − B response onto the A + B response, which is accomplished in the RF monopulse tracking processor. The modulation then can go through its zero point at the null. Bias and drift in this null are corrected by storing an offset pointing angle in the onboard attitude control processor. It also is advantageous to use a single receiver to perform all commanding, ranging, and tracking. Separation of the attitude-error signal from the command signal usually is accomplished in the command-track receiver, shown at the bottom of Figure 8.6. RF tracking systems have measurement accuracies in the range of ±0.01 to ±0.04 degree. Performance of the star tracker is of this order, or possibly better.

8.3.1.3   Antenna Pointing

The real measure of ACS performance is its effect on the accuracy and stability of spacecraft antenna pointing. This is a complex subject because it involves the performance of several elements of the payload and the bus and the interaction thereof. Pointing performance is the most critical for GEO missions for which pointing angles are the smallest. In contrast, pointing for a LEO satellite is nearly

one order of magnitude less critical; for that reason, bus operation can be somewhat simplified.

Overall pointing accuracy is predicted by identifying each contributor and assessing its impact in the form of a pointing budget. Sensor accuracy contributes approximately 10–20% of the total accuracy of pointing. A second big contributor is the movement of parts of the physical antenna due to variations in solar heating of the reflector and feed supports. Referred to as thermal distortion, that effect can produce as much as one-half of the total error, depending on the thermal stability of the material used in antenna construction. Another error is that due to the precision with which the antenna and sensor are attached to the spacecraft, a process that requires the use of optical alignment instruments during assembly on the ground. By providing a ground commanding reflector alignment mechanism, the static beam pointing error can be removed almost entirely. Last of all, the transient behavior of the ACS during thruster firing will cause the antenna to be misdirected momentarily. Analysis of those effects involves the individual accuracies, the interaction of the elements, and the operation of the spacecraft in different stages of the overall mission. That places considerable demands on the prelaunch design and test of the spacecraft, the operating principles of the spacecraft when in orbit, and the means to respond to contingency situations that arise from time to time.

Figure 8.7 provides the general view of the spacecraft, indicating the three classical axes of attitude: roll, pitch, and yaw. The axes are based on the view of the spacecraft as it flies over the Earth like an airplane. The roll axis, which typically is in the orbit plane, is pointed in the direction of orbital flight; pitch corresponds to the up-down motion of the nose of the spacecraft along an axis perpendicular to flight and in the orbit plane; and yaw is a twisting motion of the antenna about the axis pointed directly down from the spacecraft (e.g., sideways motion of the nose).

For a typical operating satellite, the overall pointing error of the antenna beam is in the range of ±0.05 to ±0.15 degree in either the north-south or east-west
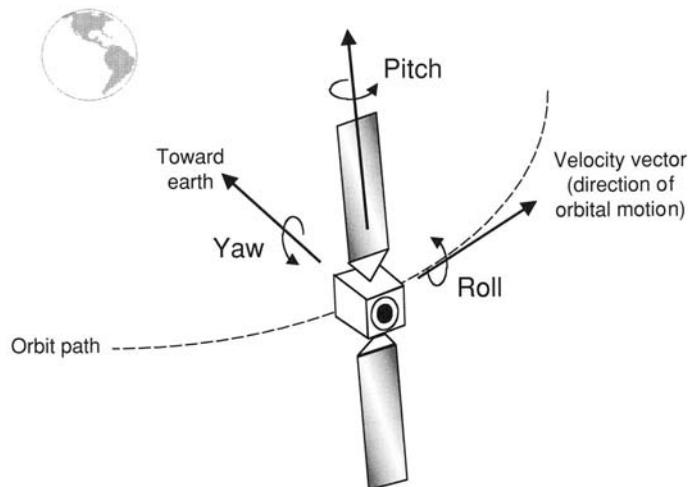
Figure 8.7   Definition of the three axes of satellite change: roll, pitch, and yaw.

directions. Motion about the third axis of the spacecraft (yaw) causes the antenna pattern to twist with respect to the Earth and therefore represents a third error-prone direction of motion. A benefit of the standard ACS approach is that yaw error is automatically removed in one-quarter orbit (6 hours later for GEO) because yaw error turns into roll error at that point. For coverage regions some distance from the equator, yaw can be nearly compensated for by east-west and north-south ACS control.

Examine the antenna coverage of Mexico in Figure 8.8, which is based on the shaped beam antenna design described in Chapter 7. The pattern provides 30 dBi of gain (the black contour) over the Mexican landmass and reaches 33.5 dBi at the peak. The approach commonly used to evaluate antenna pointing error is to employ circles or boxes with dimensions corresponding to the maximum pointing error anticipated. Because Mexico is considerably north of the equator, yaw error has been broken down into east-west and north-south components, which are already included. In the example, the ACS pointing error is assumed to be ±0.10 degree in both directions, which makes the error box 0.2 degree on each side. The conservative (worst case) approach is to use boxes at the test points (circles can be used under the assumption that north-south and east-west errors are uncorrelated). The circles then are placed at critical locations such as key cities or at points along the border of the region or country being served. In Figure 8.8, error circles have been placed at Mexico City as well as two points along Mexico's coast that protrude into the water. The Mexico City circle shows that the gain at that point nominally will be 32 dBi but can be expected at times to decrease as much as 0.5 dB (obtained by interpolation between the 32- and 30-dBi contours). The lower right circle (located at the tip of the Yucatan Peninsula) shows that even with maximum error, a gain of 30 dB is maintained. However, the error circle at the left (Pacific coast) extends outside the contour; therefore, the gain at times will drop below 30 dB. As an alternative to using boxes or circles, the size of the map of Mexico can be increased by the extent of the maximum pointing error.
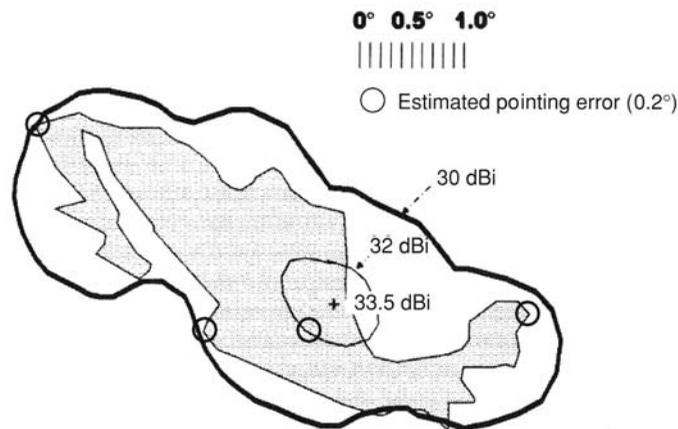
Figure 8.8   Spacecraft antenna gain performance in decibels indicating how pointing error is taken into account.

This procedure is used to compare the predicted performance of the spacecraft to the specification. While values of gain are shown in Figure 8.8, it is more appropriate to label the contours with EIRP and G/T, determined in accordance with the procedures given in Chapter 4. That then brings into play the combined performance of the communications payload and attitude-control system.

8.3.2   Solar Cells and Panels

All Earth-orbiting communication satellites use panels of solar cells to provide prime power to the payload and the bus. Over the years, performance of silicon solar cells has improved; however, high-performance solar panels are being constructed of GaAs solar cells, which can outperform silicon by a factor of two to three. The power output of a solar cell is proportional to the intensity of solar radiation reaching the cell. The intensity is maximum with the rays arriving perpendicular to the cell and decreases with the cosine of the angle of the ray with respect to the perpendicular. A condition of zero output occurs when the rays are parallel to the cell's surface. There is a seasonal variation that results from the angle between the ecliptic plane and the satellite's orbital plane.

In the case of the type of flat panel on a three-axis spacecraft, a drive motor maintains the array in the direction of the sun and all cells receive identical illumination. Alignment in that manner is optimum for a GEO spacecraft; however, for a non-GEO mission, the spacecraft must be maintained with proper attitude for panel illumination even though the orbit is not in or near the equatorial plane.

Cells are connected in vertical strings to increase voltage and permit redundancy. As noted above, the output power from a solar panel varies over the course of a year due to the orbital geometry. A simplified drawing of the geometry is shown in Figure 8.9, indicating that the equatorial plane at GEO altitude is tilted approximately 23 degrees with respect to the ecliptic plane (i.e., the plane of the orbit of the Earth around the sun). That means a satellite experiences the seasons just as we do on Earth. (The figure uses summer and winter as seen from the northern latitudes.) In summer and winter, the sun's rays hit the panel at an angle,
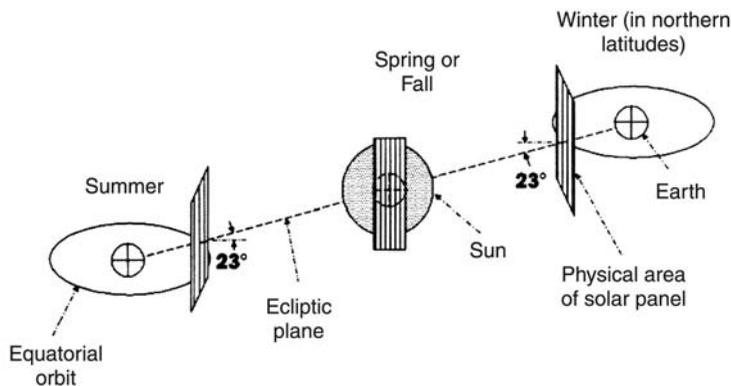
Figure 8.9   Variation of the illumination of the solar panel as the Earth and a satellite in GEO revolve around the Sun.

reducing the efficiency of the solar cells. The worst case occurs on the solstices, when the efficiency factor is equal to cos(23 degrees), that is, 92%. In the spring and fall, on the other hand, the sun provides better illumination of the panel, and the cells give their maximum output at the equinoxes. It happens that the point of minimum panel output occurs at the summer solstice, when the distance between Earth and sun is greater than in winter.

For non-GEO spacecraft and constellations, the question of solar panel illumination can be much more complicated. Among non-GEO systems, the arrangement of orbits and communication coverage requirements differ significantly, imposing unique constraints on solar power generation. That would not be a problem if the spacecraft was a sphere covered with solar cells to catch the sun independent of attitude. However, flat-panel arrays are both more convenient and more efficient, so some means must be found to provide separate alignments throughout the orbits and the mission. In an inclined orbit system like Globalstar (Figure 8.10), the dual flat panels align with the sun, and the Earth-facing phased array antenna creates 16 overlapping beams. To keep the panels on all satellites in proper alignment with the sun, the beams must rotate about the yaw axis as the satellite passes over the Earth.

In contrast, the polar-orbiting Iridium satellites each keep the orientation of their 48 beams fixed, which means the panels cannot be maintained in a perpendicular orientation toward the sun (Figure 8.11). Without yaw motion, satellites with their orbit plane nearly perpendicular to the sun line experience poor solar illumination and therefore deliver only a fraction of their potential. The panels are oversized so that adequate power is generated for the bus and the payload. A fortunate consequence of this particular orbit orientation is that no power is required for battery charging since the satellite does not go into eclipse during this time period. Optimum alignment, shown in Figure 8.12, brings with it daily eclipse periods and a requirement for rapid battery charging when the satellite is in daylight. Iridium satellites have the additional capability to twist their solar panels along a second axis to produce enough power during periods of poor sun illumination.
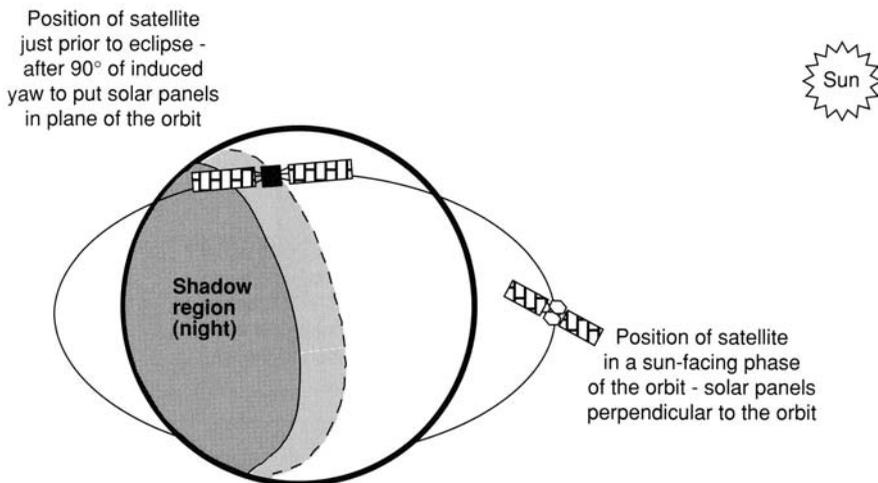
Figure 8.10   Sun illumination for an inclined orbit constellation such as Globalstar.
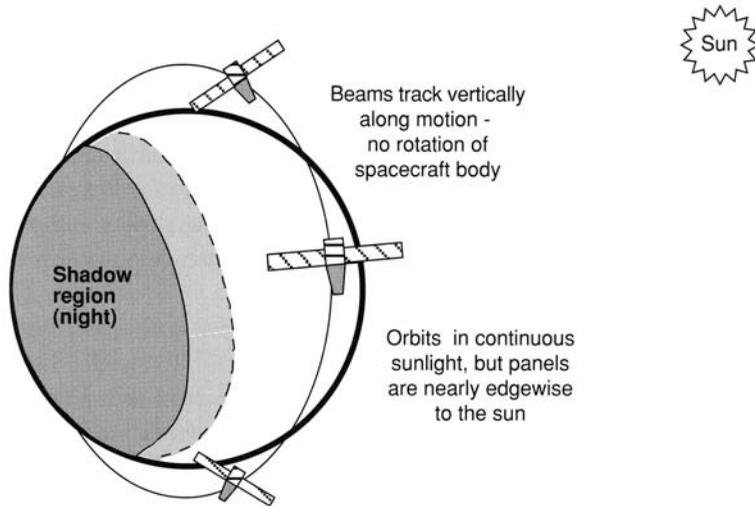
Figure 8.11 Sun illumination for a polar orbit constellation such as Iridium. This orbit has the poorest solar-panel orientation but does not experience eclipse.
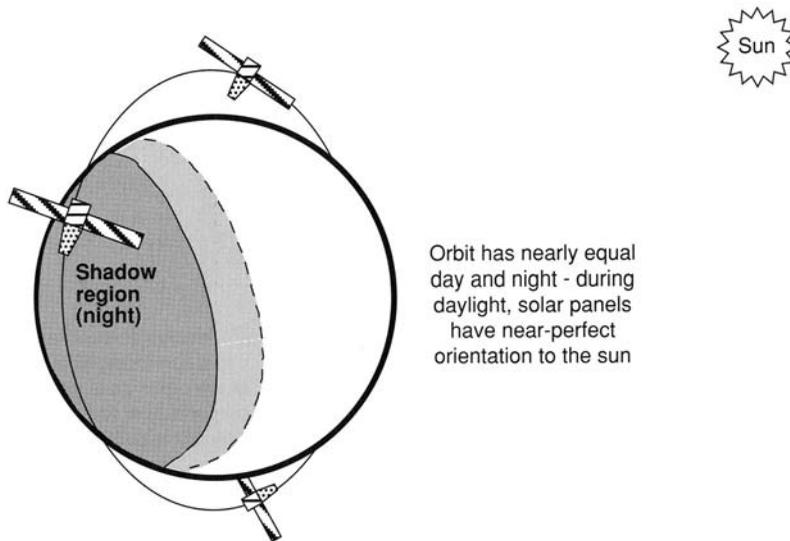
Figure 8.12 Sun illumination for a polar orbit such as Iridium; this orbit has ideal solar-panel orientation but experiences frequent eclipse.

These geometrical factors should provide some appreciation of the basic requirements for solar panel design. There are, however, many other factors to consider. The space environment degrades panel performance over time. For the geostationary orbit region, local electrons impinge on the solar panel and gradually degrade cell performance and solar storms erupt in blasts of protons whose energy is particularly damaging to solar cells. LEO satellites experience less electron degradation. By proper cell and panel design, it is possible to reduce the degradation to the range of 15–20% over the life of the satellite. Most of that power loss occurs

in the first few years of operation, while the loss in later years tapers off. Mechanisms and drive motors that are used to deploy and orient the panels must be very reliable with redundancy built in where possible.

8.3.3   Battery Design and Configuration

The solar array described in Section 8.3.2 is only a part of the spacecraft electrical power system that provides continuous power to all electrical loads. During the eclipse period, the sun is concealed by the Earth, and the spacecraft must derive power from a battery system. The frequency and duration of eclipse as a function of orbit altitude is graphed in Figure 8.13. In a LEO constellation, some fraction of the orbits will subject their satellites to eclipse periods, which is a natural consequence of entering and exiting the side of the Earth experiencing nighttime. There is a polar orbit called sun synchronous that avoids eclipse entirely by adjusting the orbit plane to be perpendicular to the sun. In the Iridium system, for example, one or two orbits are in effect sun synchronous while the rest are not. MEO and GEO satellites experience eclipse much less frequently and also have seasons in which there is no eclipse at all.

8.3.3.1   Electrical Power Bus

Figure 8.14 is a simplified block diagram of a typical electrical system incorporating dual independent electrical busses. Separate busses provide redundancy and give assurance that a major failure (such as a short circuit) could eliminate only half the system. Alternatively, a single power bus architecture can be used for the primary electrical source with the split to diverse systems occurring downstream. To understand the construction of the electrical system, examine the arrangement of the upper half of Figure 8.14 (i.e., the part corresponding to bus 1).

   Prime power is generated in the solar arrays and delivered to the bus. The limiter immediately to the right of the array is provided to drop the voltage during periods of maximum solar illumination (e.g., the equinoxes in GEO missions) or minimum power demand in the spacecraft. Switching in or out of the limiter is
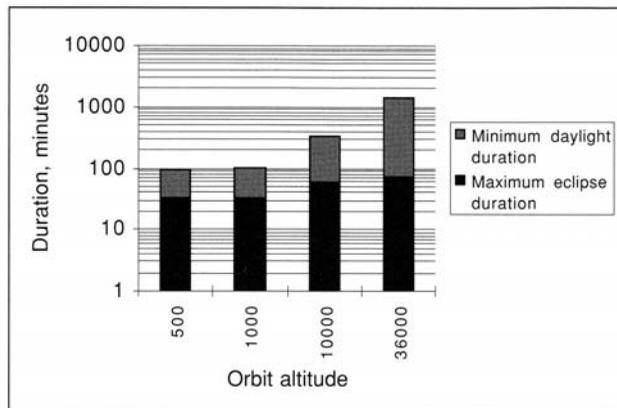
Figure 8.13   Maximum eclipse duration for a circular orbit as a function of altitude.
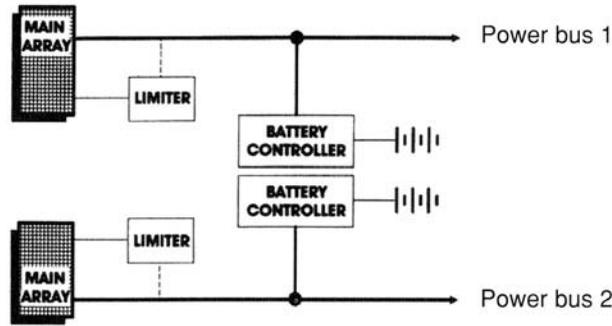
Figure 8.14 Spacecraft electrical power system, including solar arrays and batteries with dual independent power buses.

accomplished automatically by sensing the voltage on the bus (see the dashed line). The downstream electronic units further regulate their internal voltage, consistent with the specific circuit requirements. Alternatively, the primary voltage can be tightly regulated on a central basis to simplify the design of the powered units in the spacecraft.

In Figure 8.14, the symbol composed of alternating short and long lines represents the battery composed of electrochemical cells used to store energy for eclipse operation. Nickel-cadmium cells were used successfully for many years, but the more efficient nickel-hydrogen (NiH_2) cells became standard 20 years ago. Illustration of a typical NiH_3 cell is provided in Figure 8.15 [3]. The basic operation of this type of cell depends on the chemistry of gaseous hydrogen and water. The hydrogen gas stored in the pressure vessel increases in pressure as the cell is charged. When the cell subsequently is discharged, the nickel-active material is reduced to its original preoxidation state and the hydrogen gas is oxidized to form water. The cell pressure gradually decreases to its low value as the hydrogen gas is consumed. The energy storage efficiency and lifetime of this class of cell are the standard measure in the industry.

More recently, spacecraft manufacturers have introduced lithium-ion cell technology into their products. Lithium ion (Li-Ion) batteries are replacing traditional NiH_2 batteries as the standard in geostationary satellites as a result of their mass
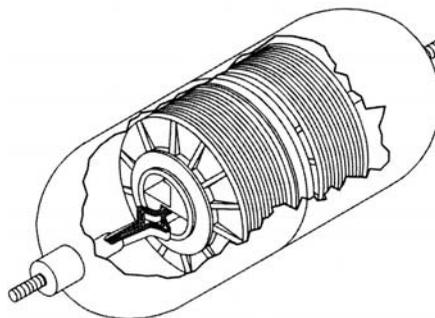
Figure 8.15 Standard NiH_2 dual stack independent pressure vessel.

efficiency. For example, after its usual careful qualification and thorough testing process, Space Systems/Loral introduced Li-Ion batteries into its 1,300 product line with the launch of Thaicom-4/IPSTAR in August 2005. The Thaicom-4 power subsystem provides 14 kW of power throughout the satellite's 15-year life. The Thaicom-4 Li-Ion battery system is shown in Figure 8.16. By January 2008, SS/L had launched one satellite with Li-Ion batteries, and has 14 more under construction. SS/L now considers Li-Ion batteries baseline for its satellites of power levels to 20 and 25 kW and beyond, and orbital lifetimes of 15 years.

Battery charge is maintained by directing excess solar array current to the battery system during sunlight. A battery controller switches the electrical bus from the array to the battery at the start of eclipse and back at the conclusion. The action can be automatic or by ground command, depending on the particular approach taken in the design of the system. The battery controller also is used for maintenance of the batteries, which consists of delivering a light "trickle" charge most of the time and periodically commanding a rapid discharge into resistors for cell reconditioning (for older battery designs). The overall process of battery maintenance is a complex subject, particularly in light of the criticality of energy storage for the entire mission.

8.3.3.2 Electrostatic Discharge

The spacecraft electrical system provides conduction paths to every electronic and electrical unit in the spacecraft and can allow potentially damaging electrical transients to propagate through the system. Low-level transients generated aboard the
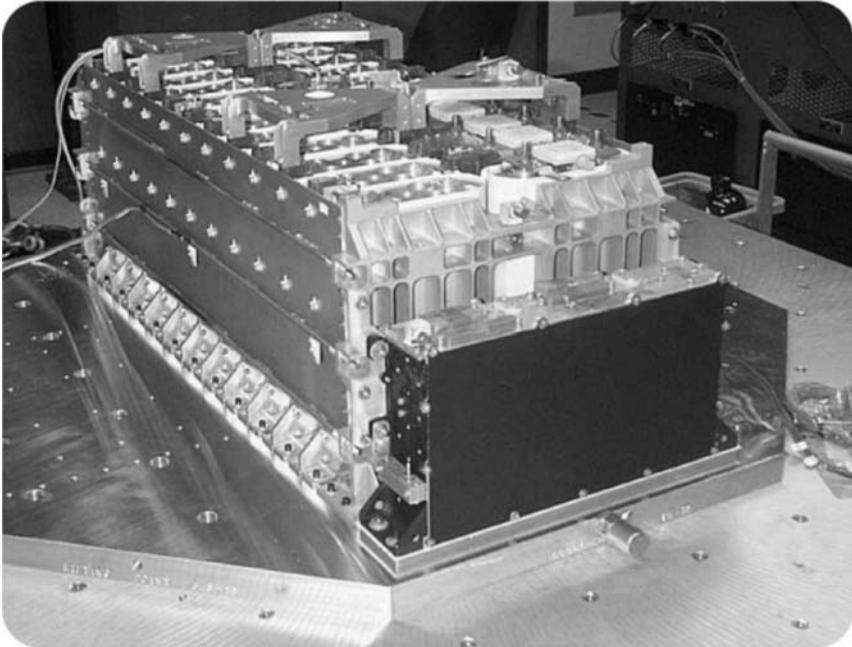
Figure 8.16   The Lithium-Ion battery system incorporated by Space Systems/Loral on Thaicom-4. (*Source:* Space Systems/Loral.)

satellite by normal operation of the payload can be rendered harmless by proper power supply design. However, it is possible for spikes in the range of hundreds to thousands of volts to occur from Electrostatic Discharge (ESD), a phenomenon not unlike lightning.

ESD is the result of a buildup on external surfaces of the spacecraft of a static charge from electrons traveling in space. In a well-designed and constructed satellite, all conducting surfaces are electrically connected to provide a common ground running through the spacecraft structure. That means all points inside the spacecraft rise in negative voltage potential together. However, if any external surface (conducting or nonconducting) is not connected to ground, its voltage potential differs from the rest of the spacecraft. As the charge builds up and potential difference increases, an electrical arc could leap across any gap that separates the external surface and another (grounded) element such as a piece of metal waveguide or structure. The large surge of electrical current that would ensue could pass from ground into the power system and be carried into sensitive electronic equipment.

The effect of ESD has been experienced many times in the past, particularly on new spacecraft designs. A seemingly random shutdown of an operating transmitter could be a symptom of ESD as could a "glitch" (a wild variation) in the telemetered output of a sensor. In some cases, ESD has been known to permanently damage digital electronic equipment and power supplies. Therefore, it is vital that ESD be carefully taken into account by the use of appropriate materials and by proper grounding.

8.3.4 Liquid Propulsion System

The onboard propulsion system uses the principle of expelling mass at some velocity in one direction to produce thrust in the opposite direction (Newton's first law). As with any rocket, there are principles and practices that govern how safely and effectively propulsion is achieved. Three basic types of mass can be ejected: solid fuel (e.g., a controlled explosion), liquid fuel (which decomposes into a gas jet), and ion propulsion (which is based on accelerating charged particles with an electric field). We begin with the liquid type of propulsion system, the most common.

The majority of commercial communications spacecraft control their orbital positions by using small liquid fuel rocket engines. Propellant options include a cold gas like nitrogen, which is low in cost and performance, a single combustible propellant like hydrazine that produces hot gas thrust when exposed to a catalyst, or two liquids (fuel and oxidizer) that produce combustion on contact. Low levels of thrust are needed for small corrections in orbit, while high levels are needed for major orbit changes or corrections. Large amounts of thrust required for major orbit changes can be provided either by solid rocket motors or by large fuel engines. The general arrangement of a three-axis spacecraft propulsion system is shown in Figure 8.17.

The launch vehicle provides the boost to place the spacecraft into an initial orbit from which adjustments can be made using the onboard propulsion system. In the case of most GEO missions, that is an elliptical orbit with perigee at
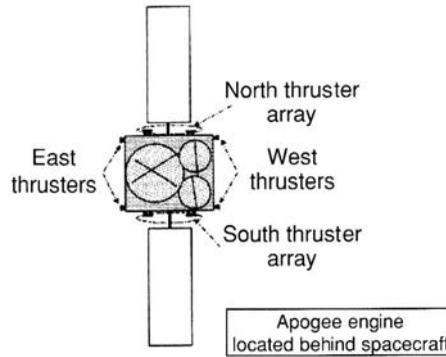
Figure 8.17   General arrangement of propulsion systems for a three-axis spacecraft.

approximately 300 km and apogee at near-synchronous altitude (36,000 to 37,000 km). The configuration in Figure 8.17 assumes the use of a solid fuel apogee engine that is permanently integrated into the spacecraft and fired by ground command.

The fact that a three-axis satellite does not spin to provide stability leads to the need for many more thrusters and a much lower level of thrusting than employed in the classic spinning satellites discussed in Chapter 2. As shown in Figure 8.17, the typical three-axis uses 12 thrusters arranged in arrays on the north, south, east, and west faces. Low thrust levels are needed to minimize disturbance of spacecraft attitude during thrusting.

The performance of any thruster or rocket engine (liquid, solid, or electric) is measured in terms of two key parameters:

1. Thrust force, F, measured in Newton (N) or pounds (force), there being approximately 4.45 N/lb (force);
2. Specific impulse, I_{sp}, measured in seconds, is the ratio of thrust force to the mass expelled by the thruster to produce the thrust. This performance parameter determines how much mass must be ejected to produce a given orbit velocity increment. Multiplied by gravitational acceleration, g, I_{sp} gives the thruster mean velocity, c, which is the true meaningful physical parameter in orbital maneuver calculations.

For reference, Table 8.2 is a summary of key characteristics for the types of thrusters described here. They break down into chemical thrusters, electric augmentation and ion thrusters, and solid-fuel rocket motors. The range of I_{sp} values is from under 100 seconds on the low end for cold gas up to 2,000 seconds or more for the different forms of electric and ion propulsion. As is usually the case in high technology, the lower performance approaches are simple and very reliable; as we move toward the higher performing systems, the complexity and cost increase dramatically. However, the trend in large satellites is toward the higher performing types of thrusters that also provide relatively low thrust force levels. Table 8.2, summarized from [4], is for illustrative purposes and should not be used as a design guideline.

Table 8.2 Basic Comparison of Onboard Thrusters Used for Various Commercial Satellite Missions

Type	Propellant	Energy	Typical I_{sp} (sec)	Thrust Force (N)
Cold gas	N_2, NH_3, He	Gas pressure	100	0.01–50
Solid fuel rocket motor	Dry mixture of fuel, oxidizer, binder	Chemical	300	$10–10^6$
Monopropellant	Hydrazine	Exothermic decomposition	230	0.05–25
Bipropellant	MMH and N_2O_4	Chemical	300	$10–10^6$
EHT	Hydrazine, NH_3	Exothermic decomposition and electrothermal	300	0.05 to 3
ARCJET	NH_3, N_2H_4	Electric arc heating	800	0.05–5
Ion propulsion	Hg, Xe, Ce	Electrostatic	3,000	$10^{-6}–0.5$

8.3.4.1 Reaction Control System

The liquid fuel tanks, fuel lines, values, and thrusters make up the reaction control system (RCS), which is an integral part of the bus. Figure 8.18 shows a conceptual drawing of a basic spinner's RCS, which uses a single liquid fuel called hydrazine. With the tanks having been pressurized prior to launch and with the spacecraft spinning on orbit, fuel exits the tanks and remains in the lines. The drawing is not accurate in that the lines are actually connected to the tanks on the side facing the exterior (solar panel) so that rotation forces the fuel out of the tanks.

The system shown in Figure 8.18 is split into redundant halves, either of which could support the mission with one axial and one radial thruster and a pair of tanks. An interconnect latch valve, which normally blocks connection between the two halves, can be commanded open to allow fuel from one side to move to the other. It normally is used during the mission to equalize fuel load and pressure, but it also can allow access to half the fuel in the event of a thruster open failure.
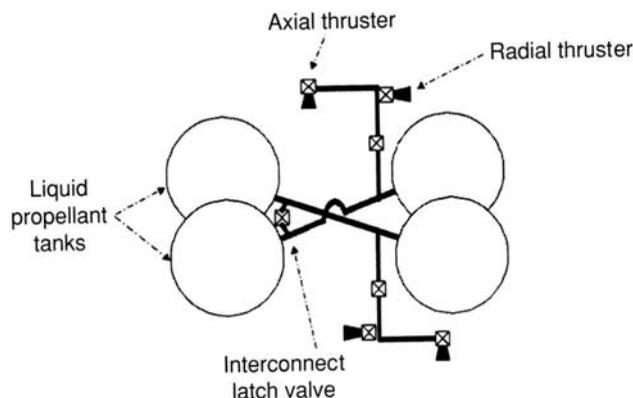
Figure 8.18 Simplified diagram of a spinner's reaction control subsystem employing monopropellant hydrazine fuel.

Note the use of an isolation valve in the line between the radial thruster and the tanks, which can be closed if a thruster fails in the open position. Tanks and lines typically are made from welded titanium and have been thoroughly tested on the ground to preclude even the slightest possibility of a leak.

In the case of a three-axis satellite, the flow of fuel in the tank is controlled by use of a device in the tank rather than the force of rotation. One approach is to use a rubber diaphragm separating the fuel from the pressurant gas, allowing the gas to push the fuel out of the tank, much the way we squeeze a tube of toothpaste. Another more common technique employs vanes and spongelike surface-tension devices to direct liquid fuel toward the tank exit by capillary action.

Conventional liquid monopropellant systems use hydrazine fuel alone, which on contact with a catalyst decomposes into its constituents. The process releases energy, resulting in high-pressure gas at the thruster nozzle. Another important innovation in spacecraft RCS was the introduction of bipropellant technology, that is, the use of a separate fuel and oxidizer. That is now the standard for GEO satellites. Monopropellant and bipropellant thrusters are compared in Figure 8.19. Having separate fuel and oxidizer produces greater thrust for the same weight of fuel, which allows a bipropellant system to yield longer life (or conversely save fuel weight for the same life). Of necessity, bipropellant systems have more tanks, lines, and valves; therefore, there is some penalty in mass, cost and increased complexity.

An improvement in monopropellant performance has been achieved with the electrically heated thruster (EHT), which increases thrust by heating the propellant with an electrical winding. Thrusting with the EHT has to be done in such a way as to use excess electrical power; otherwise, there would be a weight penalty for the additional solar cells. The benefits of bipropellant and EHT technology
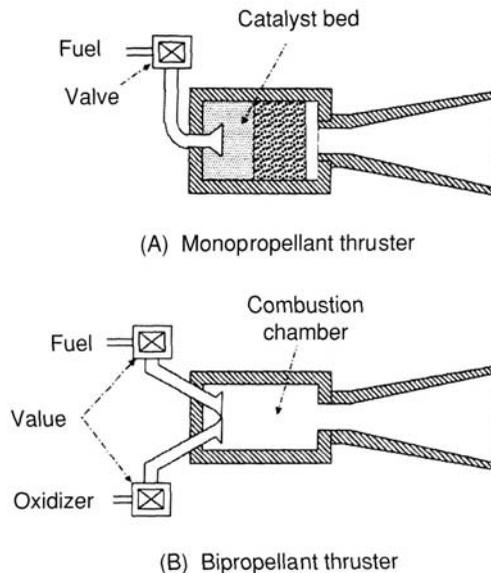
Figure 8.19   Propellant thrusters employing monopropellant and bipropellant technologies.

outweigh the costs for larger spacecraft and with missions that extend beyond 10 years.

8.3.5 Electric and Ion Propulsion

Electric propulsion, which depends on power derived from the solar panels to produce most of the thrust, is being used in commercial missions, particularly at GEO. The specific technologies under this category include the following:

- ARCJET generates intense heat using between 1 kW and 20 kW of prime power to produce a wide-field, high-energy exhaust plume using ammonia as the elemental fuel. The engine incorporates a nozzle to control the exhaust plume (while the other electric propulsion technologies, discussed below, do not). I_{sp} available is in the range of 500 to 800 seconds, and thrust level is approximately one order of magnitude below (1/10) that of hydrazine.
- Pulsed plasma thruster (PPT) utilizes Teflon propellant to provide over 1,000 seconds of I_{sp} while operating at power levels of 100–200W. Thrust can be conveniently adjusted by varying the pulse rate. PPTs have demonstrated thrust levels of only three orders of magnitude below (1/1,000) standard hydrazine thrusters.
- Hall thruster (HT) uses a low-pressure propellant gas that supports an electric discharge between two electrodes. A radial magnetic field generates a Hall-effect current that reacts with the magnetic field to place a force on the propellant in the downstream direction [4]. These devices are developed and produced in Russia and deliver I_{sp} of the order of 1,600 seconds.
- Ion propulsion is not a new technology; devices in this class have existed since the early 1960s. Ion propulsion carries very little fuel, relying instead on acceleration of a plasma to very high velocity. I_{sp} up to 3,000 seconds at a power of 1 kW represents the potential. Current devices deliver about 2,500 seconds and are capable of operating throughout a typical GEO mission of 15 years. Thrust level is quite low, requiring that the thruster be operated for extended periods at a time to provide sufficient velocity deltas.

The basic arrangement of an ion thruster is shown in Figure 8.20. That particular model ionizes elemental Xenon gas propellant to form a charged plasma. Thrust is produced by accelerating the plasma in an intense electric field, causing gas molecules to be expelled at high velocity. The thruster nozzle can be directed along two axes to facilitate precise pointing of the entire spacecraft. Typically, a pair of thrusters and control systems ensure redundancy. It can be attractive to use ion propulsion to augment the launch rocket for raising the final orbit because the I_{sp} of ion propulsion is substantially higher than any chemical system. The consideration here is the extra time (typically months) needed to achieve final orbit because of the low thrust force afforded by ion propulsion compared to chemical systems.

In GEO systems, being able to maintain orbit position throughout its life is critical to a satellite's operating success. There are significant benefits to reducing the weight of stationkeeping fuel, which represents a substantial part of the mass to be delivered to orbit. Even in the absence of stationkeeping, electric propulsion
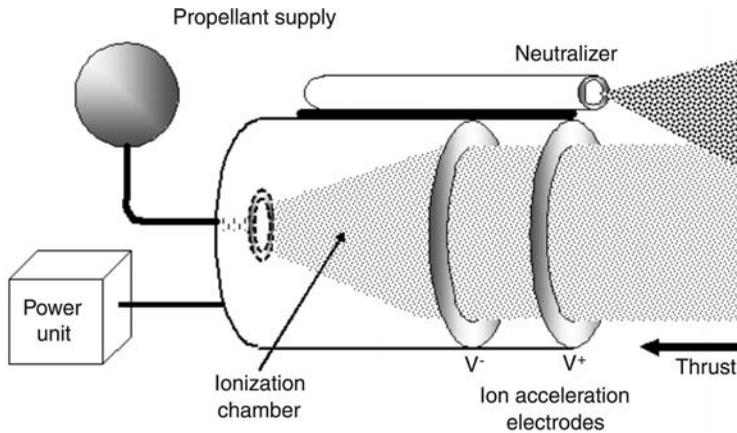
Figure 8.20   The principle of ion propulsion.

improves the price and/or the performance of satellites that must be raised from the initial launch vehicle altitude to the final orbit. There are many considerations in the use of electric propulsion, the most significant of which are the lifetime of the thruster itself, the possibility of contamination of various critical spacecraft surfaces from thruster effluent, and the need for high voltage and power to operate the thruster itself. In particular, the wide plumes of the engines result in cant angles of 45 degrees or more, reducing the achieved I_{sp} and complicating the orbit and attitude control methodology.

8.3.6   Solid-Fuel Rocket Motors

Many generations of GEO satellites used solid-fuel rocket motors for the injection into the final circular orbit. We provide a review of this approach, although most current designs employ liquid fuel for the reasons given in Section 10.1.3 and summarized below.

Solid-fuel rocket motors are filled with a relatively hard, rubbery combustible mixture of fuel, oxidizer, and binder. When ignited by a pyrotechnic device in the motor case, the mixture burns rapidly, producing intense thrust lasting 1 minute, more or less. Figure 8.21 shows a cross-section of a typical small solid motor. This type of rocket motor delivers an I_{sp} of approximately 300 seconds and is suitable for major orbit changes, such as for perigee or apogee kick. Redundancy is applied to the igniter, as shown. The case is constructed of titanium, although high-strength fibers can be used when wound in the proper shape. Attached at the exit is a nozzle assembly made of mixtures of carbon and other materials. During firing, the nozzle gets red hot from the heat of the expelled gasses and is eroded by the particles in the combustion products. Nozzle design is as much an art as a science, and the importance of quality control and testing cannot be overemphasized. Failures of such nozzles in the past have left satellites in unusable orbits.

As indicated in Figure 8.17, a single solid motor can be integrated into the spacecraft and the empty case carried for the remainder of the satellite's lifetime. Other solid motors can be attached at the bottom and discarded after use. There
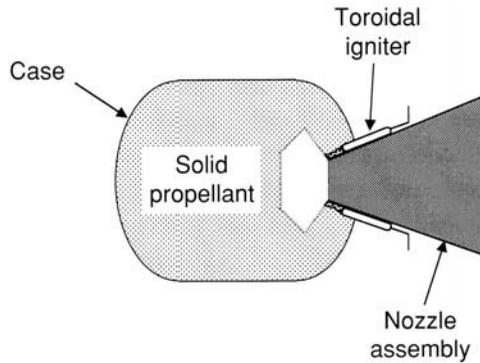
Figure 8.21 Typical solid fuel rocket motor to be part of the spacecraft for use in apogee kick or perigee kick.

has been a recent trend to use relatively large liquid bipropellant motors that may or may not be part of the RCS because such a motor can be restarted and fired at successive apogees. Solid motors will, however, still find application during launch and orbit change simply because of their thrust, efficiency and simplicity.

8.3.7 Tracking, Telemetry, and Command (TT&C)

The communications payload and spacecraft bus, although designed to serve for a period of many years, require ground support and intervention even under normal circumstances. Consequently, there is need for a two-way system of radio communication to allow ground computers and support personnel to determine the status of the spacecraft and to control the various subsystems of the payload and the bus. That is the purpose of the TT&C operations, which, as illustrated in Figure 8.22, consist of both ground and space elements.

The TT&C ground station and its associated SCC provide a sophisticated facility consisting of RF equipment, baseband and digital processing equipment, computers, and operator display consoles. (All aspects of the TT&C ground facilities are described in Chapters 9 and 11.) The RF equipment is in the form of a fairly typical Earth station with the exception that the antenna has the ability to track the satellite in transfer orbit and to measure its position throughout the mission. Commands generated by operators or computers are sent to the satellite on the command uplink. Onboard the satellite, the commands are decoded and activate some aspect of satellite operation. The telemetry downlink carries what is called housekeeping data, which are measurements of temperature, electrical voltage, RCS tank pressure, and other important parameters. Telemetry data are received by the ground facility, processed in the computer system to provide a current picture of satellite health, and stored for evaluation of long-term performance.

Most communications spacecraft employ two different antenna systems for TT&C. As shown in Figure 8.22, an omnidirectional antenna (discussed later on in this section) is used during transfer orbit, at which time the main reflector is not pointed toward the Earth. The high-gain (reflector) antenna subsequently is
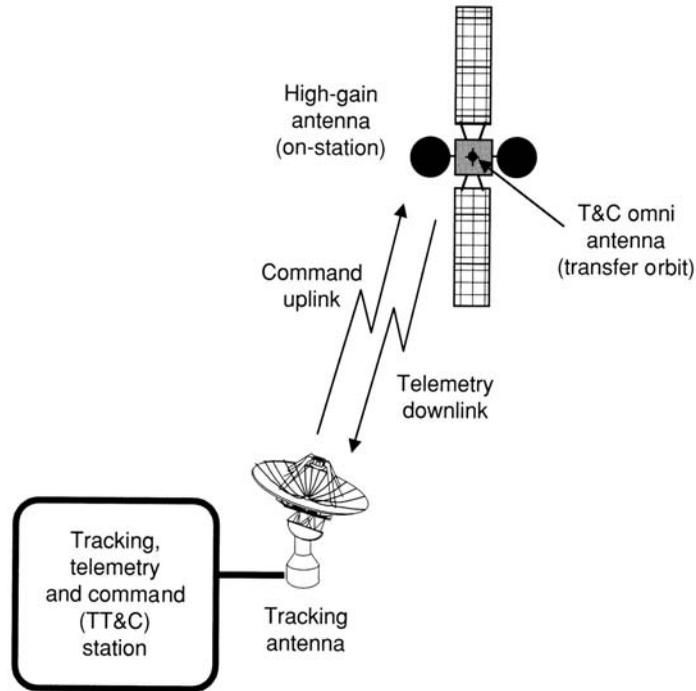
Figure 8.22 TT&C operations from the ground using the omnidirectional and high-gain antennas of the satellite.

used because it permits the TT&C station to reduce its power significantly. In the case of RF tracking, the high-gain antenna includes the feedhorns, which simultaneously receive the command carrier and generate pointing error signals for use by the ACS. LEO spacecraft may rely on omnidirectional antennas through-out their mission due to the close range and rapid motion relative to TT&C ground stations.

8.3.7.1 Satellite Tracking and Ranging

One important function of the TT&C station is to provide for the measurement of satellite position. The specific techniques employed are called tracking and ranging, illustrated in Figure 8.23. For a GEO satellite, it is a relatively simple matter to fix its position by measuring the azimuth and elevation of the Earth station antenna when its beam is aligned to receive the telemetry signal. Note also at the top of Figure 8.23 that there is only one unique value of range (the distance to the satellite from the ground station) for a given GEO satellite once we know whether it lies to the east or west of the Earth station's longitude. The precise range is obtained by measuring the time delay at the TT&C station between transmission of a reference signal (like the pulse in the bottom drawing) and its reception after being relayed by the range and satellite. Taking into account many factors in the link, the range can be calculated from the delay because the speed of propagation is fixed at the speed of light. The amount of time delay through
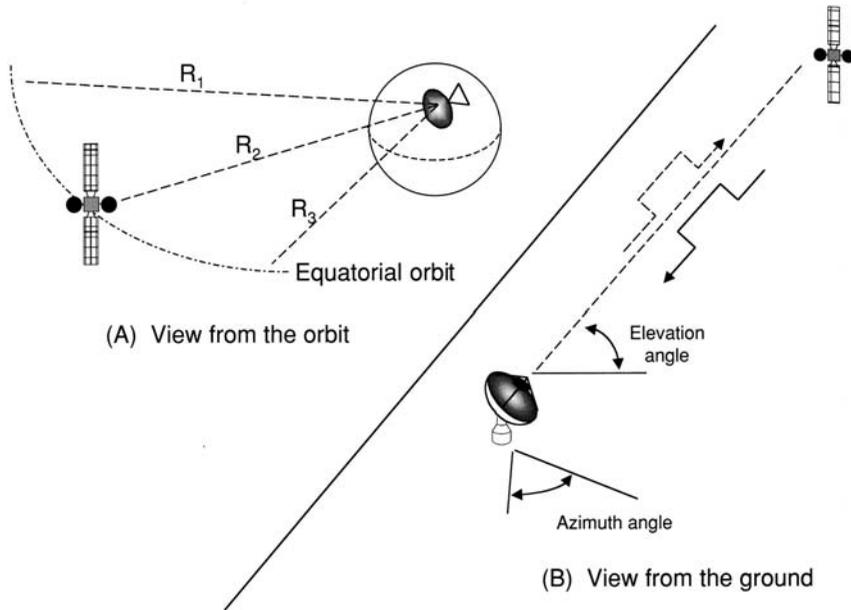
Figure 8.23 (a, b) A single TT&C Earth station measures the range and direction to the satellite so the satellite's position can be computed.

the satellite and Earth station equipment, which must be removed from the measurement, is determined by ground tests and by long-term analysis of ranging data. The combined accuracy of range measurement for GEO satellites typically is 25m or less.

Tracking and ranging of non-GEO satellites apply the same measurement techniques. Because the associated ground stations require multiple tracking antennas to ensure continuous communication, it becomes a relatively simple matter to include tracking and ranging as added functions. The data subsequently are relayed by suitable communication means to the SCC, where the satellite performance and location are determined. Position accuracy requirements depend on the particular mission but could be relatively simple compared to GEO because all ground stations are required to track moving satellites.

8.3.7.2 Spacecraft Telemetry System

The telemetry link between the spacecraft and the ground station is used to provide data concerning the status and health of the payload and bus subsystems. To get a feel for how the data are gathered onboard the spacecraft and transmitted to the ground, examine Figure 8.24, which is a simplified block diagram of the telemetry system. At the left are examples of some of the critical elements of the spacecraft: a battery, a fuel tank, and an RF power amplifier. Measurement of a key parameter of each is made through a transducing device, which senses the parameter and produces a corresponding output voltage. Battery voltage is taken directly and adjusted to a standard level range with a set of voltage-dropping resistors. A pressure transducer is connected into the fuel lines to provide information about the
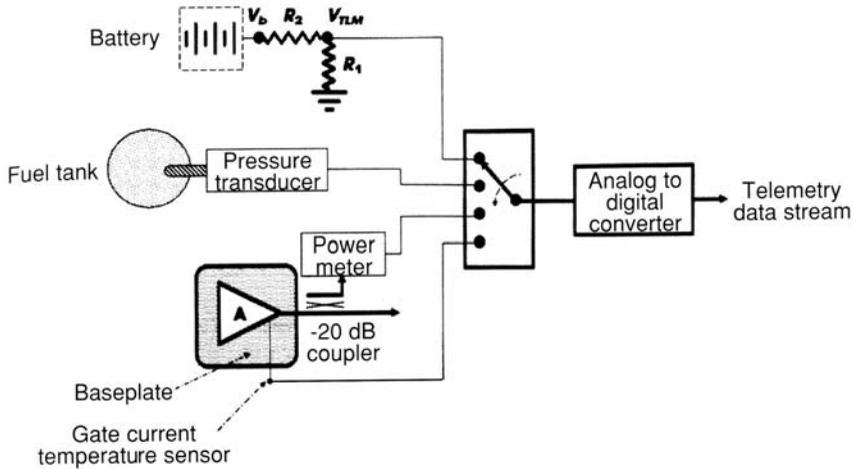
Figure 8.24   The spacecraft telemetry subsystem makes analog measurements and converts them to digital data before transmission to the ground.

amount of fuel and pressurant gas remaining. The device outputs a corresponding voltage, which also can be telemetered. The last example in Figure 8.24 is an SSPA for which we might want to telemeter the output power and GaAsFET gate current. A power meter detects a small amount of RF output power taken from a 20-dB coupler (that steals only 1% of the total output of the amplifier, equivalent to an output loss of 0.04 dB) and again produces a corresponding voltage to be telemetered.

The simple system depicted in Figure 8.24 requires the transmission of four channels of telemetry. On a real spacecraft, several thousand separate channels could be involved, some being simple on/off indications and others being complex analog measurements, like the four previously discussed. Therefore, time division multiplex (TDM) is used to combine the channels into a single stream for downlink transmission. While some of the early satellites transmitted the analog samples directly (i.e., pulse amplitude modulation), all modern commercial telemetry systems encode the samples using binary PCM. That is the same technique used in digital voice and data communication systems.

The output at the right of Figure 8.24 is, therefore, a multiplexed stream of digitized telemetry samples repeated in intervals of typically 1 second or multiples of 1 second. The frame structure allows some parameters to be measured more frequently than others (the A/D converter does the conversion from analog samples to PCM). In addition, it might be possible to add parameters when required for mission support. That could include the ability to increase the frequency of measurement of critical parameters, particularly if an anomaly is detected. All the telemetry information is transmitted to the ground, where it is recovered by the baseband and digital processing equipment in the TT&C station.

It was mentioned previously that reception of non-GEO satellite telemetry must, of necessity, be intermittent. That limitation is overcome by storing vital telemetry data onboard and downloading when in view of a suitable ground station. In contrast, Iridium satellites provide continuous telemetry by virtue of the cross

links. Systems engineers need to determine the mission requirements for telemetry coverage as part of the specification of in-service operations concepts and, ultimately, the design of the spacecraft and TT&C network. Some approaches in that regard were suggested at the beginning of this chapter and are reviewed further in Chapter 11.

8.3.7.3 Spacecraft Command System

The command link is received onboard the satellite and processed by the spacecraft command system. Unlike the continuously operating telemetry system described above, the command link need only operate when the spacecraft must be actively controlled. The one exception is the RF tracking system in which the command link is continuously on the air. While the basic structure of the command link is reasonably secure, many TT&C systems incorporate various forms of added protection, such as command encryption to prevent unauthorized parties from taking control of the satellite.

Shown in the simplified diagram of Figure 8.25 is an example of how the command system might control some elements of the spacecraft (in this case, the payload). Each unit to be controlled has associated with it a control driver, which can be as simple as the single transistor circuit shown at the top. The control driver requires two inputs before it will activate the unit under control. First, the COMMAND line from the spacecraft command decoder must go "high" (meaning that a positive voltage must be applied to the base of the transistor). Then, while COMMAND remains high, the EXECUTE line must be grounded so the transistor will conduct. Under those circumstances, the output to the unit under control will change from the high voltage, V, to near ground potential, activating the unit. The
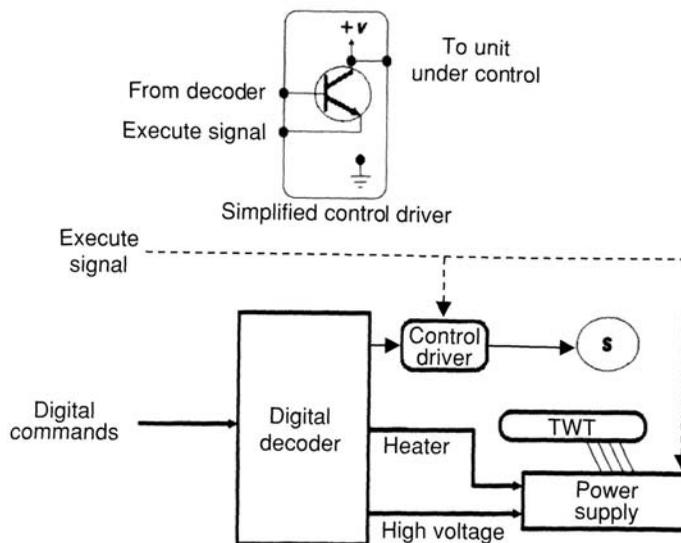
Figure 8.25 Simplified block diagram of the spacecraft command subsystems showing how the COMMAND and EXECUTE functions are used.

reason the dual process is used is to provide an extra safeguard against operator mistakes and equipment anomalies.

Just how the COMMAND and EXECUTE signals are processed is depicted in the lower part of Figure 8.25. Each function to be controlled has assigned to it a unique digital command word, like an IP address. Those words are maintained in a list that is unique to the particular type of spacecraft. To protect against unauthorized access, the words can be encrypted prior to uplink transmission and subsequently decrypted in the spacecraft. Commands are selected by the spacecraft control operator or software procedure, transferring them to the command link. In the spacecraft, the specific command is identified by the command decoder, which has a direct and unique connection to each control driver. Three typical examples are shown at the right. The S switch can be commanded between its two positions using its control driver. Activation of the TWTA requires two separate ON commands, first for the heater and then (after a predetermined warmup period) for the high voltage.

The standard procedure for commanding a unit ON is as follows. First, the uniquely defined command word is sent to the spacecraft, where it is received, decoded, and retransmitted to the ground via telemetry for verification. The ground station computer automatically checks to see that the received command word is the same as that which was sent. At that point, the command operation can be sure the spacecraft is ready for the EXECUTE, independently of the COMMAND. Only that unit for which the proper command word was sent will respond to the EXECUTE. Autonomous onboard verfification of the command (with retransmission to the ground) is accomplished through a rigorous data integrity check in the command decoder similar to the CRC sequences used in computer communications. Most spacecraft use stored commands, which are either maintained in computer memory onboard or uploaded ahead of the desired function. That involves command sequences needed to alter spacecraft configuration or even orbit.

8.3.7.4   TT&C Omnidirectional Antenna

The TT&C system on a spacecraft may use two different antennas during the mission lifetime. For most of the initial operations in transfer orbit and before the main communications antenna is functional, the TT&C omnidirectional antenna provides all access to the spacecraft. Similar comments are appropriate for non-GEO satellites, which could, because of shortened path length, employ omniantennas for an entire mission.

An ideal omniantenna would function like a point source of light, radiating equally in all directions. That is a physical impossibility for a real antenna, due to blockage from the spacecraft body and its appendages, coupled with the difficulty of designing an appropriate microwave structure. Figure 8.26 shows some more practical antenna patterns currently used on commercial satellites: the toroidal and the cardioid. Placement of the omniantenna on a three-axis spacecraft can, at times, be difficut. That is due to the potential RF blockage from reflectors, solar panels, and other appendages. The solution has been to incorporate two or more cardioid or horn antennas that combine signals in the command receiver or telemetry transmitter to produce the desired near-complete coverage. That is particularly critical
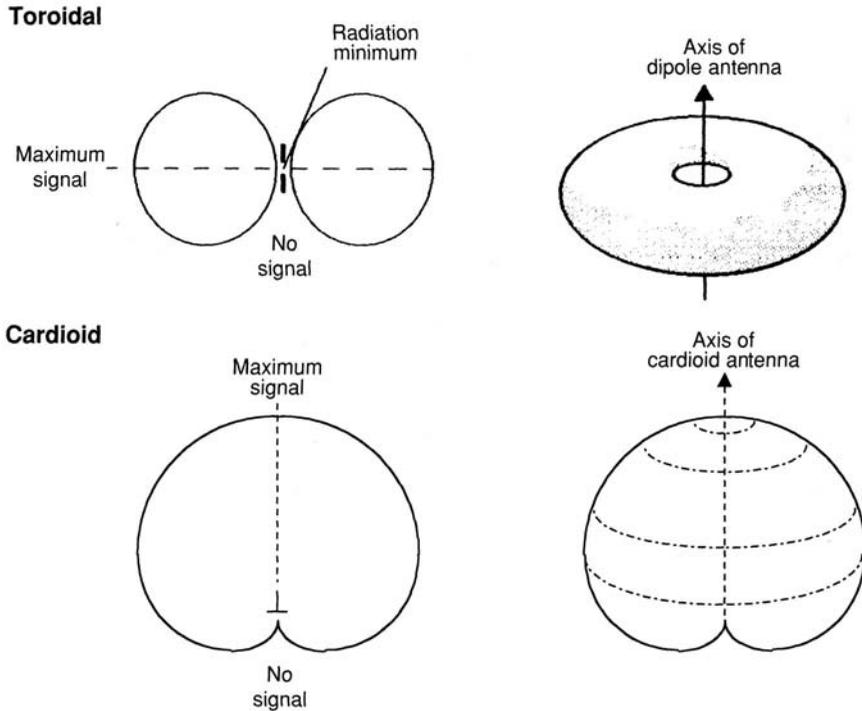
Figure 8.26   Radiation patterns of near-omnidirectional antennas.

because spacecraft may experience off-pointing during an anomaly, requiring careful monitoring and fast response by ground controllers. In the absence of adequate coverage, the spacecraft may have to rely on preprogramming of the control processor to try to put the spacecraft into a suitable attitude for TT&C operation.

8.3.8   Thermal Control

One of the little appreciated yet vital areas of spacecraft design is thermal control. Space can be a harsh environment in terms of temperature, both hot and cold. Active components would prefer to operate at or near room temperature (22°C), while external appendages, like reflectors, feeds, solar panels, and thrusters, can expect to experience temperature extremes of −100°C to +100°C. Therefore, the thermal control system represents a common denominator for all operating elements of the spacecraft. Unexpectedly high temperatures cause great concern for such components as moving bearings and solid state electronic devices, which may have been designed and tested for a more nominal temperature range. Failure of the thermal control system can be caused by inaccurate analysis of heat distribution followed by insufficient ground environmental testing. Once a satellite reaches orbit, there is little that can be done to alter the thermal performance.

To get an appreciation for the elements of the thermal control system, examine the cross-section of a hypothetical spacecraft configuration shown in Figure 8.27. The picture does not represent a particular spacecraft design but shows the key components and their arrangement. There are two fundamental appoaches to
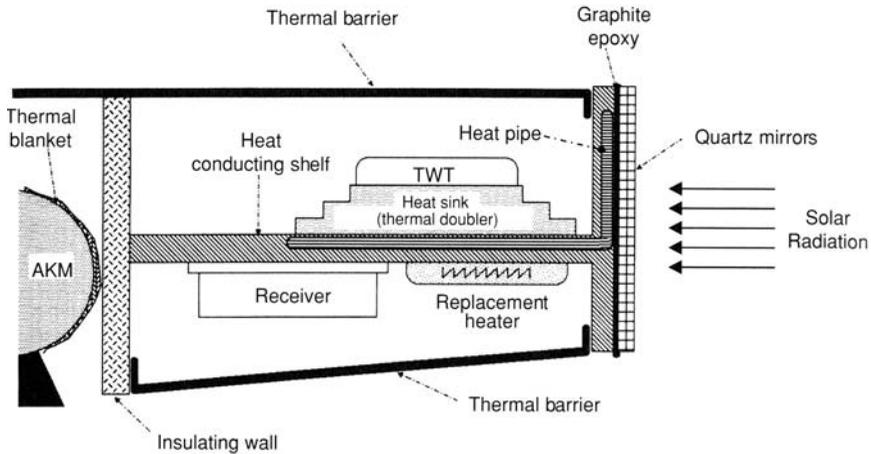
Figure 8.27   Overview of passive and active thermal control techniques, which rely on the conduction and radiation of heat.

thermal design: passive thermal control and active thermal control. Passive thermal control means that the desired temperature range is maintained using appropriate thermal surface coatings (paint or deposited material such as aluminum or gold), blankets, and mirrors. Those techniques are supported by simple conduction and radiation.

Active thermal control depends on thermostatically controlled electrical heaters or mechanical devices of one sort or another that can enhance or directly alter the thermal properties of the spacecraft. The primary example is the heat pipe, which uses a gas-liquid thermal cycle to move heat from a warm place (the interior) to a cooler place (exterior of the spacecraft). High-power GEO three-axis spacecraft in the 8-kW and above class rely on heat pipes embedded in the equipment-mounting surfaces. The absorptance and emissivity of the outer surfaces of the north and south panels of three-axis spacecraft are enhanced through the use of optical solar reflectors, which are specially designed quartz mirrors. The east, west, nadir (Earth-facing), and zenith surfaces are usually covered with multilayer insulation blankets. Satellites designed for GEO service with more than 10 kW are designed as closed cages of heat pipes.

Reviewing the elements of a passive design in Figure 8.27, the heat producing components (TWTs and other electronics) typically are attached to a heat-conducting surface such as an aluminum honeycomb equipment shelf. The primary source of external heat is the sun (shining on the spacecraft from the right), which must be precluded from raising the temperature in the compartment above acceptable limits. In addition, the heat generated in the compartment by the electronics must be transmitted to space, or it, too, will cause the internal temperature to rise. Both those necessities can be accomplished with a quartz mirror radiator located on the side facing the sun. The mirror acts like a filter, reflecting visible and ultraviolet radiation away and transmitting infrared radiation from inside to outer space. Mirrors have to be placed carefully on the spacecraft because any contamination or clouding of the surface during the life of the satellite reduces effectiveness.

Because of unavoidable degradation of the mirror surface, the temperature in the spacecraft gradually rises on a yearly basis, which must be taken into account in the overall design of the thermal control system and other components.

If the spacecraft contains an apogee kick motor (AKM), it is necessary to provide considerable thermal control measures for it. As shown at the left of Figure 8.27, the AKM is surrounded by a thermal blanket. Electrical heater wires can be placed under the blanket if needed. The purpose of these measures is to keep the motor and propellant temperatures above a prescribed minimum prior to firing. To protect the rest of the spacecraft from the heat of the motor firing, an insulating wall and thermal barriers are placed between the motor and the sensitive components.

A similar design consideration exists for the batteries, which need to be kept cooler and within a tighter temperature range than other equipment. On GEO satellites, the prevailing design is to give the batteries their own isolated thermal environment, usually at the aft end of the spacecraft.

In the thermally controlled compartment, the electronics receive and generate heat during their normal operation. It is the purpose of the thermal control subsystem to achieve thermal balance in the desired temperature range during seasonal variations and when units are individually turned off and on. The TWT shown in Figure 8.27 produces localized heat, which can be distributed more evenly by metallic heat sinking using thermal doublers (that technique, unfortunately, adds considerable weight to the spacecraft). The receiver mounted below it, therefore, is heated by the TWT, aiding in its temperature control. Thermal conduction is improved by installing heat pipes under the hot portion of the TWTA. If and when the TWT is not operating, the heat necessary to warm the receiver is provided by a replacement heater located in proximity to the units. Replacement heaters, operated either automatically or by ground command, are necessary to maintain an overall ''bulk'' temperature for the spacecraft.

High-power DTH satellites can be designed to enhance heat rejection by employing TWTs with direct radiating collectors. The collector end is exposed to space by mounting of the TWT on an externally facing panel of the spacecraft. As mentioned in Chapter 6, the collector develops about half the total dissipation of the TWT, so direct radiation is beneficial in terms of reducing thermal control system complexity and weight.

This discussion of the primary aspects of thermal design is intended only as an introduction. An important aspect to such study is the creation of analytical thermal models for use in predicting temperatures in the spacecraft at various critical points and times in the mission. The models must be verified by ground testing in a simulated space environment. Once in orbit, it is difficult to compensate for errors in analysis or design, although it happens all too frequently. That is why it is important to provide a thermal design with adequate margin and to use telemetry to measure the temperature of critical elements after the satellite is operating in orbit. Satellite operators need to keep a close eye on temperature trends through alert monitoring and careful recordkeeping. The data should be fed back to the spacecraft designer to help improve future models.
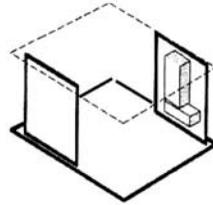
Structural panels in a modular three axis spacecraft

Figure 8.28   Basic structural arrangement used with three-axis spacecraft.

8.3.9   Structural Arrangements

The last area to discuss in spacecraft design is the structural arrangement or physical support system for the spacecraft. Figure 8.28 shows one arrangement for the three-axis.

The box structure is used for many three-axis satellites because it facilitates the attachment of solar panels and antennas. A monocoque cylindrical can be placed at the center to handle launch loads, although flat panels and struts may be preferable to maximize mounting area and volume. Equipment is mounted on panels, which are integrated with the spacecraft. A practice is to configure the structure as separable payload and bus modules to facilitate assembly and final integration.

Structural design of the spacecraft must consider all phases of the mission, ranging from buildup during manufacture, on-ground testing, transportation to the launch site, launch and final orbit injection, deployment on station, and operation for the lifetime of the satellite. A major consideration is that the spacecraft must fit properly within the launch vehicle shroud, as will be discussed in Chapter 10. All components must be placed appropriately in the spacecraft and be capable of operating properly under all expected conditions. The structure also functions as a shield cage to protect sensitive electronics from space radiation and ESD. Another area of complexity is the continuing drive to reduce weight because that is what determines lifetime (due to fuel availability) and cost (associated with launching of the satellite or constellation). There is a need to find lightweight materials, both metallic and nonmetallic, that can be properly fashioned into the structural components of the vehicle and withstand the rigors of space flight.

References

[1]   Larson, W. J., and J. R. Wertz, *Space Mission Analysis and Design*, Torrance, CA: Microcosm, and Dordrecht, the Netherlands: Kluwer Academic Publishers, 1992.

[2]   Smith, D., "Operations Innovations for the 48-Satellite Globalstar Constellation," *AIAA 16th Internat. Communications Satellite Systems Conf.*, Washington, D.C., February 25–29, 1996, Collection of Technical Papers, American Institute of Aeronautics and Astronautics, New York, paper no. 96-1051, 1996.

[3] Coates, D. K., et al., "The Design and Manufacture of Nickel-Hydrogen Batteries for Spacecraft Power Systems," *AIAA 16th Internat. Communications Satellite Systems Conf.*, Washington, D.C., February 25–29, 1996, Collection of Technical Papers, American Institute of Aeronautics and Astronautics, New York, paper no. 96-1023, 1996.

[4] NASA, "Overview of Electric Propulsion Devices and Power Systems," http://trajectory. lerc.nasa.gov/aig7820/projects/e.

Earth Stations and Network Technology

Earth stations provide access to the space segment, interconnecting users with one another and with terrestrial networks such as the Internet and the public telephone network. In the first decades of this industry, Earth stations were large, both in physical and financial terms. Individual ownership of a satellite terminal was within the reach only of major organizations and perhaps a few wealthy individuals. By the late 1980s, low-cost TV receive-only (TVRO) terminals cost as little as $1,000; many people who lived in remote areas were pleased to be able to receive commercial television programming through this means. During the 1990s and past 2000, we have the widespread adoption of low-cost user terminals (UTs) that provide access to two-way as well as one-way services. Interestingly, the electronic complexity of the modern UT is no less than what formerly took up several equipment racks in the large Earth stations.

The overall RF and baseband performance of the Earth station has been covered in previous chapters. In this chapter, much more will be said about the various Earth station architectures in relation to the types of services that are provided. Included are digital media content, data communications, and fixed and mobile telephony. A discussion is provided of the TT&C Earth station and satellite control facility, both of which are needed to implement and operate the space segment. They are not typically used for communication services, but their design and operation are no less important to meeting the technical and financial objectives of a satellite mission. Tradeoffs in design of the various types of Earth stations are also described. The terrestrial "tail" that connects the Earth station with the outside world is reviewed, as are considerations for implementing the building and other support facilities.

9.1 Basic Earth Station Configuration

For the purposes of this discussion, an Earth station can be broken down into the following major elements: the RF terminal (including the antenna), the baseband and the control equipment, and the user interface. In addition to these electronic systems, every fixed Earth station must be supported by a physical facility capable of housing the equipment, supplying reliable electrical power, and maintaining the temperature and humidity of the electronics within acceptable limits. These latter functions are similar to the ones provided by the spacecraft bus (while the functions in the previous sentence are comparable to those of the communications payload).

The key difference here is that the majority of operating Earth stations can be serviced by maintenance personnel and externally powered. That allows the larger Earth stations to be designed from building blocks that consume lots of power and are easily repaired or changed.

The previous remarks obviously apply to the traditional type of permanent Earth station such as what is used to uplink digital TV. Modern UTs, such as advanced TVRO terminals and MSS phones, are self-contained and some are battery powered. Just like consumer electronics, UTs of this class do not require servicing except for setup (such as a user would do with a mobile phone) and replacement of batteries. Repairs are performed by taking the unit to a repair facility or, in the case of the ultimate consumable electronic item, simply throwing it away as one would an obsolete PC or mobile phone. VSATs, however, fall in between the two extremes, and it is common for repairs to be made by roving technicians who are called on to bring a terminal back into service.

The major elements of a generic Earth station are illustrated in the block diagram in Figure 9.1. The dotted box is the RF terminal, encompassing the antenna, the HPA system, the LNA, and upconverters and downconverters. In the figure, the LNA and the HPA are connected to the cross-polarized ports of the antenna feed. (The operation and performance of those components were covered in Chapter 4.) Where an Earth station is used as a major hub of a network or is required to provide reliable service, redundant, or backup, equipment (as shown) is included in the RF terminal.

The RF terminal provides an intermediate frequency (IF) interface to baseband equipment that performs the modulation/demodulation function, along with base-
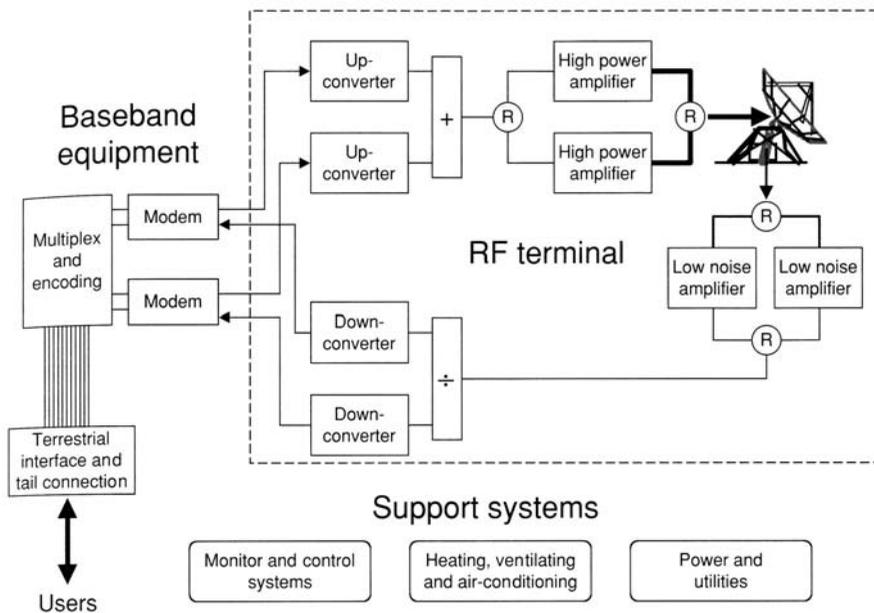
Figure 9.1   Major elements of an Earth station used to provide two-way digital communications links.

band processing and interface with the terrestrial tail. The specific configuration of the baseband equipment depends on the modulation and multiple access method employed (covered in Chapter 5). In Figure 9.1, baseband equipment consists of digital modems and a time division multiplexer. Multiple lines interconnect those facilities because there would be several independent data channels. Also, equipment redundancy in some form could be included. The last electronic element of the station is the terrestrial interface along with the terminal end of a tail connection of some kind. Interfaces must be properly engineered and specified so that communication services can be carried over the satellite link without disrupting the operation of the overall network. It is generally a good practice to standardize this interface to something common like Ethernet or the T1/E1.

The tail connection itself would be composed of a fiber optic cable or microwave link, extending to the point of service such as an office building, telephone switching office, or television studio. Tail connections deserve more attention than they usually get because they affect quality and reliability in a direct and significant way. Poorly engineered or installed tails introduce noise, bit errors, and, in the case of analog voice, echo. Furthermore, the high reliability of the satellite-to-Earth station path is easily compromised by the susceptibility of buried cable tails to breakage, deterioration, and electrical discharges (due to power line faults and lightning). These remarks about the terrestrial interface and tail do not apply to VSATs and UTs, because they typically are self-contained and may not need to extend the interface to external devices (if they do, the interfaces again need to be properly considered).

The power and the building facilities for the Earth station (analogous to the bus of the spacecraft), often get overlooked but are important nevertheless. Prime ac power for the station usually is obtained from the local electric utility company. It is common to utilize battery backup for critical parts of a station that provide vital communications service. To do that, a multiple sequence of ac-to-dc-to-ac conversion is followed. For example, in a typical uninterruptable power system (UPS), commercial power (or a diesel generator in the case of a remote installation) is converted in a rectifier to dc and tied to a lead-acid battery system capable of powering the essential electronics (called the critical load). Nonessential loads, like office lighting and air conditioning, are not protected in this manner. In the event of an outage, nonessential loads would be transferred to a generator.

The function of the UPS is available in a small package suitable for a VSAT-type installation. It consists of the ac-dc-ac electronics conversion system plus battery backup. A diesel generator is not appropriate because of the desire for compactness and low cost. However, it would be entirely feasible to provide enough backup battery power to supply the terminal for many hours. That usually is sufficient time for commercial power to be restored or to connect a portable emergency generator.

A properly engineered facility also provides heating and cooling to maintain the equipment within its prescribed operating temperature and humidity range. Electronic racks often are mounted on top of a raised floor (called computer flooring) so that conditioned air can be conveniently fed from below; that also allows cables and waveguides to be neatly routed between racks. When done properly, a large Earth station facility looks attractive and allows easy and conve-

nient maintenance of the electronic equipment. A small Earth station such as a VSAT, on the other hand, can be treated like telephone or data processing equipment installed in an office building. Building modifications may be needed to access the outdoor RF terminal, often roof mounted, and reliable power with UPS provisions may still be required.

Special consideration should be given here to the RF terminal, which often is left outside in the elements. The RF equipment could be contained in a waterproof box similar to the cabinets used by telephone companies for street installation. As long as the door is closed, this type of container can withstand almost anything nature can throw at it (except a massive earthquake or hurricane). The problem can come if the equipment must be serviced during bad weather, particularly heavy rainfall and wind. The high voltages and ac power that may exist in the box should not be exposed to the elements or to unprotected human beings. The solution is to utilize an equipment shelter large enough for the service person to work inside and be protected from the weather. Also, at least two people should work on high-voltage power supplies and amplifiers, with one person observing and ready to disconnect the power in case of threat to life.

Monitoring and control (M&C) functions of an Earth station allow local and remote control operation. M&C capabilities are advancing rapidly and involve sophisticated networks of their own. Earth stations that incorporate routers and TDMA require a centralized system of control to ensure that the proper channel and data packet routing are established. That can be done once or twice daily in a preprogrammed fashion or may be changed on user demand (i.e., demand assignment). The operation and status of RF and baseband equipment also can be determined, and in the event of failure, redundant equipment can be switched by remote control. In principle, M&C is to an Earth station what TT&C is to the satellite itself.

9.2   Performance Requirements

The performance requirements of the Earth station mirror those of the satellite. That is to ensure that there is a satisfactory RF link between the ground and the space segments under all expected conditions and for the range of required services. In addition, the Earth station determines the baseband quality and much of the end-to-end communication performance of the services being provided. We first consider the key parameters of EIRP and G/T as they relate to the ground.

9.2.1   Transmit EIRP

The Earth station EIRP simply is the value that applies to the uplink. As with the satellite (downlink) EIRP, it is obtained by multiplying the antenna gain (as a ratio) by the transmit power at the input to the antenna. In decibel terms, it is the antenna gain plus the transmit power in decibels relative to 1W (dBW). As shown in simplified block diagram in Figure 9.2, the EIRP also considers any RF loss between the HPA and the antenna. The loss factor, l_t, is a number greater than 1, and the ratio $1/l_t$ is actually the transmission factor (e.g., a gain less than 1). Major Earth
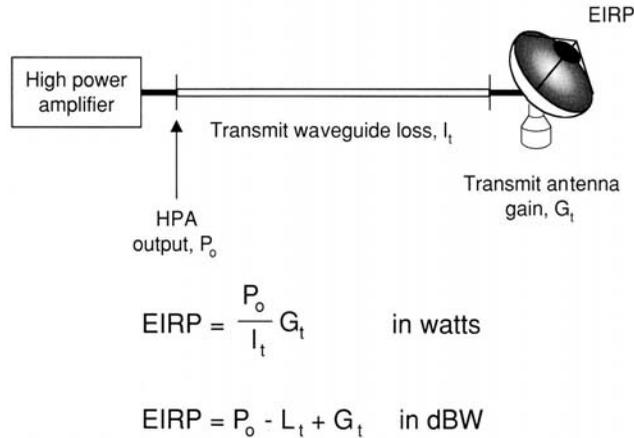
Figure 9.2 Definition of Earth station transmit EIRP, indicating the primary contributors to performance.

stations typically are designed for reliable uplink operation with low outage due to rainfall. Therefore, the required uplink EIRP is substantially higher than that of the downlink (possibly including uplink power control). That may require the use of vacuum tube HPAs capable of transmitting several kilowatts or at least a substantial fraction of 1 kW. The antenna gain will, of course, depend on the uplink frequency and the size of the antenna. While the behemoth 30-m antennas of the early INTELSAT system no longer are being constructed, it is common to find antennas in the 7-m to 13-m class for service at C-, X-, Ku-, and Ka-bands.

VSAT networks are designed to not require vacuum tube HPAs at the remote sites, favoring instead lower power SSPAs. EIRP for a hand-held UT is going to be low because of low transmit power and low antenna gain. In non-GEO systems, the satellite can be literally anywhere in the sky, so the UT antenna should provide nearly uniform radiation in all directions. The expected elevation angles toward the satellites as they pass overhead, as well as the variety of physical orientations that the user might impose, must be considered.

High-power uplinks bring with them a conern about radiation hazard to humans and other living things. That is one of the reasons major Earth stations are not located in highly populated areas or are shielded from such exposure. Careful consideration must be given to that aspect of safety. Low-power uplinks in VSATs and MSS terminals may pose less of a risk but still should be considered potentially harmful to humans. Handheld UTs expose the user to direct radiation, but power is kept to levels as low as typical cell phones. This is an area of active research and is still controversial with regard to the operation of common cell phones, which transmit similar power levels. In all cases, appropriate research and testing are prudent.

9.2.2 Receive G/T

The figure-of-merit of overall Earth station performance in the downlink is the G/T. G/T is the ratio of the antenna gain (itself as a ratio) to the total system noise

temperature (in Kelvin, which is with respect to absolute zero) of the receiving Earth station, converted to decibels per Kelvin. The contributors to G/T are indicated in Figure 9.3. Again, l_r is a factor greater than 1, and $1/l_r$ is the transmission factor (e.g., less than 1). Higher antenna gain improves sensitivity as does lower noise, hence the use of a ratio. The G/T of the station is used in link budget calculations (discussed in Chapter 4) to compute the ratio of carrier to noise. In general, Earth station downlink performance can be adequately measured and optimized by consideration of the G/T figure.

The figure of merit (G/T in dB/K) of a C-band Earth station is plotted in Figure 9.4 as a function of antenna diameter for three common levels of LNA noise temperature: 30K, 70K, and 110K. We assume that the antenna temperature in each case is a typical value of 40K, common for parabolic reflectors at elevation angles greater than about 20 degrees. The other factors that make up G/T consist

$$\frac{G}{T} = \frac{\dfrac{G_r}{l_r}}{\dfrac{T_a}{l_r} + \left(1 - \dfrac{1}{l_r}\right)\cdot 290 + T_{re}}$$

Receive antenna gain, G_r

Antenna temperature, T_a

Receive waveguide loss, l_r

Low noise amplifier

Receiver noise temperature, T_{re}
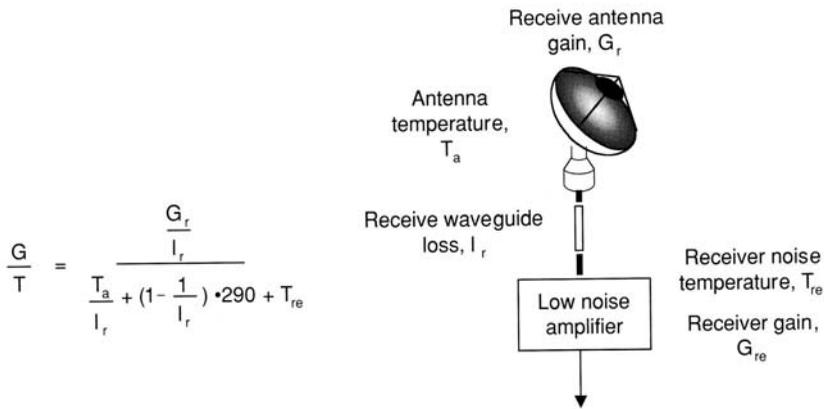
Receiver gain, G_{re}

Figure 9.3   Definition of Earth station receive G/T, indicating how noise contributes to performance. Receiver gain, G_{re}, is sufficiently high to reduce the downstream noise (due to down-conversion and so forth) to less than the equivalent of 1K.

System noise temperature

80K
120K
160K

Receiving system figure-of-merit, G/T, dB/K

Antenna diameter, meters
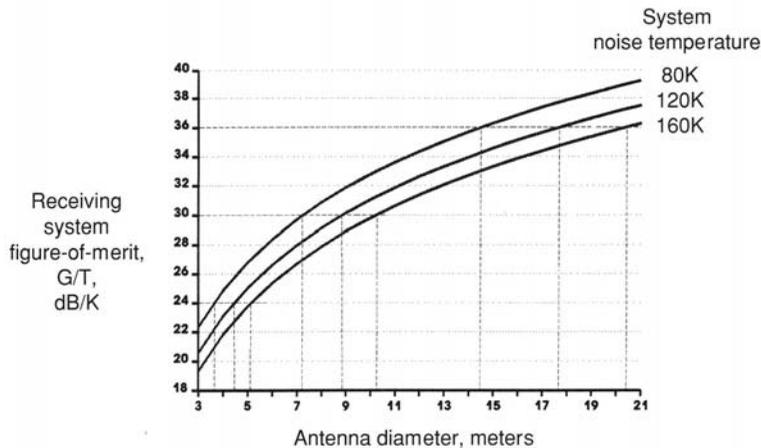
Figure 9.4   Earth station receiving figure of merit (G/T) at C-band as a function of antenna diameter and system noise temperature.

of the noise contribution and loss of the connecting waveguide (assumed to be 0.2 dB), as indicated in Figure 9.3. An observation from Figure 9.4 is that 40K reduction in system noise temperature (i.e., selecting an 80K LNA over a 120K model) allows the use of a 7-m antenna instead of a 9-m antenna.

Larger diameter antennas are expensive to build because of the structural strength needed to maintain reflector alignment, particularly in high wind. Also significant is the effect on gain of surface accuracy, which is the accuracy with which the surface fits a true paraboloid (e.g., the surface obtained by rotating an ideal parabola about its axis). Likewise, the mount and the motor drive must withstand greater antenna weight and wind load to keep the beam pointed in the proper direction. The cost of an antenna increases rapidly with the diameter, following an exponential relationship. For example, doubling of the diameter from 5m to 10m (a 6-dB increase in gain) can increase cost by a factor of 10 or greater. Cost consideration generally encourages the system designer to use the smallest antenna with an appropriate LNA to achieve the necessary G/T. If the station is to transmit as well as receive, there is a lower limit on antenna size imposed by the highest available HPA power. Another consideration is adjacent satellite interference, where a larger antenna could be required for isolation reasons.

The G/T performance of small antennas is defined in precisely the same manner as for the large Earth station. Terminals with reflector antennas, like VSATs and TVROs, literally are miniature versions of their bigger cousins but are only able to deliver G/T consistent with their receive gain. A more complex situation exists for the hand-held UT to be used in non-GEO MSS services, since the antenna points in many directions at the same time. Here, the antenna temperature often is over 100K, due to pickup from the ground and local obstructions. The G/T is a negative number due to low antenna gain.

9.2.3   Location and Platform Requirements

Satellite communications has greatest benefits for remote, portable, and mobile applications. Therefore, the Earth station may need to be designed to allow flexibility in its operation no matter where it might be located. The industry has produced a wide variety of supports and integrated packages that allow the user to obtain broadband satellite communications almost independent of location.

Portable terminals like the one shown in Figure 9.5 can be carried on board a vehicle or airplane and set up in the field. Once power is connected, this terminal can be commanded by the operator to automatically acquire a GEO satellite. In Figure 9.6 is a shipboard terminal with similar communications capabilities. In addition to auto-acquire, the shipboard terminal will also correct for ship direction and motion.

While satellite newsgathering (SNG) vehicles have been used for nearly three decades, they have moved from large vans to something compatible with an SUV. Naturally, modern SNGs have the same auto-point capability as the portable and shipboard terminals already described. An approach used by the military employs this type of antenna because it can be operated while the vehicle is in motion. A prototype of such a system is under the radome of the vehicle shown in Figure 9.7.

Figure 9.5   Portable Ku-band Earth station for two-way broadband services; this terminal is capable of automatically acquiring a satellite once set up.

Figure 9.6   A shipboard Earth station capable of providing broadband service under normal sea conditions. (Courtesy of SeaTel.)

Figure 9.7 Ku-band Earth station mounted on a military troop carrier capable of communications on the move. (Courtesy of Datron L-3 Communications.)

9.3 Radio Frequency Equipment

A simplified block diagram of the typical RF portion of the Earth station is shown in Figure 9.8. In it, we see many of the components of the standard bent-pipe type
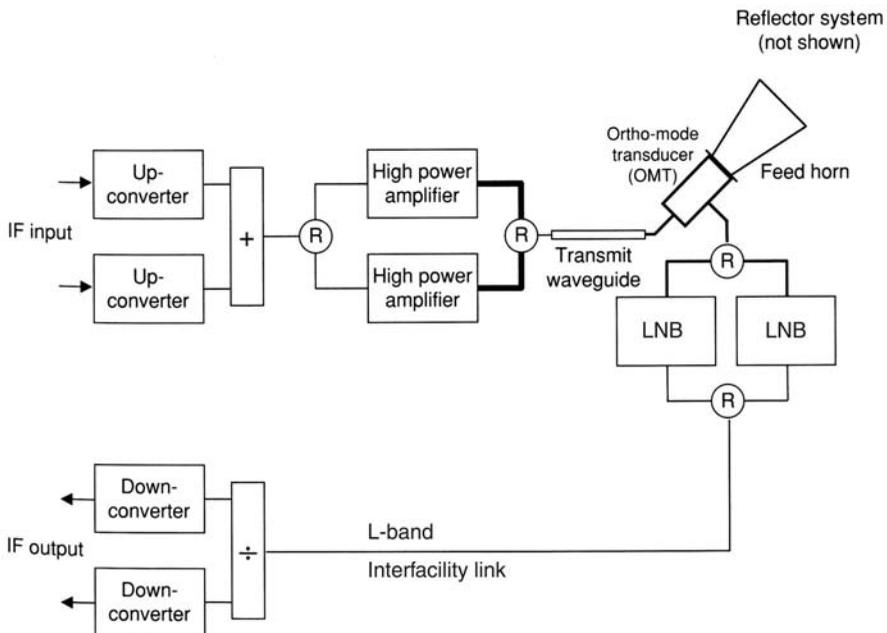
Figure 9.8 Block diagram of a typical RF terminal.

of repeater: a receive chain consisting of the antenna, LNB, and downconverter, and a transmit chain consisting of an upconverter, an HPA, and, again, the antenna. It is almost always the case that a transmit Earth station uses the same antenna to receive, since we are establishing communication with a common satellite. The LNB and HPA are connected to the same feed horn using a piece of waveguide plumbing called an ortho-mode transducer (OMT). The OMT then derives linear horizontal and vertical polarization for the antenna. The upconverters and downconverters interface with the rest of the station equipment at the IF, which typically is 70 MHz. An alternative is to package a block upconverter (70 MHz to L-band) with the HPA to allow the use of low cost coaxial transmission line for uplink as well as downlink. The upconverters and downconverters provide the translation from IF (typically 70 MHz) to the necessary RF frequencies assigned for the uplink and the downlink, respectively. An exception to using the standard IF is likely for compact UTs such as applied in MSS because of the potential for economies in the details of the internal design. For example, it might be possible for the modem to transmit directly on the uplink L-band frequency. The RF terminal often is contained in a separate shelter or cabinet adjacent to the antenna. That reduces cable and waveguide losses, which is desirable from the standpoint of EIRP and G/T.

9.3.1 Antennas for Earth Stations

A considerable amount of technical information on the properties and performance of antennas was given in Chapters 4 and 7. In this section, we focus on the physical characteristics of antennas specifically used in Earth stations: the prime focus fed parabola and the cassegrain, shown in Figure 9.9(a) and Figure 9.9(b), respectively. An alternate feed system called the splash plate is shown in Figure 9.9(c). As discussed in Section 9.3.1.3, a splash plate is a flat reflecting plate introduced into the antenna, usually to allow a different mounting geometry.

9.3.1.1 Prime Focus Fed Parabola

Antennas with a diameter less than 4.5m and particularly those used in RO service may employ the prime focus fed configuration shown in Figure 9.9(a). Within this
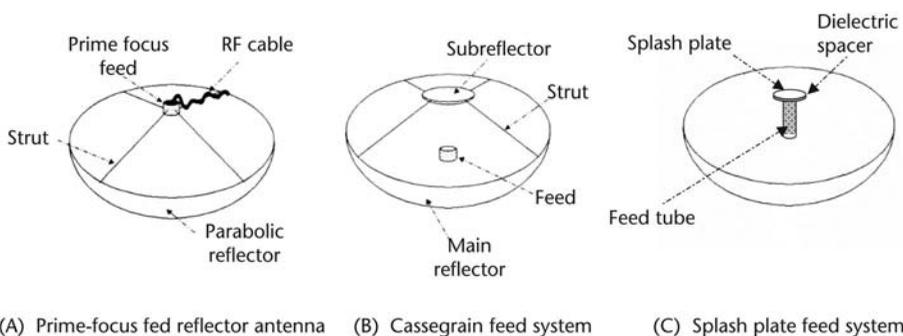
Figure 9.9 Earth station antenna configurations commonly used for receive and two-way communications links. (a) Prime-focus fed reflector antenna, (b) Cassegrain feed system, and (c) splash plate feed system.

class is the offset fed parabaloid, used extensively for antennas less than about 2m in diameter. The cost usually is the lowest because of its simplicity and ease of installation. To minimize RF loss and noise, a compact LNB package is located directly on the waveguide port of the feed to minimize loss between the front-end amplifier and the feed. The RF cable to connect between LNB and indoor receiving equipment is attached to one of the struts. Unless the electronics can be fitted at the feed, the prime focus configuration is not attractive for transmit/receive applications.

In a variant of the prime focus approach, the feedhorn is at the end of a hook-shaped piece of waveguide that extends from the vertex of the parabolic reflector. The LNB is attached to the end of the waveguide behind the reflector, and the assembly is supported at the feedhorn with either struts or guy wires. That has the obvious advantage of allowing the electronics to be accessed more easily and provides placement out of the way of the main beam. The mechanical rigidity and thermal stability of the bent waveguide system are limited, but small-diameter antennas used for TVRO purposes can get by with that approach.

9.3.1.2 Cassegrain Optics

Many of the shortcomings of the prime focus fed antenna are overcome with the cassegrain system, shown in Figure 9.9(B). Using the folded optics principle of the reflector telescope, the focus of the parabola is directed back by a hyperbolic subreflector to a point at the vertex of the center of the main reflector. Proper illumination is still provided, but the feed and the electronics can be joined at an accessible point in the antenna structure. In large antennas, a hub enclosure is provided at the back for placement of essentially the entire RF terminal in close proximity to the feed. The cassegrain system eliminates the cable or waveguide runs between RF electronics and feed.

9.3.1.3 Splash Plate

The splash plate feed system, shown in Figure 9.9(c), relies on a subreflecting surface to direct energy back toward the reflector. In this case, the surface is a flat disk that is supported by a dielectric material like styrene plastic positioned in front of a feed tube. There is no need for a feed horn or feed support because the assembly is supported by the feed tube itself and the RF electronics can be located behind the reflector. This arrangement benefits from its mechanical simplicity and strength, as well as its avoidance of blockage from support struts.

9.3.1.4 Offset Fed Geometry

The offset fed geometry, which was reviewed in Chapter 7 for spacecraft antennas, also has been applied to Earth stations. Shown in Figure 9.10 is the common home Ku-band TVRO used in high power DBS. A main benefit of the offset approach is the elimination of main beam blockage by the feed and struts, which tends to improve antenna efficiency and reduce sidelobe levels. It also permits rain and snow to fall off the surface, because the reflector is at a more acute angle relative to the ground. In a particular design, there may be mechanical or cost advantages

Figure 9.10 Typical DBS home TVRO (*Source:* Courtesy of Sony.)

as well. While the figure shows a single feed and LNB, configurations with up to three feed-LNB combinations (to view three orbit positions) are available on the market.

9.3.2 Antenna Beam Pointing

The relatively narrow beamwidth provided by a fixed Earth station antenna results in some variation in signal strength as either the satellite or the antenna physically moves with respect to each other. As discussed in Chapter 4, the 3-dB beamwidth of an antenna defines the angle for which the gain will be one-half its peak value (or 3 dB down from maximum). For a GEO satellite, the antenna can be fixed to its mount as long as the stationkeeping error is well within the 3-dB beamwidth (more typically 0.5 dB). If that is not the case, some form of tracking will be needed.

Many antenna mounts are configured like an artillery piece, with adjustments being provided for azimuth (0 to 360 degrees) and elevation (0 to 90 degrees). That makes sense if the antenna must be capable of pointing anywhere in the sky, which is the case for antennas used to track a satellite in transfer orbit or for LEO operations. In moving the antenna between GEO satellites, the azimuth and elevation axis have restricted pointing ranges and must be adjusted in a coordinated way. For example, the mount may limit its travel to a 60° azimuth arc. Another common design for GEO satellites is the polar mount, named for the fact that the main axis of rotation is aligned with that of the Earth (e.g., pointed toward the poles). The reflector then can be moved from satellite to satellite by rotation along one axis.

9.3.3 High-Power Amplifiers

The performance of an Earth station uplink is gauged by the EIRP, the same parameter that applies to the satellite downlink. As discussed earlier, the EIRP is

the product of the HPA output power, the loss of the waveguide between HPA and antenna (expressed as a ratio less than 1), and the antenna gain. The result is expressed in decibels relative to 1W, so it also is convenient to add the component performances in decibel terms. For example, an EIRP of 80 dBW results from an HPA output of 30 dBW (i.e., 1,000W), a waveguide loss of 2 dB, and an antenna gain of 52 dB. Since the dimensions of an Earth station antenna are not subject to the physical constraints of the launch vehicle, the diameter and, consequently, the gain can be set at a more convenient point. That could be the optimum, which occurs where the cost of the antenna plus HPA is minimum. In VSATs and small UTs, HPA power must be held to an absolute minimum to save cost and minimize radiation, so the optimization usually is reversed to rely on the satellite G/T and EIRP to produce a satisfactory link. Antenna size or diameter may also be constrained by the location, such as the vehicle system in Figure 9.7.

Figure 9.11 presents a tradeoff curve for the design of an Earth station uplink for use in full transponder digital video service (in C-, Ku-, or Ka-band). To deliver an EIRP of 80 dBW, the curve gives the requisite antenna diameter and HPA power. An allowance of 2 dB for waveguide loss has been made in the curve. At Ku-band, relatively low HPA powers (less than 200W) are acceptable with antennas 6m in diameter or greater. To be able to employ a more compact reflector of 3m in diameter (typical for a truck-mounted transportable Earth station), the HPA must be capable of outputting 1,000W. There is an additional problem with the smaller diameter antennas, namely, the excessive power also will be radiated through the antenna sidelobes, producing unacceptable interference in the adjacent satellite. Also, these numbers do not allow for Ku- or Ka-band uplink power control (UPC) which requires up to 10 dB of HPA power margin.

The discussion of HPA technology in Chapter 4 provides another perspective with which to evaluate the tradeoff in Figure 9.11. Sufficient antenna size permits the use of lower powered, less expensive HPAs. In FDMA services, it is possible to use an SSPA to achieve the required Earth station EIRP for a single carrier provided sufficient antenna gain is available. Likewise, a klystron HPA could be avoided if the power for video service is less than 600W, which is possible in the
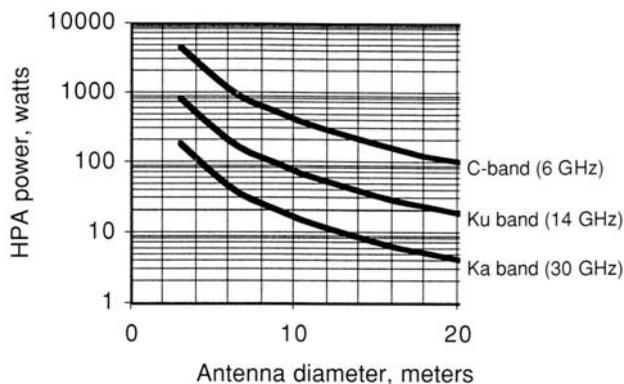
Figure 9.11   Earth station HPA power versus antenna diameter for full transponder uplinking (e.g., EIRP of 80 dBW).

example with a 6-m Ku-band antenna. There is an important consideration for multiple carrier service in that the HPA would have to be backed off to control RF IMD. That increases the HPA power rating at saturation by between 7 and 10 dB.

9.3.4 Upconverters and Downconverters

Downconverter and upconverter equipment, illustrated in Figure 9.12(a) and Figure 9.12(b), respectively, is needed to transfer the communication carriers between the operating RF frequencies and the IF used by the modems and baseband equipment. A typical upconverter has a first stage of amplification to provide adequate gain for the operation of the station equipment. The actual frequency conversion is accomplished in a mixer and LO combination such as that shown. A frequency-agile upconverter employs a frequency synthesizer to generate the LO so any carrier frequency within the satellite uplink band can be used. In a synthesizer, a standard frequency reference at 10 MHz is typical. Proper filtering is needed to prevent the LO and its harmonics from reaching the uplink path, a common consideration in the design of frequency translation stages.

The design of the downconverter is much the same as the comparable stage of the satellite receiver. Here, the first amplifier stage provides the needed overall gain and reduces the noise contribution of mixer and IF equipment. A synthesizer also has to be used to provide agility in receive frequency operation. In small UTs, the upconverter and the downconverter can be driven by the same synthesized LO to reduce cost and to access the same channel of operation over the satellite.
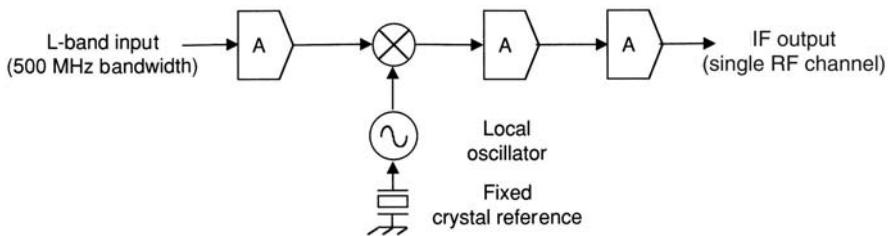
(A) Downconverter block diagram (L-band to IF)
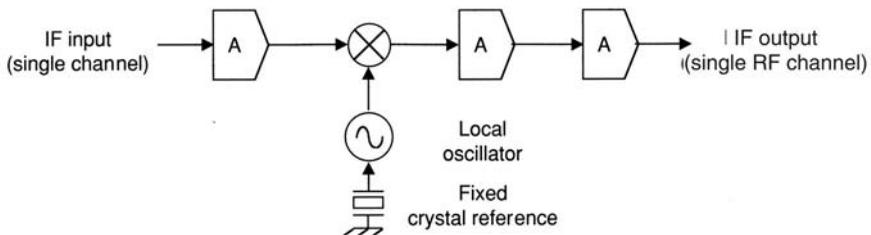
(B) Upconverter block diagram (IF to RF output)

Figure 9.12 (a, b) Typical upconverter and downconverter block diagrams used in fixed Earth stations.

9.3.5   RF Combining

In many large Earth stations, there may be a requirement to transmit two or more simultaneous carriers and to receive multiple carriers as well. The simple approach is to use passive microwave components such as the hybrid and filter multiplexers to combine and separate carriers, as appropriate. That principle was addressed in Chapter 6 in connection with the satellite repeater. An example of a passive power combiner is provided in Figure 9.13. On the uplink side at the RF frequencies, microwave filters must be used, one for each transmit frequency. Power combiners of this type look much like the output multiplexer in a bent-pipe satellite repeater, which is not surprising because they perform the same function. Some operators prefer to use 3-dB hybrids for greater flexibility because new microwave filters are unnecessary.

9.3.6   Uplink Power Control

Ku- and Ka-band transmit Earth stations and some VSATs require automatic uplink power control (UPC) to compensate for link fading. The extreme nature of the fading problem at Ka band makes UPC a standard feature. The principle behind UPC is that rain-induced fading is compensated in direct response to the attenuation. It is not a true closed-loop system because the uplink itself can be measured precisely only at the satellite, which is something generally not provided. Instead, the uplink Earth station measures downlink power and uses that as an indication of what is happening to the uplink. A workable approach in VSAT networks is for the hub to measure the return signal and command adjustments by the remote site.
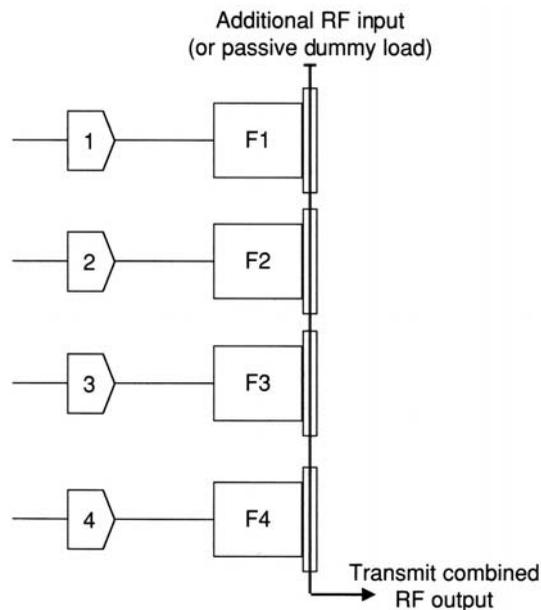
Figure 9.13   Example of a four-channel power combiner for an Earth station uplink, employing directional filters (alternatively, circulator coupling can be employed).

9.4   Intermediate Frequency and Baseband Equipment

The IF and baseband equipment, illustrated in Figure 9.1, establish the basic service capabilities of the Earth station, including modulation and FEC, multiple access method, and interface to the user or terrestrial network. In many cases, this portion represents the largest investment on the ground, particularly for a major gateway or network control facility. That is because, while the RF terminal usually is fixed for a particular application, the quantity and the design of the baseband equipment must be adjusted to support the bandwidth and communication applications the stations serve. This section reviews typical configurations of the major elements of this section of the Earth station.

9.4.1   Modulators, Demodulators, and Modems

The most essential portion of the baseband-to-IF chain consists of the modulator and demodulator or a combination of the two, called a modem. On the receive side, the demodulator detects the incoming carrier, synchronizes the data, performs error correction and outputs a clean bitstream for the particular application. The transmit side works in the opposite direction. Threshold performance in terms of error rate and synchronization is determined at this level of the station. Because there must be one modem for each frequency, multicarrier FDMA stations can contain large quantities of these units, representing a sizable investment. In TDMA star networks, some frequency channels are shared by several Earth stations. There would need to be only a single modem per Earth station, since it can receive the bursts from the other Earth stations so long as they do not overlap in time. That is, of course, central to the operation of a TDMA network (as opposed to a CDMA network, where stations may transmit simultaneously on the same frequency).

Figure 9.14 is a block diagram of a typical BPSK digital modem, such as that used in a TDMA VSAT. Modulation is provided by mixing the data with a 70-MHz carrier, inverting the sinewave when a change from 1 to 0 (or vice versa) is desired. We obtain QPSK by adding a 90-degree phase shift to a second modulated stream of data, thus doubling the use of the uplink frequency channel. Reception works in the reverse direction, using the same mixing process but resulting in the original data being recovered. Additional circuitry is needed on the receive side to remove satellite link noise and resolve the bit pattern. The latter is accomplished by a decision circuit that, simply enough, decides if the received bit is a 1 or a 0. The presence of noise and distortion in the demodulated bits causes the decision circuit to make a bad decision occasionally, a process that gives bit errors. It is the average rate of those errors, the BER, that determines received data quality.

The typical satellite modem (and terrestrial modem, as well) also contains the FEC coding and decoding circuitry that reduces the BER by factors of 100 to 10,000, depending on the strength of the coding technique. Other functions include burst synchronization for TDMA, baseband processing such as compression/decompression, and encryption/decryption, if appropriate. More options for the modem, including bandwidth-efficient modulation (BEM), can be found in Chapter 5.

(A) Biphase PSK (BPSK) modulator

(B) Biphase PSK (BPSK) demodulator
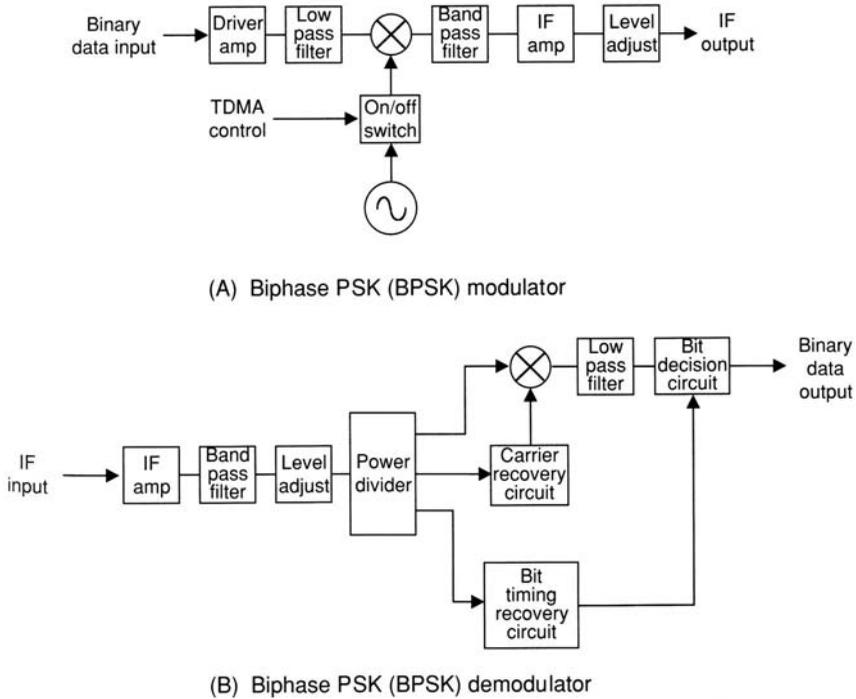
Figure 9.14   (a, b) Functional elements of the typical PSK modem.

9.4.2   Multiplexing and Packet Processing

It is common practice to use the outbound and inbound data bitstream for multiple functions, such as multiple information channel transmission, network control, and facility monitoring. The convenient way to do this is with TDM or packet data transfer. Both rely on inserting different blocks of data into the common stream to and from the satellite. In TDM, the individual streams or information channels are inserted according to prearranged timing, usually periodic and of fixed length. That works for a standard format without compression. On the other hand, many applications use dynamic bandwidth allocation, so fixed timing cannot be assumed. In that case, it is better to use some form of packet transmission, which is a standard feature in most telecommunications networks, satellite and terrestrial. The equipment usually is designed specifically for satellite network use, but in some cases, conventional multiplexers and packet switches (e.g., IP routers) are employed.

9.5   Tail Links and Terrestrial Interface

The terrestrial interface can be called the ''business end'' of the Earth station because it is where the actual user services are made available. In larger stations, the terrestrial interface can be quite complex, as suggested in Figure 9.1. A gateway station looks to the outside world like a traffic node or a portion of a broadcasting

studio, depending on the application. On the other hand, simple terminals such as a TVRO or a VSAT render their services directly to consumer equipment, like a TV set or a PC, respectively. A handheld UT may have no external interface since it is a self-contained unit. The requirements of the terrestrial interface, therefore, vary widely based on the application and user facility provided.

9.5.1 Terrestrial Tail Options

A terrestrial tail may be needed to connect the communications Earth station to one or more remote user locations. The distance to be covered can range from a few meters to hundreds of kilometers. In C-band satellite systems, the Earth station is often isolated from a city to reduce terrestrial interference, in which case an elaborate tail is required. On the other hand, terrestrial interference is not present in most Ku- and Ka-band systems; therefore, tails can be relatively short. An exception is the case where a large Earth station (i.e., a teleport) is shared by several users each of which must be reached by local terrestrial transmission links, often provided by local telecommunications carriers. Site diversity may be needed at Ka-band and higher to provide reception when one location is severely affected by rain. The tail link between the diverse sites can represent a considerable investment in fiber optics or microwave towers due to the requirement to transfer a large bandwidth.

   A few examples of the tail configuration for a major Earth station are illustrated in Figure 9.15. A cable interfacility link (IFL) connects each of two RF terminals (lower left of the figure) to the main Earth station building. A good option for this IFL is fiber optic cable terminated in optical analog transceivers, as discussed later. Within the building can be found the baseband and interface equipment appropriate for the types of services being provided. It is assumed that all the traffic is to be
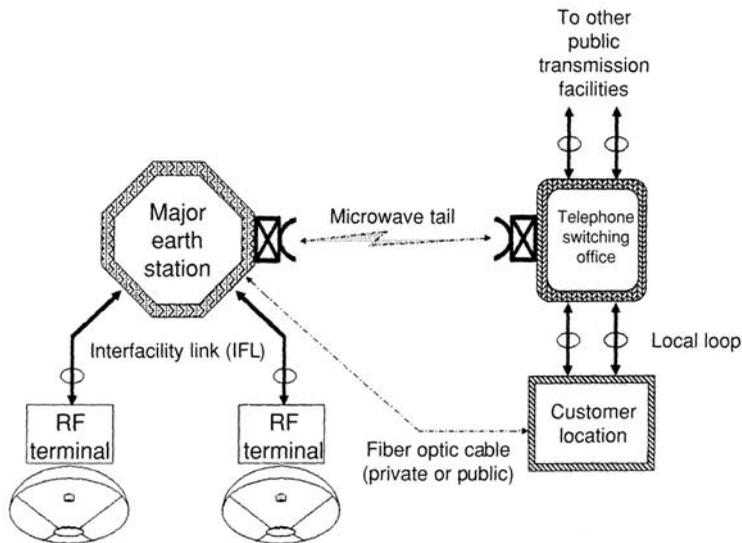
Figure 9.15   Use of terrestrial tails between Earth station facilities, switching offices, and customer locations (distances not to scale).

transported to the nearest city. These transmission requirements are met in the example with a single-hop terrestrial microwave link, equipped with sufficient receiver-transmitter units to carry the video voice, and data traffic. The customer location or locations access the first telecommunications office through either public or private fiber optic cable.

The telecommunications office and local loops can be bypassed (the dotted line between the customer location and the major Earth station) with a private tail fitting the requirement. The reason for bypassing can be based on economics, it possibly being less expensive to own the tail facility rather than leasing capacity from the local telecommunications utility. The private bypass, which permits the user to maintain control of the resources, eliminates some of the time necessary to implement new public facilities. Reliability can be greatly improved with the path diversity of using both the public access lines and the private bypass link.

The two principal types of terrestrial tails for bulk transmission purposes are line-of-microwave and fiber optic cable. Both are effective and reliable and can be economical when applied properly. Short distances between buildings, such as for the IFL within an Earth station site or when a VSAT is connected to one or more users, are best traversed with fiber optic cable. It has become a practice in large Earth station design to rely on fiber optic cabling because of its low-noise performance and its ability to reduce radiated electromagnetic interference and conducted electrical surges. Another option for short tails is single-hop microwave communication. Available in capacities of a few megabits per second to over 1 Gbps, single-hop microwave systems are relatively inexpensive to install and operate. They, too, have the benefit of not transferring electrical surges and high levels of interference.

Focusing on large-capacity terrestrial tails, Figure 9.16 compares fiber optic with microwave on the basis of relative investment cost. A fiber optic link is relatively costly per kilometer of construction, particularly in a metropolitan area. The black line of constant slope is the cost of the cable, electronics, and installation for a basic point-to-point link, assuming that the right-of-way is already available at essentially no cost. Most of the expense is for the repeater and terminal electronics
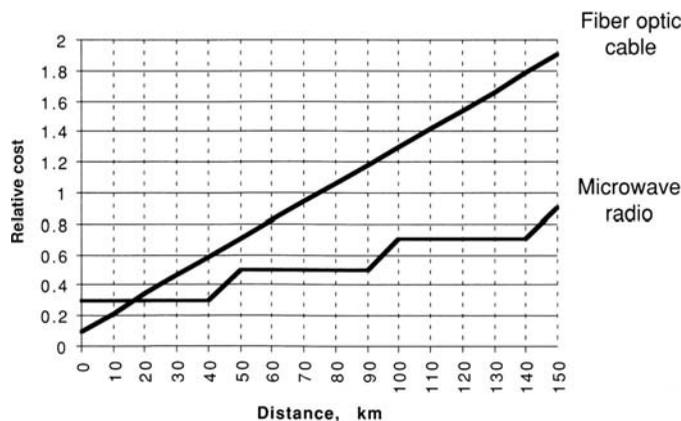
Figure 9.16 Comparison of fiber optic and terrestrial microwave links on the basis of relative investment cost for a terrestrial tail (excluding land and right-of-way).

and for installation, which is labor intensive. Fiber optic cable itself is a small part of the cost, so it usually is wise to facilitate future expansion and repairs by including more fiber pairs than are necessary for the current demand.

The cost of putting in a high-capacity terrestrial microwave system, including receiver-transmitter units, towers, and antennas, is shown with the stair step, where the step occurs at the assumed repeater spacing of 50 km. The cost of preparing and obtaining repeater sites is not included. In comparing the two technologies, it is clear that fiber optic is attractive for relatively short tail lengths, that is, less than the extent of one microwave hop. On the other hand, the capacity of the fiber optic link can be made many times greater. That is because the cable contains extra fiber pairs that are available for expansion (and that were included in the original cable at little extra cost). In addition, transmission rates on a given fiber pair can be increased as newer optical modems and electronics are introduced.

It should be kept in mind that any real-world situation should be examined in detail before the tail technology is selected. That is because installation and right-of-way costs can vary widely, as can the cost of the equipment itself. Figure 9.16 illustrates a valid framework for making such a comparison. New approaches, both through purchase or lease, constantly appear on the market, so the tradeoff will change over time. Consequently, an investigation of this type should be conducted prior to commitment to a particular approach.

9.5.2 Terrestrial Network Interfaces

Earth stations are a means of providing access to a satellite communication network of some type. An important part of their design deals with the interface between the station and terrestrial network infrastructure, illustrated as the last box at the lower left of Figure 9.1. It was emphasized at the beginning of this chapter that an improperly designed or installed interface will seriously degrade the quality of service. Often, access to the satellite network will be blocked entirely, particularly in the case of digital traffic and control information. Detailed information can be obtained from technical specifications of equipment and international standards such as those promulgated by the technical committees of the American National Standards Institute (ANSI), the Institute of Electrical and Electronics Engineers (IEEE), and the ITU.

Interfacing with the public telephone network is often on a digital basis using either the Pleisiochronous Digital Hierarchy (PDH) or the Synchronous Digital Hierarchy (SDH). Having been in use for more than five decades, PDH is offered in North America by the T1 at 1.544 Mbps and DS3 at 45 Mbps, and in Europe and much of the rest of the world by the E1 at 2.048 Mbps and E3 at 38 Mbps. It was designed for digital data transfer at a constant bit rate but in a manner that allows for some inaccuracy in the clock timing on the ends of the link. Direct connection between the T1 and the E1 are accomplished by a device called a *transmux,* which aggregates sufficient bandwidth on each side to balance the link. These access circuit types are still quite prevalent because of the large quantity of existing infrastructure on the public network side and telecommunications equipment on the user side. For example, private branch exchanges (PBXs) and Internet

routers all offer T1, DS3, E1, and E3 interfaces due to availability of these service offerings. As a result, many Earth stations support PDH services as well.

The SDH is associated with fiber optic transmission and was developed to exploit the reliability, bandwidth, and accuracy of modern telecommunications systems. The interface to SDH begins at 155 Mbps, which is referred to as SDH-1. From there, it builds rapidly to gigabit per second speeds that correspond to the capacity of glass fiber. In the case of a satellite path, SDH-1 is the most common level as it can be accommodated by one or a group of standard transponders. Note that in North America, SDH is referred to as the Synchronous Optical Network (SONET) at starts at one third of the SDH-1 rate (i.e., 53.5 Mbps).

Another interface standard found in high-speed networks is the Asynchronous Transfer Mode (ATM), a standard originally intended to form the broadband public network of the future. What has happened instead is that ATM has established itself within certain niches where high speed and flexible adaptation to applications are important. The technology is similar to the Internet Protocol in concept be very different within its construction and features. For example, ATM is based on a fixed length packet of 53 bytes, called a cell. The cells are not addressed in the same manner as access points in the Internet, but rely on the concept of the virtual circuit that is set up and taken down as needed to satisfy a particular demand. The Earth station interface to an ATM network could be with SDH or PDH access lines.

Our final example of a data interface is the IEEE 802.3 standard, commonly referred to as Ethernet. Within this overall family we find the familiar standards: 10BaseT (supporting 10 Mbps) and 100Base T (supporting 100 Mbps), both used on unshielded twisted pair (UTP) with ubiquitous RJ-45 connector. These employ eight individual wires and are known as Category 5 UTP (CAT 5) and Category 6 UTP (CAT 6), the difference relating to the quality of wire and resulting ability to carry up to 1 Gbps in the case of CAT 6. This approach has rationalized and simplified the interface from VSAT and data UT to the end user. It can also be employed at a hub Earth station used to connect the satellite network into a private Intranet or the public Internet.

Within the domain of digital video, the common standards are referred to as Asynchronous Serial Interface (ASI) and Serial Data Interface (SDI). The two standards are similar in general concept but cannot be connected directly to each other. The first generation of MPEG-2 digital video equipment employed SDI but later generations adopted ASI as it results in some cost savings and simplifications in interconnection. Both are used for equipment in close proximity and must be adapted to ATM or SDH if distances measured in kilometers are involved.

9.6   Earth Station Facility Design

The last practical consideration to be covered is the physical installation of the Earth station, that is, the design of the building and the arrangement of equipment therein. It is instructive to take as an example the layout of the site and building for a major Earth station, shown in Figure 9.17. Many of the same principles can be applied to smaller stations, including VSATs. The drawing assumes operation
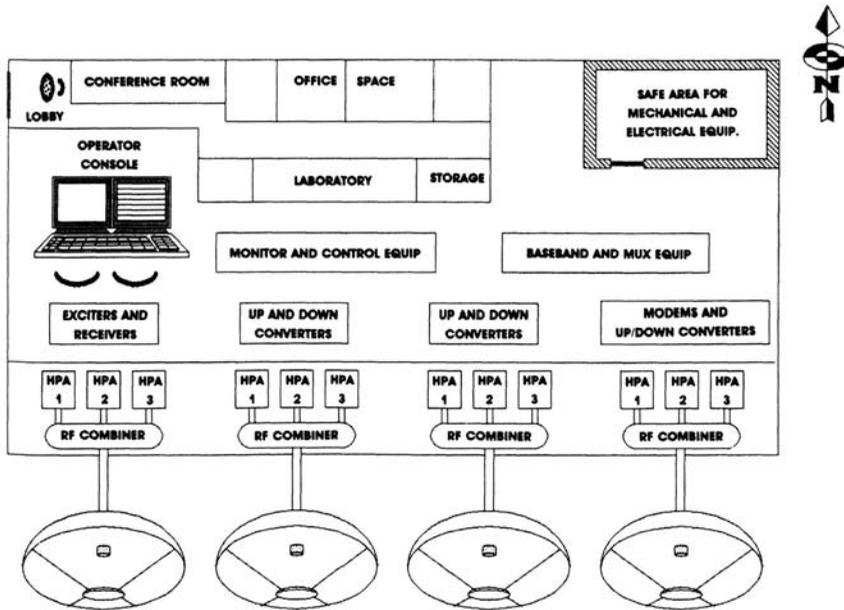
Figure 9.17 Site plan and equipment arrangement for a typical major Earth station.

with GEO satellites with the station located north of the equator. Consequently, the antennas are aligned along the southern perimeter of the site to afford good visibility of the orbit arc. This arrangement of the antennas is usually the most flexible if there is sufficient space between the reflectors to allow independent pointing of each main beam to any orbit position. Care in placing the foundations must be taken so that antennas never touch each other during repointing and that the beams do not overlap one another to a significant degree. The latter minimizes the interaction of the electromagnetic fields around the antennas, including the effects of diffraction. With proper isolation between antennas, the beams will be properly formed, giving the desired cross-polarization isolation and contol of sidelobes.

The next important aspect is the location and arrangement of the RF equipment. The LNAs or LNBs should be mounted directly on the receive port of the feedhorn to minimize noise and loss. That also can be done for the HPAs as well if the power level is in the range of 5W to 150W, because compact SSPAs and TWTAs can be mounted near or on the reflector. Power levels in the hundreds to thousands of watts require high-power TWT and klystron amplifiers, respectively, which must be protected against the elements. Operation in two or more transponders at the same time is accomplished by connecting the HPAs in the building to an RF combiner of the type described in a previous section. The HPAs and RF combiner should be located as close as possible to the antenna to minimize waveguide losses. That may require that this portion of the RF terminal be contained in an equipment shelter away from the building but close to the antenna.

The baseband and control equipment can be located anywhere in the building since baseband and IF cable lengths are not critical to the performance of the Earth station. An IFL that extends between buildings should be constructed with fiber-

optic cable using analog optical transceivers. The main considerations here deal with allowing station personnel to operate and maintain the equipment. The appearance of the equipment in the room also is important both to the people who work there and to guests who may come to visit. Equipment that could be hazardous to humans, such as batteries, high-voltage electrical equipment, and diesel generators, should be isolated from the general flow of people, as shown in the upper right of the drawing. A major Earth station also deserves to have a front lobby to control access to the facility and office space for administration and management personnel. Test, calibration, and repair of electronic equipment can be done in a small laboratory.

The importance of providing adequate space between adjacent antennas to minimize RF interaction has been stressed. Figure 9.18 illustrates another important aspect in antenna placement dealing with the protection of personnel on site and the local community. The far-field pattern of the ground antenna contains the familiar main beam and sidelobes. For a high power uplink, the near-field region can produce elevated levels of RF radiation that are potentially harmful to humans and other living beings that enter the near field. The normal practice is to fence in the site to prevent people and animals from ever reaching such close proximity; in crowded metropolitan areas, however, that may be impractical.

A shield wall of metal or concrete such as that shown in Figure 9.18 can greatly reduce the radiation exposure in the surrounding area. The top of the wall must reach above the highest point on the antenna for any angle of elevation it may take. Usually, the worst case occurs when the antenna is pointed at the horizon, which also maximizes the absolute level of radiation along the particular azimuth. A high wall minimizes the diffraction over the top, which is the same consideration in using natural or man-made shielding to prevent RFI. Analytical techniques using wave and diffraction theory can predict radiation levels in the vicinity of the antenna and the wall, provided that the detailed structure of the antenna and the wall are included in the analysis.

9.7   Major Classes of Earth Stations

A unique class of Earth station is employed for each type of communication service being rendered. In addition, the size and the complexity of a station depend on
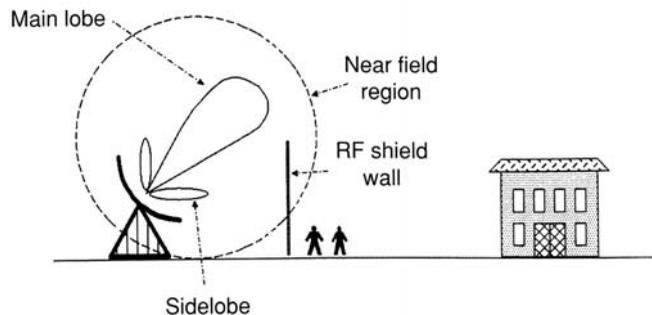
Figure 9.18   Protection of personnel and buildings from local RF fields around the transmitting Earth station antenna.

the quantity of traffic and, if appropriate, the variety of services offered. This section briefly describes some of the more common station configurations in use for broadband communications. To start off, the TT&C Earth station and its associated satellite control center are presented to give the reader a feel for the scope of the ground facility used to operate the space segment. The rest of the section covers stations ranging in size from the largest class (used for television uplinking, bulk data transmission, and MSS gateway service) to the smallest (VSATs, TVROs, and tabletop and handheld MSS UTs).

9.7.1 TT&C Ground Facilities

The TT&C ground facilities of a satellite operator cover a variety of functions and capabilities, as depicted in Figure 9.19. In some systems, all the functions are centralized at one location to minimize capital and operating expenses. That approach occupies the least land and building space and can economize on equipment. Additional savings accrue from minimizing expenses for running the buildings and by being efficient in the use of personnel. For example, a single maintenance team can service all the electronic equipment. A second approach splits the facility between a TT&C Earth station and an SCC, as indicated by the vertical broken line in Figure 9.19. Using a tail link such as microwave or leased private lines, the SCC is established at a distant point, such as the headquarters of the satellite operator. That is advantageous where much of the business is carried out in a central city and the TT&C is located remotely to avoid RFI and environmental hazards. It also is the case in non-GEO systems because a given TT&C station can view only a few satellites at any time. This section reviews the design and capability of each half in the TT&C facility. More detail on the operation of these facilities is provided in Chapter 11.

9.7.1.1 TT&C Earth Station

The TT&C station has essentially the same capability as the generic Earth station given in Figure 9.1, incorporating the RF terminal, baseband equipment, and terrestrial interface (appropriate to a separated SCC). The antennas employed for GEO satellites are usually 10–13m in diameter (whether C-, Ku-, or Ka-band) to provide the maximum possible link margin because transfer orbit and potential abnormal on-orbit conditions need to be accommodated. In a typical TT&C station for GEO operations, there is one limited motion antenna pointed at each operational satellite (two are assumed in the illustration), while the full tracking antenna is used for transfer orbit operations and for testing purposes during the life of the satellites. A TT&C station used for a non-GEO constellation must, of necessity, rely on tracking antennas, which also function for communications. An example of a TT&C station with full and limited motion antennas is shown in Figure 9.20. All antennas aid in tracking the satellites' positions by providing azimuth and elevation readouts to the data processing system from their respective mounts. The full tracking and limited motion RF terminals interface with the rest of the station at intermediate frequency.
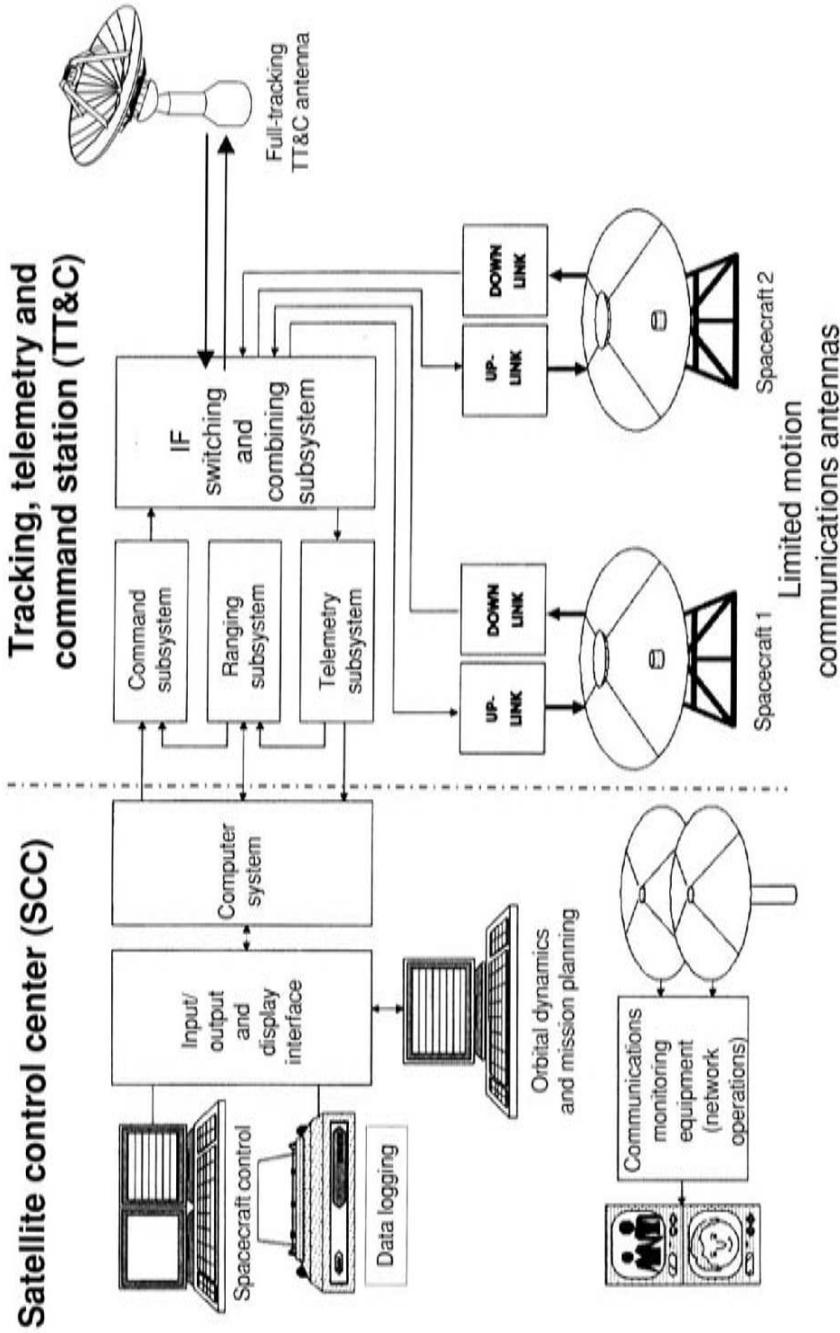
Figure 9.19   General arrangement of TT&C ground facilities for a space segment consisting of two geostationary satellites.
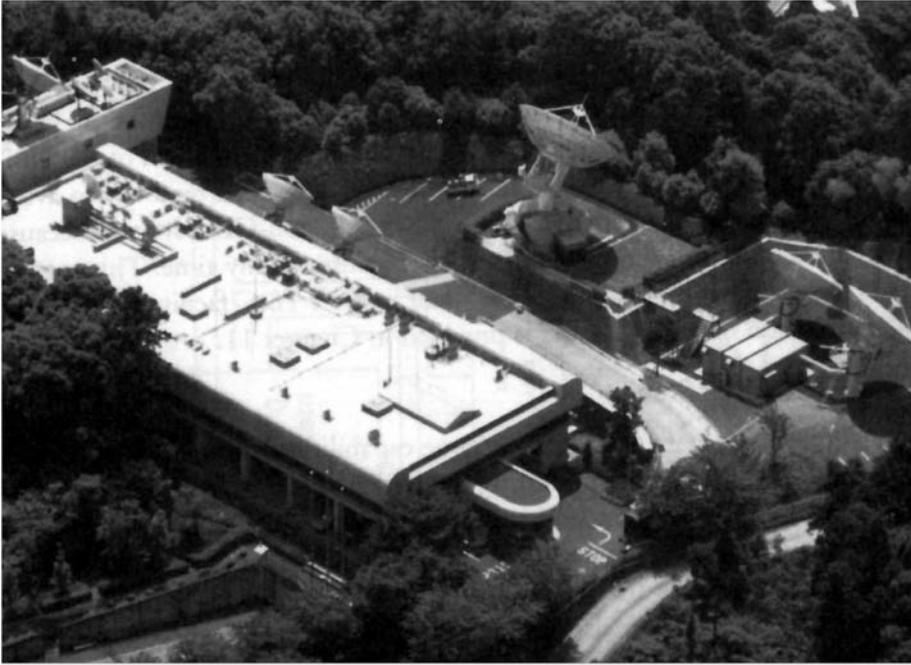
Figure 9.20   A typical TT&C Earth station with a full-motion antenna for tracking and testing as well as several limited motion antennas used for full-time TT&C links and communications services. (*Source:* Sky Perfect JSAT, Ltd.)

The baseband equipment is shown at the upper center of Figure 9.19. In the command subsystem, the actual digital commands are taken from the incoming data stream and modulated onto the IF carrier using either FSK or PSK, depending on the design of the spacecraft command receiver. Other special features can be incorporated in the command subsystem, such as providing the RF pilot carrier modulation used by a spacecraft antenna tracking system. For commanding purposes, the common carrier need be on the air only when it is sending commands and for ranging. That allows the uplink chain to be used on another satellite or for other purposes when commanding is not being performed.

The telemetry subsystem receives the telemetry carrier at IF from the downlink and demodulates the actual stream of telemetry data. The process is straightforward and directly analogous to the reception of data communication in a standard Earth station. The third function of the baseband equipment is for ranging, which is the process outlined in Chapter 8 for measuring the position and velocity of the satellite. The ranging subsystem generates a baseband ranging signal, which the command subsystem modulates on the uplink carrier. After passing through the TT&C subsystem of the satellite, the ranging signal is received at the TT&C station, demodulated, and the baseband compared with what was transmitted. The measure time delay and frequency provide the basis for computing the range and the range rate. In an alternative ranging approach, the ranging baseband signals are modulated on an independent carrier and uplinked to a frequency within one of the operating transponders of the communication payload. Referred to as transponder ranging, the technique eliminates the need to switch the command link between the commanding

and ranging functions but consumes some of the communications bandwidth of the satellite repeater. In yet another approach, the ranging signal rides along with a communication carrier, thus consuming less transponder resource.

The interconnection and switching of command, telemetry, and ranging links between each other and among different satellites is facilitated by the IF switching and control subsystem. For example, during transfer orbit operations, the TT&C link to the satellite is established through the full tracking antenna. A switchover to the assigned limited-motion antenna normally is accomplished with the IF switching and control subsystem after the satellite is located in its assigned orbit slot and has been fully deployed and tested. The TT&C station must be capable, however, of switching the full-motion antenna back in line for testing purposes or in the event of an emergency situation, which could happen at any time in the operating lifetime of the satellite.

In a global TT&C network for non-GEO operation, several TT&C stations transfer their data to a central location, such as the SCC. Coordination and switching of signals from literally dozens of satellites at a cental facility can be a complex task, requiring special computer equipment and software.

9.7.1.2 Satellite Control Center

The SCC is the brain of the satellite operation, providing the computing power and human intelligence necessary to operate and control a system of several satellites and TT&C Earth stations. Depicted in the left half of Figure 9.19 is the configuration of a typical SCC. The baseband signals between the TT&C and the SCC can interface directly with the computer system. Additional pieces of specialized digital processing equipment may be placed between computer and baseband in the integrated system. An example of such a device is the command generator used to format and modulate command transmissions for the uplink. A complementary device called the telemetry decommutator accepts the telemetry TDM data stream and demultiplexes the telemetry channels. Another type of special device is the spacecraft simulator, usually a computer workstation programmed to behave like an operating satellite (from a bus standpoint). The simulator is used to prepare for critical maneuvers and for training operators to deal with normal and contingency situations, such as a satellite temporarily losing attitude pointing control.

The single box labeled "computer system" in Figure 9.19 contains sufficient computing power, redundancy and storage to support all the requirements of the SCC, including reliability. The computer system performs real-time functions of command generation, telemetry reception and processing, and ranging. Additionally, several SCC activities are handled by the computer system in the background. Orbital dynamics deals with the determination of the satellite orbit and the planning of orbit correction maneuvers necessary to maintain satellite position. Using ranging data as an input to the orbital software, the orbital dynamics personnel generate maneuver plans, which are presented to the spacecraft controllers (i.e., the people who "fly" the satellites) for action by the commanad system. The orbital dynamics personnel and spacecraft controllers access the computer system with the workstations shown at the upper left of Figure 9.19. In the photograph in Figure 9.21, spacecraft controllers can enter spacecraft commands through workstations in front

Figure 9.21   The satellite control console where the JC Sat satellites are controlled and monitored through the video display terminals. The Yokohoma TT&C station is used as the primary operation center for the JC Sat systems. (*Source:* Photograph courtesy of Sky Perfect JSAT, Ltd.)

of them. The computer system keeps an ever watchful eye over the telemetry data and compares the reading to previously stored threshold alarm values. The data logging device in Figure 9.19 stores every action taken by the controllers and records unusual spacecraft data identified by the computer.

The last function of the TT&C ground facility is the monitoring of the communication transmissions to and from the satellite. This particular example is for a bent-pipe type of satellite, where all uplink transmission to the satellite can be monitored in the downlink. Prior to the start of service, the communication payload is thoroughly tested and every major component checked out from the TT&C Earth station, using automatic test equipment associated with the full tracking antenna. With commercial communications services being provided over the satellite, it is important that the satellite operator have a system to continuously monitor the entire downlink spectrum of each satellite. That is needed to ensure that all RF carriers are set to their prescribed power levels and are located at their assigned frequencies. If the satellite employs dual polarization to achieve frequency reuse, the purity of polarization of every carrier also must be checked prior to start of service and monitored during operation. Most of the transmissions to the satellite do not emanate from the TT&C station; therefore, the communications controllers must have telephone access to the uplinking Earth stations no matter where they may be located.

If the communications control console is located at the TT&C site, the limited-motion antennas can receive the downlink spectrum and deliver it to the SCC for inspection with spectrum analyzers and video monitors. Figure 9.19 indicates the other approach, in which the communications control is colocated with the SCC some distance from the TT&C Earth station. In such cases, one or more separate RO antennas can be provided but only if the satellite footprint covers the location of the SCC. Satellites that provide frequency reuse using narrow spot beams, such as applied at Ku- and Ka-band, do pose a bit of a problem in downlink monitoring since a separate Earth station is required in each independent footprint. Some

approaches to this problem are discussed in Chapter 11. The communications console used to monitor transmissions from the JC Sat satellites is shown in the photograph in Figure 9.22. Note the use of video monitors to augment RF measurements and the availability of telephone communications for contacting uplinking Earth stations and other satellite operators.

Monitoring of communication transmissions for digital processing satellites and non-GEO constellations used for MSS services is considerably more complex than the preceding discussion would indicate. That is because the downlink from the satellite comes in many separate beams and would involve reformatting of information onboard the satellite. The inclusion of intersatellite links adds another level of difficulty since the signals cannot be observed on the ground. Monitoring of services, therefore, depends on special facilities onboard each satellite and the relay of data either through the telemetry system or by a special downlink for that purpose. On the ground, the data form part of the overall network management system used by the satellite operation; this function could be handed off to a network service provider.

9.7.2 TV Uplinks and Broadcast Centers

Earth stations used for commercial video transmission services generally are large in class because of the high bandwidths, playback abilities, and extensive signal routing and switching capabilities. Video/program distribution refers to the basic function of the uplink in using the point-to-multipoint feature of a GEO satellite. A video uplink provides the origination path for TV as well as the capability to receive signals through satellite and terrestrial means, and retransmit them to the satellite. In addition, a broadcast center adds a whole range of other technical and production activities and so mirrors aspects of a TV studio. A typical broadcast center is the core of a DTH network, including the ability to retransmit and originate as many as 300 independent video channels. Any of the above classes of video

Figure 9.22   The downlinks of the JC Sat satellites are monitored on a 24-hour basis by communications controllers at the JC Sat network operations center. (*Source:* Photograph courtesy of Sky Perfect JSAT, Ltd.)

Earth stations has the full range of facilities discussed at the beginning of this chapter and shown in Figure 9.1, including a full complement of redundant equipment.

The configuration of the electronic systems of a basic video uplink station with the capability to transmit four simultaneous TV carriers is shown in Figure 9.23. There are separate interfaces for the video (color picture) and stereo and audio services. Video and audioswitching and distribution are provided prior to the uplink equipment to allow operators to redirect specfic programming to the various uplink channels without changing the physical connections. That is often under computer control according to a prearranged schedule. Four of six video exciters provide the RF input to klystron HPAs. On the receive side, the antenna delivers the downlink band to video receivers that are used to monitor the uplink originated channels and to acquire programming from other sources (sources that potentially include other satellites, if additional antennas are available). Video servers can be included to play back received TV programs as well as from physical drives. The design of the overall broadcast center is as much science as art, relying heavily on the designer's experience and understanding of the transmission and business needs. Conversion from analog to digital could be performed remotely at the studio and the interface be entirely digital in nature. Modulation for the uplink would be with QPSK or a more bandwidth-efficient modulation such as 8PSK or 16APSK.

9.7.2.1   Video Equipment Chain

The encoded video signal is accepted from the terrestrial interface by the video and audio monitoring and switching unit, shown at the upper left of Figure 9.23.
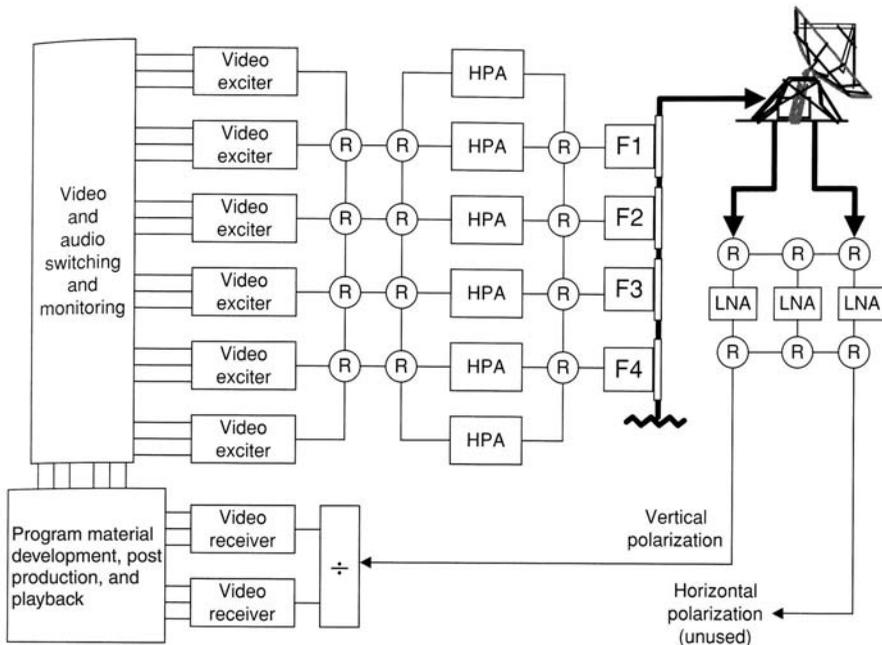
Figure 9.23   Example of a major communications Earth station with the capability to provide full duplex analog or digital video transmission.

Along with the video comes one or more channels of high-quality audio (monaural or stereo, possibly with additional channels for other languages) as well as data channels that support the management of the TV network. The monitoring and switching equipment is used in both the transmit and receive directions to route the baseband signals between uplinks, downlinks, redundant equipment, and TV test equipment. Service quality is extremely important in television broadcasting, so Earth station personnel must have the facilities and training to perform tests and to help in troubleshooting technical problems.

Digital video compression and transmission systems are part of most video uplink stations and broadcast centers. The basic architecture includes the necessary digital compression and TDM equipment. The latter is used to multiplex video and audio information in digital form and to multiplex channels together for multiple-channel per carrier (MCPC) operation. The quantity of simultaneous digital video channels can be nearly doubled on each carrier using statistical multiplexing. Another important feature of the broadcast center is conditional access to control delivery of the video programming. That makes use of encryption technology to prevent pirating of the programming or unauthorized viewing.

For video service, the HPA has an output power in the range of 400W (TWTA) to 1,500W (KPA), depending on the size and the gain of the ground antenna and the sensitivity of the satellite. The configuration in Figure 9.23 shows additional HPAs that operate in other transponders; hence, the HPA outputs are combined in an RF combiner. These very high power levels demand the lowest possible RF loss and adequate cooling of the waveguide filters of the combiner. Redundancy switching is accomplished with waveguide rotary "baseball" switches (discussed in Chapter 6) to cross-connect online and standby HPAs The resistor symbols indicate high-power dummy loads that absorb the output of an HPA when it is not switched into the operating chain and antenna.

The downlink equipment mirrors the uplink with the only difference being in the configuration of the LNAs. A separate operating LNA is provided for each polarization, although the station in Figure 9.23 is not using both simultaneously. When receiving a backhaul video point-to-point feed, either polarization may come into use. The LNA connected to the vertical antenna port provides sufficient gain to overcome the loss of a passive power splitter that feeds the entire frequency range to the video receivers. Each receiver can be tuned to any transponder in the downlink frequency range, including that of the uplink for monitoring purposes.

9.7.2.2   Video Distribution, Routing and Switching

The quantity and the arrangement of the video distribution, routing and switching equipment depend heavily on the particular application of the station or broadcast center. The basic devices are sold to cross-connect a particular quantity of video and audio signals. For example, an eight-by-eight router can connect each of eight input video/audio ports to any one of the eight output ports. It is akin to a telephone cross-bar switch, with the added complexity that multiple signals are switched simultaneously and the device is computer controlled to support a dynamic programming schedule common to the TV industry.

In a broadcast center, the routing and switching function is central to the operation of the distribution network; hence, its design and installation are critical to success. The function requires a large capacity in terms of the number of simultaneous video/audio channels (numbering in the hundreds). Both computer and human control are involved: the computers are programmed ahead of time with the requirements of the network, and humans are always present to observe that the system is working correctly. Numerous backup modes are provided, including dual feeding of the same (important) video channels, and human intervention may be needed at any time to ensure that the right programs are being transferred to the right channels. More recently, computer automation has been added to reduce the amount of manual labor involved and to improve the overall reliability of switching operation. Mistakes at this level can be detrimental to the quality of service and, ultimately, the revenue derived from programming and advertising.

9.7.2.3 Program Acquisition and Playback

Whether we are talking about a basic video uplink or a comprehensive broadcast center, there is always a need to obtain the programming input from multiple sources. Program acquisition in this context means that the video signal is originating outside the station from one or more remote sources. An uplink that serves a TV network acquires its programming from the originating studio through a terrestrial link. As discussed earlier in this chapter, tail links use either line-of-sight microwave communication or fiber optic cable. Fiber optic cable transmission usually is preferred because of the wider bandwidths for either analog or digital video formats and because of the absence of signal fading (except, of course, for a backhoe fade). If we are talking about receiving from other satellites, there must be sufficient TVRO antennas to provide the desired access. For that reason, a large uplink station, such as that shown in Figure 9.24, includes a considerable quantity of transmit and receive antennas.

The other potential source of programming is from video files, which are either developed at the site or delivered to it (a process called ''bicycling''). Files on digital tape, hard drives and DVDs are received, cataloged, and stored on servers for subsequent playback. The advantage of video server technology is that it precludes mistakes in setting up the transmission and allows content to be selected for particular transmissions (e.g., the insertion of promotions and advertising targeted to a particular audience).

9.7.3 FDMA Digital Communications Service

The traditional type of Earth station used for wideband data transmission and backhaul links in cellular telephone often employs the FDMA approach. There are two ways to implement FDMA: single-destination carriers and multiple-destination carriers. In the single-destination approach, there are two carriers for each bidirectional (duplex) link, thus connecting a pair of Earth stations. We would expect each carrier to contain multiple voice and data channels using TDM or possibly statistical multiplexing. Time assignments on the respective carriers are also paired

Figure 9.24   The DIRECTV broadcasting center in Castle Rock, Colorado. (*Source:* Courtesy of DIRECTV, Inc.)

so that each voice or data channel is duplex. The other technologies identified in Section 9.5.2 may be applied as well.

The multiple-destination type of carrier contains traffic not for one station but rather for the group of stations that are to be networked. Figure 3.4 presents a four-station network where each station transmits one carrier that contains multiplexed traffic for all other stations. Each Earth station must, in turn, have one receive chain for each transmitting station providing traffic to it, for a maximum of three receive chains. In total, there is a requirement that four carriers access the satellite. In comparison, using single-destination carriers and four stations, there would have to be a total of 12 carriers to provide full interconnectivity between all stations.

A digital communication subsystem capable of transmitting four T1 (or E1) carriers is shown at the left of Figure 9.25. Each T1 has its own modem, tuned to a separate frequency in the transponder. Redundancy is provided for the modems on a five-for-four basis using an integrated switching system. The T1 channels interface with the terrestrial network through a TDM multiplexer that combines the four streams into one 6-Mbps (T2) channel. The transponder contains carriers for each station in the network. Operation in a single transponder allows the use of one operating upconverter and one operating downconverter, with full redundancy provided to maintain high reliability.

Because Earth station transmissions must be on different frequencies, FDMA can be complicated to arrange and even more difficult to change once the network goes into operation. One inherent advantage, however, is that the transmit power of the station can be tailored to its traffic requirement (i.e., the total number of carriers and channel capacity). Another advantage is that station transmissions are independent of each other (because they are on separate frequencies) so they can be mixed in with other unrelated carriers in the same transponder.
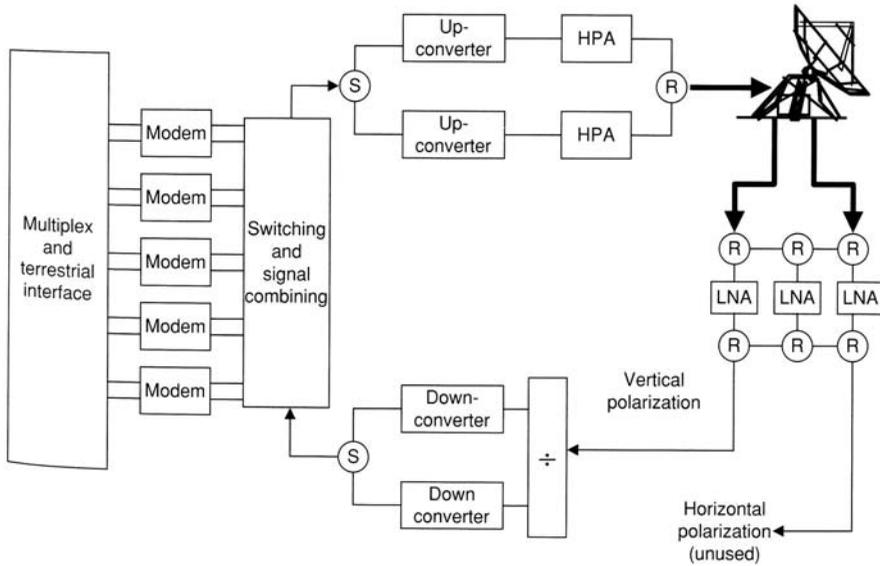
Figure 9.25 Example of a major communications Earth station with the capability to provide full duplex digital transmission FDMA service.

9.7.4 Full-Mesh TDMA Earth Station

TDMA Earth stations were first developed to use an entire transponder in the most efficient manner, transmitting at 60 to 120 Mbps to achieve the highest throughput possible [2]. Later on, narrowband TDMA was introduced to allow the use of small, lower cost RF terminals much like VSATs. As discussed in Chapter 5, there is some overhead in terms of guardtime and burst synchronization, so the actual throughput is somewhat less than the bandwidth would support. Also, it is desirable to use FEC to reduce the error rate for sensitive applications like digital video and corporate data transmission; therefore, throughput is further reduced by the coding rate. Still, TDMA is widely accepted in satellite communications because its mesh capability provides flexibility as well as its ease of integration into digital networks. Also, the digital elements of a TDMA terminal can be configured for different services and new locations can be added on a moment's notice.

Figure 9.26 presents a simplified block diagram of a full-mesh TDMA Earth station; the approach for star-network VSATs is covered in a later section.

9.7.4.1 RF Equipment for TDMA Service

The configuration of the RF terminal is similar to that of the video uplink. In full-transponder TDMA service, the G/T and the EIRP of this type of station usually demand an antenna of 10m or 7m for C- or Ku-band service, respectively. The reason for that is that TDMA establishes a true mesh network in which stations can transmit directly to one another through the same transponder. That is in contrast to the VSAT hub (described later in this chapter), which is the center point of the star network. For the type of TDMA station using a portion of a transponder, the antenna size is reduced by a factor of two to four, while the HPA must have
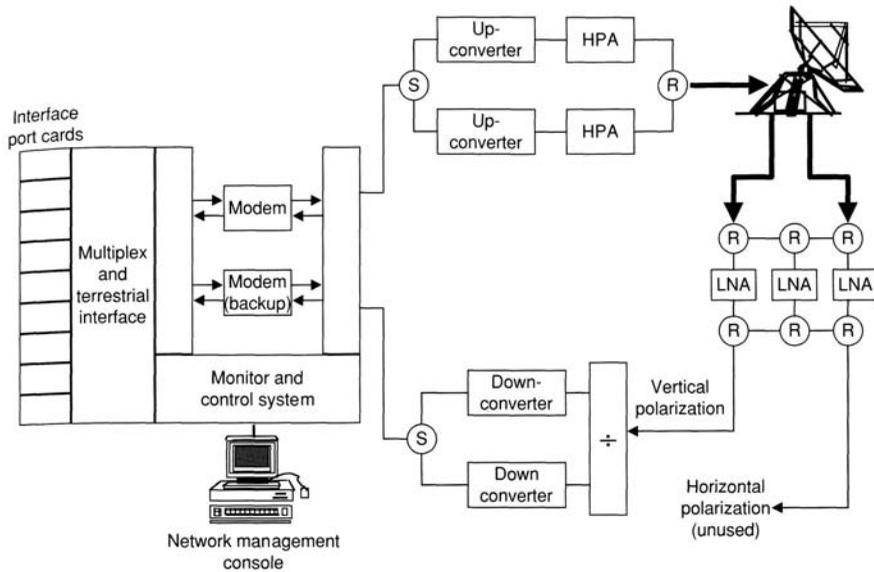
Figure 9.26 TDMA Earth station for full-transponder or partial-transponder operation with the capability to provide a variety of digital communications services on a full mesh basis.

an output power in the range of 20 to 400W, depending on link requirements. The upconverter and the downconverter are typical off-the-shelf items. An exception for the downconverter is if the TDMA network supports transponder hopping, in which case a frequency-agile LO will be required.

9.7.4.2 TDMA Terminal

Moving to the left half of Figure 9.26, the TDMA terminal section consists of an operating modem and a backup (if available), baseband equipment, multiplex, and M&C equipment. It is customary to purchase the terminal as an integrated subsystem as part of an integrated network from a TDMA manufacturer. The computer workstation connected to the M&C equipment is used to monitor the network and to program the particular terminal for the required traffic arrangement or routing. Usually, one station acts as the network master and is used to create the overall traffic routing pattern. Called the network map, the routing pattern is stored in each TDMA terminal to direct individual TDM communication channels to the proper destination. Depending on the time of day or the traffic demand, the network map is modified by appropriate programming of the central M&C computer. Central control for the network can be exercised from a location not at an Earth station since the computer terminal can be connected through a terrestrial line.

The design and control of the TDMA modem are critical to the operation of the terminal. This type of modem is intended for burst operation, that is, where the carrier must be turned on and off in rapid succession. The demodulator section has the difficult task of locking onto the incoming carrier at IF and synchronizing with its modulated digital information stream. That dual process must be accomplished for each received burst before the useful data actually can be recovered.

Typical burst transmission rates are 15, 45, 60, 120 and 240 Mbps (e.g., wideband TDMA); rates below 15 Mbps (called narrowband TDMA) allow several TDMA carriers to be placed in the same transponder, sharing its power and bandwidth.

The terrestrial interface is illustrated at the left of Figure 9.26, with an arrangement similar to that of the TDM multiplexer. In the multiplex section, buffering of the burst-rate digital stream adjusts the speed to a continuous rate typically matching the local digital hierarchy (PDH or SDH). Port cards are plug-in circuit modules used to configure the terrestrial interface for specific user applications. For example, if only T1 channels are desired, then there would be sufficient port cards to support a specific number of 1.544-Mbps streams. Other port card options include 56, 64, or 256 kbps (often a common port card design can be "strapped" for the desired channel rate), RJ-45, or other interface conditions. Adaptive features such as reprogrammable coding and information rates and statistical multiplexing also are available.

9.7.4.3 Carrier Hopping

The capacity of a given TDMA carrier, which is transmitted sequentially from several Earth stations, is limited by its bandwidth and power. When the sum of the channel requirements of the individual stations is greater than the capacity of the carrier, additional carriers must be provided. Carrier hopping is the technique whereby stations can switch between several such carriers to pick out traffic destined to them. Carrier hopping in wideband TDMA must be done between transponders; hence, the RF terminal is equipped with a frequency-agile downconverter. Narrowband TDMA usually is done in the same transponder, eliminating the need for multiple downconverters. This is simplified when using an L-band IF since all TDMA carriers share the same spectrum.

9.7.5 VSAT Star Network Terminal

VSAT technology brings the features and benefits of satellite communications down to an extremely economical and usable form. Corporate users can implement VSAT networks to obtain services that otherwise would require an extensive terrestrial infrastructure. Some VSAT star networks are designed around very compact and inexpensive terminals for Internet access or remote facility monitoring. The favorable economics result from sophisticated digital hardware and software technology in support of the data communication protocols common in business and government applications. The network employs a hub station to form a star network, with the host computer connected to the hub. Such a network, indicated in Figure 9.27, can provide transmission rates of up to 4 Mbps per remote terminal, but lower average data rates are more common.

The use of Ku-band (in the FSS portion of the spectrum) simplifies Earth station siting by eliminating the need for terrestrial frequency coordination. The higher power satellites permit more services to be carried, including digital video, using antenna diameters in the range of 0.7–1.8m. Implementation of a VSAT network is practical at C-band with somewhat larger antennas, which is mainly driven by the concern over adjacent satellite interference. The antenna size also is determined
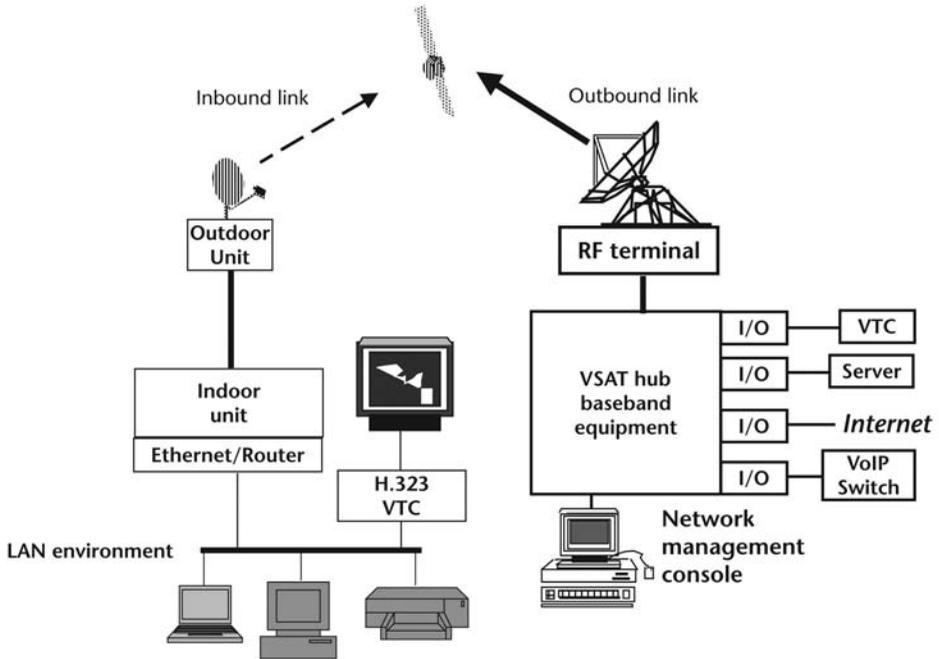
Figure 9.27 Architecture of a typical VSAT star network indicating a central hub Earth station and one remote VSAT.

by the required EIRP, indicated in Figure 9.11. The following sections discuss Earth station types that are used to implement star and mesh networks.

9.7.5.1 Hub and Network Control Station

The hub station of a VSAT star network is similar in size and capacity to the previously described TDMA terminal. That is not surprising because most VSAT star networks employ the TDMA access approach for the inbound (from VSAT to hub, also called the return link) and a relatively high-capacity TDM outroute (from the hub to the VSAT, also called the forward link). As shown in Figure 9.27, the hub consists of an RF terminal, a set of baseband equipment (including the modems, multiplexers, and encoders) and network M&C equipment. The RF terminal is designed and constructed in accordance with the principles covered previously. It provides a wideband uplink of one digital carrier per separate network and receives the narrowband channels from the remote VSATs themselves. The specific nature of these carriers depends on the architecture of the particular vendor and the parameters selected by the network operator. Generally, the outbound carrier operates at an aggregate data rate in excess of 10 Mbps and the inbound carrier at something less. The EIRP of the VSAT can thus be held to a low value, with the hub antenna being the exclusive receiving point. As a result, the HPA of the VSAT can be under 10W and may be as little as 1W. That imbalance in outbound versus inbound carrier level means there are always more inroute carriers.

In the baseband equipment, the modems provide the function of modulation of the digital outbound carrier and reception of the TDMA and ALOHA inbound

channels. The data are buffered and multiplexed in the baseband equipment to provide an appropriate interface to the host computer, which would be either close by or accessed over a tail link of some type. Software in the hub performs the necessary protocol conversion between the user interface and the space link, which typically is a proprietary multiple access and data communication scheme. That is because of the unique aspects of the satellite link, including its propagation delay and error rate performance. With proper protocol adaptation at the interface, error control over the satellite link, and buffering on the distant end, the satellite link is made to perform as well as a high-quality terrestrial link.

9.7.5.2 VSAT Terminal Configuration

The configuration of a typical VSAT with full-service capability is shown in Figure 9.28 and reviewed here. Photographs of Ku-band and C-band VSATs are presented in Figures 9.29 and 9.30, respectively. Highly reliable solid state electronics are used so that a single nonredundant string usually is adequate. Redundancy can be
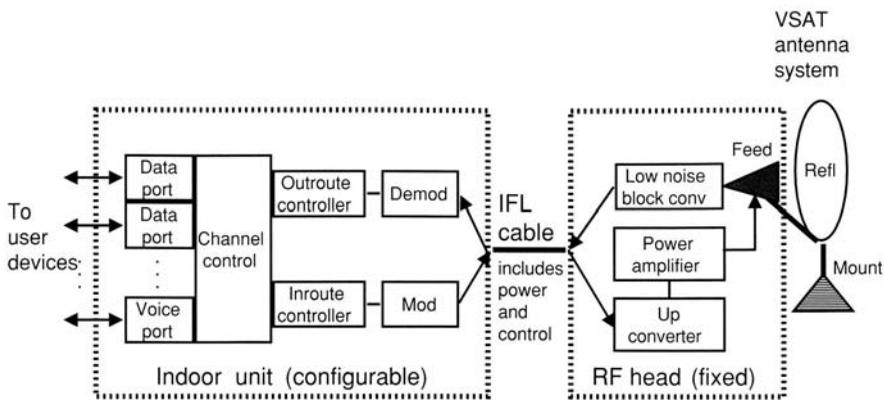
Figure 9.28 Configuration of a VSAT, which combines two-way voice and data with one-way video.

Figure 9.29 A typical fixed Ku-band VSAT capable of 18 Mbps forward and 4.2 Mbps return links. (Courtesy of iDirect Technologies.)

Figure 9.30 Photograph of a C-band VSAT installation with a 2.4m antenna.

added for applications demanding greatest reliability and where the VSAT might be difficult to reach by maintenance personnel. In situations in which the PSTN is available at the site, a dial backup facility can be added to the VSAT and hub to provide service in the event of an outdoor unit failure.

Backup can also be made available through an Internet connection, if available. The RF terminal of a VSAT is extremely compact and often attached to the antenna itself. As shown at the right of Figure 9.28, the RF equipment is composed of an LNB for reception and an upconverter and SSPA for transmission. The combination of block upconverter and SSPA is often referred to as, simply, a block upconverter (BUC). The antenna in the illustration is an offset-fed parabola with electronics mounted directly to the feed. For a GEO satellite network, a fixed mount is locked into position, although there still should be provision to repoint the antenna if a change of satellite location is ever contemplated. Other configurations were described in Section 9.2.3.

The remainder of the VSAT indoor equipment, consisting of a compact electronic box about the size of a home router, would be located in the building to which service is provided. As shown in Figure 9.28, the elements are functionally the same as those provided in the major Earth station described earlier in this chapter. All service would be rendered using the Internet Protocol at the RJ-45 interface. Because of the low transmit power and small diameter of the RF terminal, a VSAT typically is not capable of transmitting video. A way to get around that limitation is to employ compressed digital video operating at 128 kbps. The quality of such a two-way link is significantly poorer than conventional TV in terms of its response to rapid picture motion but could be acceptable for meetings and static pictorial information.

The techniques for transmitting and receiving data, both on a broadcast and an interactive basis, involve the use of random access packet transmission (e.g.,

ALOHA) as well as TDMA. For the first transmission from VSAT to hub, an ALOHA packet normally is uplinked. Messages may include a single line of a keyboard entry; a request to the hub for a file transfer; or a brief report requested by the hub. The TDMA mode is effective for long and relatively frequent transmissions and for voice and video communications, which require a constant bit rate. Switching between random access and TDMA is in response to local traffic demands, activated either by the centralized network management system or automatically when required by the VSAT itself.

The star network structure is assumed for the example in Figure 9.28. If the network is to be used extensively for switched telephone or video teleconferencing service, the configuration of the mesh is going to be far more effective. By allowing direct links between VSATs, the system avoids double hops and multiple hub stations interconnected by expensive terrestrial links. That can be achieved with TDMA depending on the network requirements. In a more or less traditonal FDMA demand-assigned scheme, a request for service enters the VSAT over the terrestrial interface and is transmitted to a hub, where the setup of the link is directed. The hub responds with digital instructions to the VSATs on both ends of the intended circuit to tune their SCPC modems to a frequency pair assigned for the duration of the call. Through that hub, the network management system monitors the use of the network and accomplishes such tasks as data traffic routing and billing.

Voice and data interface equipment is needed to make the VSAT appear to be the same as any telephone or data communication circuit. For example, the VSAT network may use a proprietary protocol to carry and process signaling information that the user's PBX does not understand. The voice interface then converts standard telephone signaling into the network protocol. Likewise, data protocols such as TCP/IP are converted to the VSAT network protocol in the data interface equipment. The appropriate protocol and address information are recreated at the distant hub station or remote VSAT. This is discussed in Chapter 3 and illustrated in Figure 3.11. In addition to signaling and protocols, the interface equipment can provide signal conditioning such as bandwidth compression and expansion.

Video teleconferencing services on a point-to-multipoint basis are well within the capabilities of VSATs. As shown in the upper left of Figure 9.28, a LAN at the distant end delivers data to a teleconference unit. The picture can then be carried to a conference room or theater. In addition to the video, the equipment might provide a voice interface for interactive question-and-answer service, using the VSAT to provide the reverse audio link.

This has been only a brief introduction to the types of Earth stations that make up a VSAT star or mesh network. The hardware designs and software technologies that facilitate these systems are undergoing an evolution to much smaller and cheaper installations. New Ka-band processing satellites are taking over some of the space segment needs. The purpose of those efforts is to expand the market into more types and sizes of businesses that can exploit private networks.

9.7.6   TV Receive-Only Design

A TVRO Earth station can take on one of a number of possible configurations, depending on the particular application. An installation used for a cable TV head

end must be capable of receiving more than one hundred channels at the same time. For a home installation, only one receiver is required (although two or more might be desired); however, that receiver should permit the viewer to change channels and orbit locations conveniently. In multiple-dwelling situations, such as an apartment building, there would be one antenna but several individual receivers for the different users.

9.7.6.1   TVRO Antenna Sizing

The size of a TVRO antenna is determined primarily by the RF power (i.e., the EIRP) of the GEO satellite. At the same time, international and domestic regulations limit C-band EIRP because of the possibility of RFI between satellites and terrestrial microwave services that share the band. No such sharing is required for much of the FSS and all of the BSS segments of Ku-band. As a consequence, the power level of C-band satellites is considerably lower than that of Ku-band satellites. That, of course, is the result of the regulatory limit on the design of the satellite transponder and has nothing to do with the physics of propagation. On the other hand, link margin (and hence EIRP) at Ku-band must be increased relative to C-band to adjust for the rain attenuation in the particular climatic region.

Figure 9.31 presents in graphical form the relationship between TVRO antenna sizing for parabolic reflectors in the range of 45 cm to 3m (shown along the X-axis) and satellite EIRP. A system noise temperature of 100K is assumed. Satellite EIRP values between 30 and 55 dBW (shown along the Y-axis) cover the typical ranges for C-band, Ku-band FSS, and Ku-band BSS. We have assumed that the satellite is transmitting a single wideband digital carrier with multiple TV channels that are encoded with the MPEG 2 format (according to the DVB-S standard). On the basis of power alone, there should be only one curve; however, the factor of rain attenuation forces the curve for Ku-band to move up the power range.
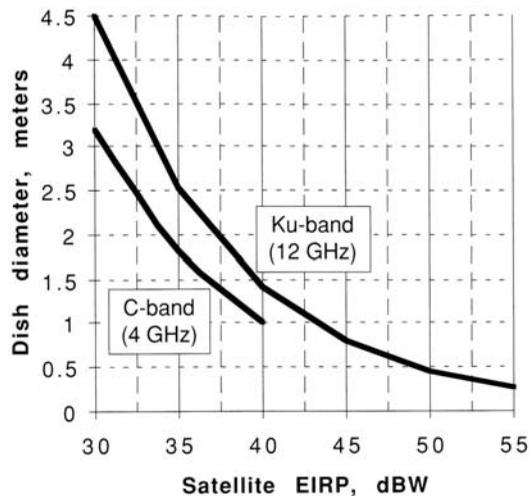
Figure 9.31   Antenna diameter of a TVRO home Earth station as a function satellite EIRP and frequency band.

The upper curve includes 2 dB of additional power to provide the same link reliability in the Ku-band range (FSS and BSS), which would be sufficient for temperate climates such as in the northeastern United States, Japan, or Europe, and with an elevation angle greater than about 30 degrees. Tropical regions with heavy thunderstorm activity and frequent torrential rains require even greater incremental power margin. The C-band curve stops at a diameter of 1m and an EIRP of approximately 40 dBW because of the international limitations on power flux density (orbital spacing is a factor as well). With the rain effect, it takes approximately 43 dBW at Ku-band to achieve the same link performance with the 1m antenna. In the Ku FSS range, an antenna diameter as small as 0.5m is adequate with an EIRP of 49 dBW. However, because FSS satellites are not intended for broadcast video services exclusively, the satellites are usually spaced 2° apart and ground antennas smaller than 0.7m may be subject to unacceptable interference.

The Ku BSS range is available for truly high-power DTH broadcasting, as shown in Figure 9.31. Satellites using the same frequency channels and polarization are to be separated in orbit sufficiently to allow the use of antennas as small as 45 cm in diameter. To achieve that at Ku-band with rain attenuation, the satellite must be capable of delivering an EIRP of approximately 50 dBW or higher. The power output of the appropriate transponder amplifier usually is around 150W (depending on the size of the footprint), nearly double that used in the FSS segment of Ku-band in North America.

9.7.6.2 Cable Head End TVRO Configuration

The cable head end TVRO, illustrated in Figure 9.32, is used to receive several commercial TV channels for distribution over a cable TV system. Four receivers are equipped in the figure; an actual CATV head end, however, would have one receiver per carrier. Each receiver, in turn, can recover several TV channels via a multiplexed group. While a single LNB is shown, it is more common to find two units: one for the vertical polarization and one for the horizontal. Online redundancy typically is not provided because of the high reliability that solid state LNBs usually provide. In the event of failure, the bad LNB can be rapidly changed out by technical personnel at the head end. The dc power for the LNB usually is carried
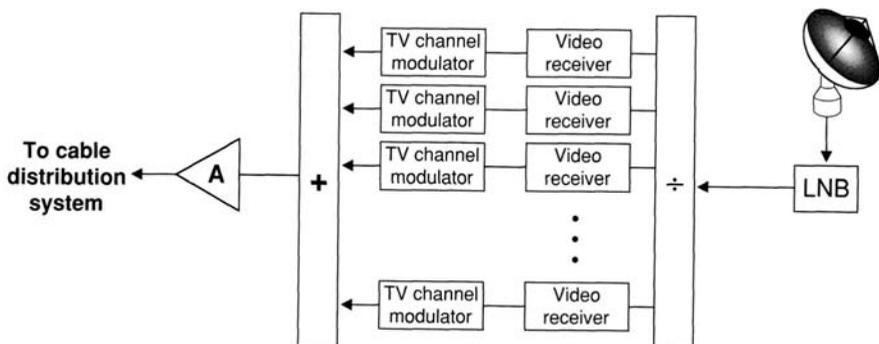
Figure 9.32 Configuration of a cable TVRO Earth station of the type located at a cable head end.

over the conductors of the same coaxial cable used to bring the RF signals back to the receivers.

The wideband output of the LNB is divided on a power basis to provide the entire spectrum to the video receivers. The channels received in a digital format are either transmitted directly over the cable by digital modulation or, in a hybrid analog/digital system, converted to analog and then remodulated as in the conventional analog TV approach. The cable distribution system (Figure 3.15) spreads the ensemble of channels out to cable TV subscribers. Modern cable systems provide each home with a settop box that allows the viewer to select from the full range of TV channels delivered over the cable. Because pay channels may be involved, the settop box has the capability to be programmed to allow the viewer to watch only those channels that have been authorized by the cable company (and for which the subscriber presumably is paying).

9.7.6.3   Direct-to-Home TVRO

In a direct-to-home installation, one video receiver also incorporates the TV channel modulator. Provision is made for convenient tuning of the receiver to any desired channel the viewer wishes to watch. The early C-band TVROs included an antenna with a motor-driven mount and adjustable polarizer. Such receiving systems were made user friendly by microprocessor control, allowing the viewer to key in the desired satellite and channel with a wireless remote control unit.

Modern digital DTH systems, such as that shown in Figure 9.33, use fixed antennas to reduce cost and operational complexity. By operating multiple satellites in the same orbit position and using digital video, the capacity of such a low-cost receive system can be over 200 full-color digital TV channels. A second feed doubles the total capacity and number of viewing options. The receiver demodulates and decodes the basic bitstream from the downlink carrier and demultiplexes the individual TV channels. The channel capacity per carrier typically is in the range of 5 to 12, depending on the total data rate, compression parameters, and whether statistical multiplexing is included. The latter has been introduced to nearly double the number

(a)

(b)
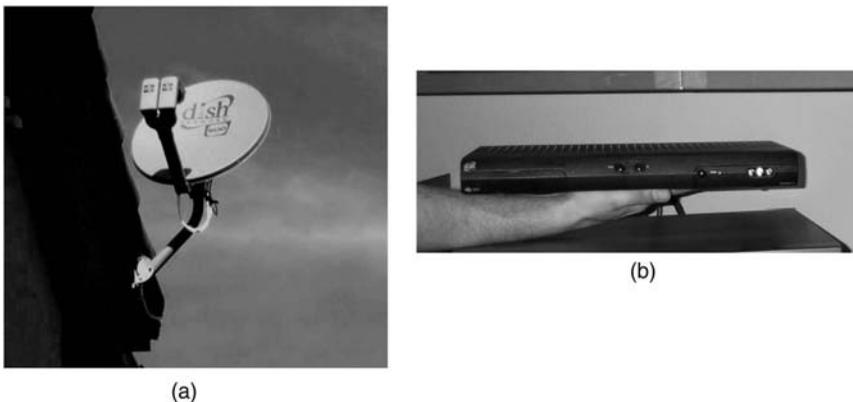
Figure 9.33   A typical DTH receiving system employing two BSS orbit locations. (a) 45-cm DTH reflector system, and (b) integrated receiver-decoder with personal video recorder.

of TV channels per carrier by dynamically adjusting the data rate to the motion demands of each channel. Expansion to three orbital positions was provided for DIRECTV by the triple LNB shown in Figure 9.34. Another important feature of the receiver is conditional access, which is discussed next.

9.7.6.4   Receiving Encrypted Programming

The TVRO must include a decoder to recover channels that have been encrypted at the uplink source. Several systems are in use in North America and Europe, each differing in the degree of disruption of the video and audio and in the degree of difficulty of unauthorized receiving of the signal. There has been a trend to adopt the DVB series of standards that were developed in Europe. Even with that standard, there are several proprietary encrypting and conditional access schemes from companies like News DataCom (NDC), and Irdeto. The actual transmission system is based on MPEG 2, with DVB providing the supporting arrangement for the multiplexing of added data for control and supervision.

9.7.7   MSS User Terminals

Mobile satellite communication is rapidly evolving, as new applications, and terminal types are introduced around the world. The types of terminals that were available at the time of this writing include marine (the forerunners of all MSS UTs), aeronautical, vehicular (primarily for trucks), laptop and handheld. The terminals are supported by the Inmarsat system of GEO satellites (particularly Inmarsat 4 with its spot beams and onboard processing) and certain regional GEO satellites.

Examples of leading types of MSS (L-band) terminals are presented in Figures 9.35 to 9.37, indicating the range of current concepts. A maritime terminal from

Figure 9.34   Triple LNB.

Figure 9.35   An example of an Inmarsat maritime broadband terminal. (Courtesy of Thrane & Thrane.)
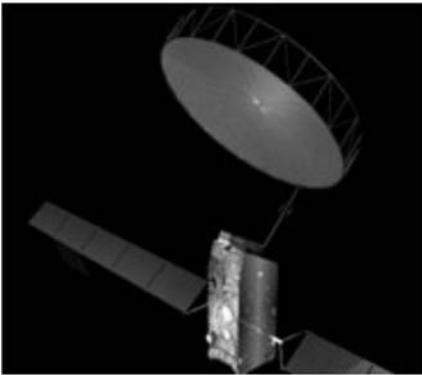
Figure 9.36   Illustration of the Inmarsat 4 GEO L-band satellite and an example of a user data terminal.

Thrane & Thrane capable of ISDN voice and medium data rate data service is presented in Figure 9.35. The antenna within the radome is 60 cm in diameter and the other equipment can be extended a distance of up to 100m. When operating over an Inmarsat 4 GEO satellite, this system provides the vessel with digital dial-up telephone service as well as Internet access for e-mail and Web-based applications. Figure 9.36 provides an illustration of the Inmarsat 4 satellite that is used to provide service both to the maritime terminal in Figure 9.35 as well as the portable land-based terminal from Hughes shown to the right. This portable terminal has very similar communications capability to the maritime terminal with the added benefit of a high degree of transportability.

Lastly, Figure 9.37 presents three different configurations of handheld MSS satellite phones. The first generation Iridium (LEO) phone, manufactured by

Figure 9.37 A selection of hand-held MSS satellite phones; the Iridium phone is shown fully deployed while the Inmarsat and Thuraya phones are collapsed in their storable configurations. (a) Iridium (Motorola) 9500, (b) Inmarsat iSatPhone, and (c) Thuraya SO-2501 (www.highspeedsat.com).

Motorola, is indicated in Figure 9.37(a) as it would be used with the antenna fully deployed. Designed and manufactured by Motorola, the model 9500 was the first to demonstrate the potential of handheld voice service via satellite communications. Newer phones from Inmarsat and Thuraya (both GEO systems), shown at Figure 9.37(b) and Figure 9.37(c), respectively, are smaller and lighter, but still require an extendable antenna much like that of the Motorola phone in order to avoid head blockage.

The approach being taken for handheld UTs is more in keeping with cell phone technology. They rely on compact circuit modules using application-specific integrated circuit (ASIC) components that operate at low power to conserve batteries. First-generation hand-held MSS UTs are about the same size as second generation cell phones (Figure 9.37). That is due to the added complexity of operating at L or S band and the fact that the initial manufacturing quantities tend to be not high enough to justify the full miniaturization of the product.

References

[1] Elbert, B. R., and B. Martyna, *Client/Server Computing: Architecture, Applications, and Distributed Systems Management*, Norwood, MA: Artech House, 1994.

[2] Schmidt, W. G., "The Application of TDMA to the Intelsat IV Satellite Series," *COMSAT Technical Review*, Vol. 3, No. 2, Fall 1973.

Launch Vehicles and Services

The technologies that make it possible to put one or more communications space-craft into Earth orbit are the result of rocket science and astrodynamics. Fortunately, we do not have to be rocket scientists to understand the use of rocketry and to work with the people and organizations that provide those services to satellite operators. Rather than attempt to cover those specialized fields, we have taken a practical approach in providing a basic understanding. Needless to say, the necessary systems work quite well, and although significant risks are involved, spacecraft have been placed into orbit, have landed on the moon, and are visiting faraway planets. Fifty years of experience with satellite operations have reduced those technologies to commercial practice, although we must never lose sight of the complexities that exist below the surface. For that reason, satellite operators rely heavily on the specialized capabilities of spacecraft manufacturers and the organizations that build and launch the rockets themselves.

In addition to a review of the technologies, this chapter also presents a summary of launch vehicle (LV) systems available at the time of this writing. This chapter should not be taken as the definitive word because the particular set of usable launch vehicles continuously changes in terms of both the particular rockets and their capabilities. Commercial launches were once the exclusive domain of the U.S. rocket manufacturers using the NASA launch facilities at Cape Canaveral, Florida. However, what has evolved is a truly international business, with good capabilities also available from European and Russian launch services organizations. While most of the launches support single GEO satellites, the advent of non-GEO satellite constellations reestablished multiple launch techniques for commercial service.

The selection of a particular LV system and service provider can be strategic to the success of the overall mission and business. The design and operation of the typical LV are complex and relate as much to technical design as they do to the consistency of the processes used in LV manufacture and operation. The three most important criteria in LV selection typically are the launch mass capability, the reliability or success record of the overall system and process, and the cost of use. A secondary but important consideration is the launch shock and vibration environment. This chapter reviews the technology and the systems in use, but it is not a substitute for a thorough investigation of the available options. Any reader needing to make such a selection should base it on current factual information provided by the LV providers, supplemented by other industry sources, such as the spacecraft manufacturers, experienced consultants, and launch insurance brokers.

10.1   The Launch Mission

Spacecraft are designed to be compatible with one or more LVs for placement into Earth orbit. Kepler's second law states that the orbital period scales as the 3/2 power of the semimajor axis. Therefore, the higher the apogee altitude, the larger the period of revolution. A good example is the geostationary transfer orbit (GTO), which has a period of approximately 16 hours. After perigee is raised to GEO altitude (around 36,000 km), the period increases to the prescribed 24 hours.

The launch mission is the sequence of steps that commence when the spacecraft and the LV leave the launch platform (on land, sea, or in the air) until the spacecraft is separated and safely on its way. An overview of that process for a GEO communications satellite is presented in Figure 10.1. Major changes in the trajectory and the orbit of the vehicle are provided by powerful liquid fuel or solid fuel rocket engines, which increase the velocity by the required amounts indicated in the figure. As discussed in Section 10.1.3, the successive burn technique is used with low-thrust liquid and electric engines. A good reason for first studying the GEO type of launch mission is that it includes other non-GEO orbit altitudes as intermediate steps. A big difference, however, is that LEO and MEO have significantly higher inclinations than would be practical for a GEO mission. That is why, as we will see later, the non-GEO spacecraft are launched toward the north from sites that are more northerly as well.

For the typical GEO launch mission, the booster lifts off from the launch platform and delivers the vehicle to an altitude of between 150 and 300 km. At that altitude, the vehicle follows a circular orbital path that is used in many LV systems as a parking orbit. The distance is exaggerated in Figure 10.1 because if shown properly to scale 300 km would appear like the thickness of a pencil line. Another rocket stage is then used to kick the vehicle into the elliptical GTO, where the farthest point (apogee) is at or near geosynchronous altitude (e.g., 36,000 km) and the closest point (perigee) is still at the same altitude as the parking orbit. That is because the satellite velocity is not high enough to sustain the altitude at apogee. Once the spacecraft is in transfer orbit, all control is exercised through the
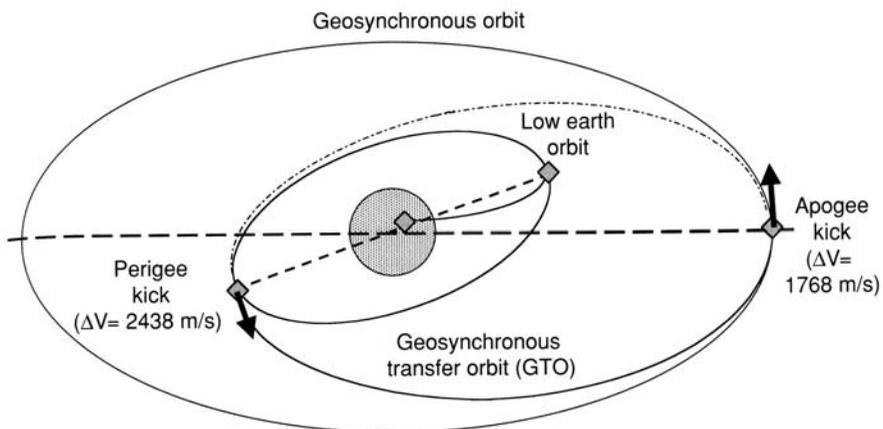
Figure 10.1   Major orbit changes to reach geosynchronous orbit.

TT&C station(s) and the SCC. Because the GTO is elliptical, its period is around 16 hours and the perigee point changes its longitude from revolution to revolution. The Earth's rotation under this orbit also contributes to the changing of the perigee point.

There are fundamentally two different strategies for moving from GTO to the 24-hour GEO. The original and simplest approach is to employ another high-thrust rocket stage to boost the satellite at an appropriate apogee (e.g., one that lies close to the final longitude). The last rocket stage should provide sufficient velocity increment (i.e., 1,768 m/s) to raise perigee to 36,000 km, thus circularizing the orbit. This stage is fired by ground command and is also used in many GEO missions to correct for orbit inclination of the booster. The latter will result if the launch platform is at a nonequatorial latitude.

An alternative to using a single high-thrust boost is to employ the spacecraft propulsion system to add velocity incrementally, through a number of successive burns. They are activated when the satellite is at apogee, so that perigee is raised by corresponding increments (Figure 10.2). The process takes from a matter of days to perhaps months, depending on the thrust available from the spacecraft propulsion system. Low-level thrusting from ion propulsion produces the longest period for orbit raising but yields greater efficiency (e.g., more useful payload weight to GEO), due to the higher I_{sp}.

Once perigee is equal to GEO apogee, the satellite will remain in orbit forever because there is insufficient atmospheric drag to cause reentry. Other initial touchup maneuvers and orbit corrections during the life of the satellite are made using the spacecraft's propulsion system (discussed in Chapter 8). Table 10.1 lists the events in a typical launch mission beginning with liftoff and ending with the start of operations in GEO. Considerations for non-GEO orbits are discussed later in this chapter.

10.1.1   The Boost Phase

To reach LEO or the parking orbit, an object must be elevated high enough above the Earth to escape the Earth's gravitational pull. That is done with a ballistic rocket stage that lofts the spacecraft, following a path that would return to Earth some distance from the launch site. A second rocket burn pushes the vehicle into Earth orbit. Kepler's first law governs the motion of the vehicle in orbit, wherein the satellite follows an elliptical path with the Earth centered at one focus. The satellite stays in orbit because the force produced by centripetal acceleration is equal to the gravitational pull of the Earth.

Because the mass of the satellite is insignificant compared to that of the Earth, the time period of the orbit is completely determined by the velocity of the satellite. (Pertinent information about this factor is provided in Chapter 1.) Launch vehicles consist of multiple rocket stages to satisfy the demand for fuel. Release of spent stages is beneficial because it reduces the dead mass that the propellant must push into orbit. The second stage of the booster is pointed along the tangent to the desired orbit, and the total thrust increases the velocity of the vehicle to the required value. At a point where the pressure induced by the atmosphere is low enough for safe deployment, the protective shroud around the spacecraft is jettisoned. In the
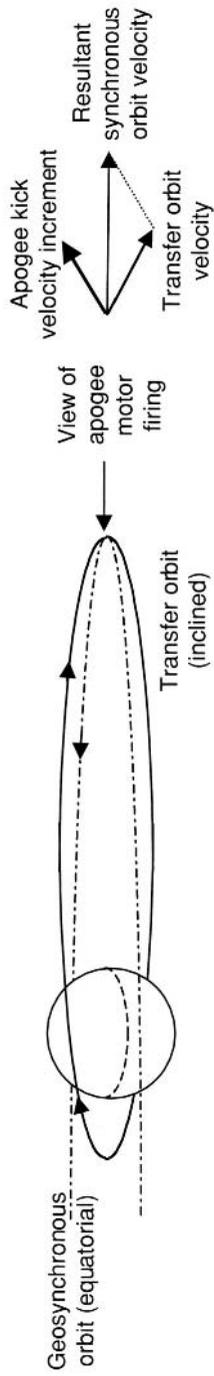
Figure 10.2 Injection from transfer orbit into geosynchronous orbit using apogee kick maneuver.

Table 10.1   Launch Sequence for a
Geostationary Orbit Mission
Identifying Key Events and Ground
Support Activities

1. Liftoff of main booster
2. Parking orbit achieved
3. Transfer orbit injection
4. TT&C link established with ground
5. Reorientation to AMF attitude
6. Preburn RCS maneuver
7. Apogee engine firing to achieve GEO
8. Drift to assigned station
9. Orbit and attitude adjustments
10. Despin or reorientation
11. Spacecraft or body deployments
12. Bus testing
13. Payload testing
14. Start of service

parking orbit, the vehicle makes one revolution around the Earth in approximately one and a half hours. Since that is too fast for conventional TT&C stations to track the LV, operation of the LV's first and other stages is automatically controlled by an onboard computer.

The thin atmosphere that extends to the parking orbit is minute, but there is sufficient drag to cause some slowing of the vehicle. If not corrected in time, the orbit gradually decays and eventually the vehicle reenters and burns up. Because that can take from months to years, there is little risk to the satellite between the time it reaches the parking orbit and when the third stage propels it to GTO. There can be circumstances in which the second stage does not provide its full increment of velocity and the vehicle is then on a suborbital or even ballistic path back into the denser atmosphere. Thus, timely firing of the third stage is extremely important to prevent loss of the mission. In a LEO mission, discussed in Section 10.1.2, this orbit is at or close to the final orbit, so preoperational activities can commence.

There is an issue that relates to the geographical latitude of the launch site. Nearly every launch site is located some distance from the equator, while the GEO is in the plane of the equator. (An exception is Sea Launch, which uses a floating platform that is towed to a location in the Pacific Ocean at the equator.) The parking orbit produced by the second stage may be at an angle with the equatorial plane; this is referred to as inclination of the orbit. This inclination is approximately equal to the latitude of the launch site and must be removed for a GEO mission. This is accomplished using rocket energy at some point prior to putting the satellite into commercial service. Obviously, the closer the launch platform is to the equator, the less additional rocket energy is needed to correct for inclination. The counter to this is that inclination is not a problem, per se, because it can be corrected by subsequent thrust maneuvers from either the LV or the spacecraft propulsion systems.

10.1.2   Non-GEO Missions

Satellite constellations that operate in LEO do not require the use of a transfer orbit. Instead, the LV injects the spacecraft into the appropriate LEO, where they

may continue to orbit the Earth. Atmospheric drag is definitely an issue, as are radiation effects from the lower extreme of the Van Allen Belt. Those considerations can be taken into account in the design of LEO spacecraft, whereby sufficient orbit maintenance fuel and electronic device shielding are provided. Between the constraints imposed by those two factors along with investment cost and satellite capacity, LEO spacecraft generally are designed for a shorter life than are GEO or even MEO spacecraft.

A standard type of LV can place substantially more spacecraft mass into LEO than GEO. More important, LEO spacecraft need to be deployed in such a way as to fill a specific orbit plane. Therefore, LEO spacecraft sometimes are launched in a bunch of four to eight. In addition, current LEO constellations use polar or highly inclined orbits. That means the launch site can be quite far from the equator with the LV trajectory pointed toward the north or northeast, as opposed to the east.

10.1.3 Geostationary Transfer Orbit

GEO is reached by first placing the satellite into GTO with apogee at or near 36,000 km and perigee at the altitude of the parking orbit (150 to 300 km). The term *perigee kick* is used to describe the action of the last LV stage, which may either be part of the LV or provided separately with the spacecraft. (The mechanical aspects of those alternatives are reviewed later in this chapter.) The LV may perform the perigee kick with a single burn of either a liquid or solid fuel rocket motor. An alternative strategy, employed by the Proton rocket, is to perform two burns: the first to provide the basic perigee kick and the second at apogee to increase perigee altitude. That also helps compensate for the relatively high latitude of the launch site (Baikunor, Kazakhstan, in the case of the Proton).

Figure 10.2 shows how the transfer orbit is initiated from parking orbit wherein the perigee kick stage is fired at the point opposite from where apogee is to occur. The mission plan usually specifies that the first apogee after injection into GTO must be in view of a TT&C station at a specific location. Conversely, the firing position may be fixed for the particular launch site and LV, forcing the location of the TT&C station to be in view of the resulting first apogee. The period of GTO is approximately 16 hours, and the satellite is easy to track near apogee when its motion is slowed considerably. Because each successive apogee occurs at a different longitude on the Earth, it is advantageous to have TT&C stations in the eastern and western hemispheres. That can be accomplished without owning additional sites by contracting with a satellite operator in the opposite hemisphere for the use of its TT&C station during the transfer orbit phase of the mission.

The final phase of transfer orbit operations is the injection into geosynchronous orbit. At the right of Figure 10.2 is shown the vector representation of the velocities before and after injection, looking toward the point of apogee motor firing (AMF) from the right. In the figure, the satellite velocity vector in the inclined transfer orbit at the point in question is at the angle with respect to the geostationary orbit plane. To cause the final synchronous orbit velocity to be in the equatorial plane, the spacecraft is oriented so the apogee kick motor fires along the vector direction shown pointed upward. The thrust increment at AMF can correct for inclination if

the velocity change is sufficiently greater than that required to produce synchronous orbit. Therefore, the resultant velocity is in the equatorial plane (i.e., horizontal in the vector diagram) and its magnitude is sufficient to circularize the orbit at 36,000 km altitude.

There will always be some inaccuracy in the amount of velocity change, including the precise direction, so the orbits achieved will have some error as well. That is multiplied by the fact that the first key rocket stages are under autonomous control and cannot be guided from the ground. There is a need to use ground-commanded corrections to produce the desired orbit geometry. That can be done only by first precisely locating the satellite using the TT&C station and then determining its orbit parameters. Touchup maneuvers are then performed with the spacecraft reaction control system. For a GEO mission, the particular transfer orbit where AMF is accomplished is selected for operational reasons. For example, it is desirable to fire the AKM at an apogee that is as close as possible to the final orbital longitude of the satellite and thus minimize the time for drift movement from AMF longitude.

The preceding discussion was for a single apogee thrust using a solid fuel rocket motor. However, the majority of spacecraft (including all spacecraft heavier than approximately 2,000 kg dry mass) use liquid propulsion to raise perigee in successive increments. As illustrated in Figure 10.3, the process requires multiple burns of a liquid apogee motor (LAM) that uses either a dedicated bipropellant propulsion system or shares tankage with the reaction control system used for stationkeeping. In addition to the reliability of a redundant RCS, the desired orbit can be achieved without specific touchup maneuvers.

10.1.4   Drift Orbit for GEO Operation

Following a large-burn apogee injection, the satellite is in a near geosynchronous orbit and drifting toward the assigned orbit location. The LAM approach typically
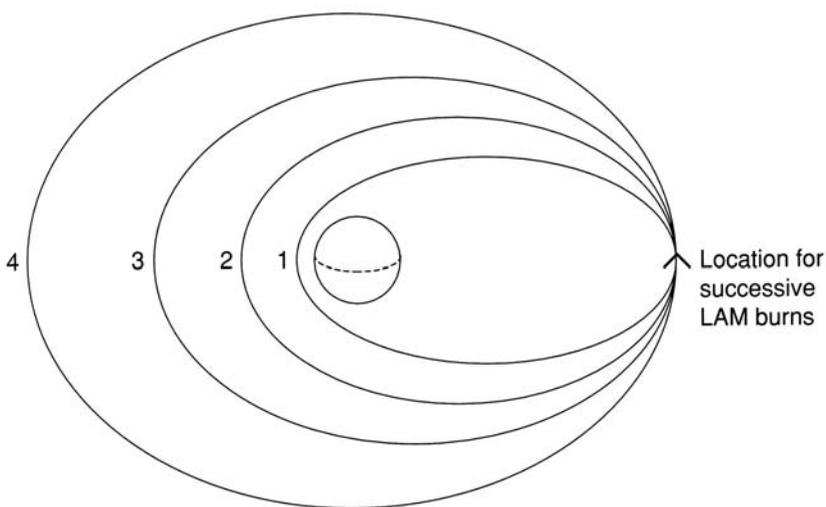
Figure 10.3   Injection from transfer orbit in geosynchronous orbit by successive burns of a liquid apogee motor (LAM).

produces a proper drift orbit without injection errors. For a solid AKM, the detailed parameters of the initial drift orbit are unknown after motor firing, so the first priority of TT&C operations is to determine the orbit as accurately as possible. The next step is to make the necessary velocity corrections to cause the satellite to drift in the desired direction. Figure 10.4 shows how a satellite in drift orbit is actually at a different altitude than the geostationary orbit, which is a consequence of Kepler's second law. That limits the possibility of a drifting satellite bumping into a stationary one. A satellite drifting westward is above the geostationary orbit, while one drifting eastward is below. However, there still remains the possibility of RFI as the drifting satellite passes through the antenna beams of the Earth stations, which point at their respective geostationary satellite longitudes.

Potential problems with RFI in drift orbit are prevented through detailed coordination between satellite operators. Prior to reaching drift orbit, the manager of the mission provides written notification to the operators of every satellite that could conceivably receive interference. Then, during drift, the uplink from the TT&C station and the satellite payload are disabled when there is inadequate orbital separation (typically 2 degrees of separation at C- and Ku-bands, but the specific amount should be determined prior to the mission). The harmful interference that could result provides the motivation for cooperation between operators, some of which are potentially competitors in business. Spacecraft deployments and testing could be conducted during drift orbit or at a safe test location when the satellite is definitely out of view of operating ground stations, but it is usually best to defer such activities until the satellite reaches its assigned orbit position.

A predetermined amount of reaction control system (RCS) fuel is consumed to stop the satellite from drifting. Final touchup maneuvers with the spacecraft RCS bring the velocity and altitude into alignment with the geostationary orbit. From that point forward, the stationkeeping phase of the mission begins and lasts through the rest of the satellite lifetime. The satellite can be relocated to another longtitude by expending two equal amounts of fuel: one to start the drift and the other to stop it when the final longitude is reached. Because the duration of thrusting determines the speed of relative motion, the amount of fuel consumed is roughly inversely proportional to the total time allotted for relocation.
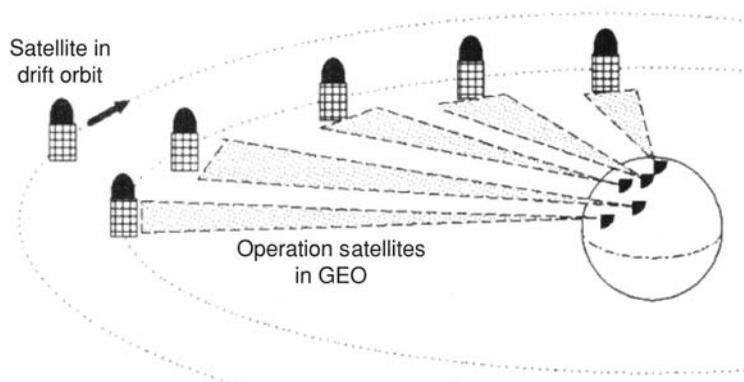
Figure 10.4   Drifting of a geosynchronous satellite from the apogee injection longitude to the final longitude position.

10.1.5   Deployments and In-Orbit Testing

Whether GEO or non-GEO, a communications satellite is delivered to orbit in a somewhat compact condition, consistent with the volume and mounting constraints of the LV. Part of the mission involves preparing the spacecraft for service by deploying appendages like solar panels and reflector antenna systems and thoroughly testing all the active subsystems of the bus and payload. Procedures to calibrate and preload data and software are usually required prior to start of service. All of that can take as little as a few days or as long as several weeks, depending on the tasks involved and the complexity of the subsystems onboard the satellite or the nature of the operation.

GEO satellites probably are the most complex and require more attention during this phase than might the non-GEO satellites. That is because modern GEO satellites tend to be larger in physical and electrical terms to overcome the distance and to provide adequate communications capacity and coverage to justify the required investment. In constrast, some of the smaller LEO satellites are quite simple in mechanical and electrical terms and hence could be placed into service almost immediately after activation in final orbit. That is also driven by the issue of the minimum required quantity of LEO satellites needed to provide the total service from a complete constellation.

To provide some guidance as to the amount of effort involved, Figure 10.5 shows the deployment sequence for a typical three-axis GEO satellite. This particular model requires that the solar panels be deployed early in the mission, and the use of a testing longitude (which could be other than the final assigned longitude if an operating satellite is already in that position). All the subsystems that make up the bus are checked for function and then tested for quantitative performance. Often, the spacecraft manufacturer or operator needs to determine the calibration of sensors and other devices before exercising the entire spacecraft.

Testing of the communications payload can commence once the bus has been tested and configured for operation. Payload testing can be carried out as a brief

1	Separation from launch vehicle	6	Deploy reflectors
2	Reorient to sun-normal	7	Deploy solar panels
3	Apogee injection boost sequence	8	Sun acquisition and bus testing
5	Reorient for deployments	9	Earth acquisition and start of operations
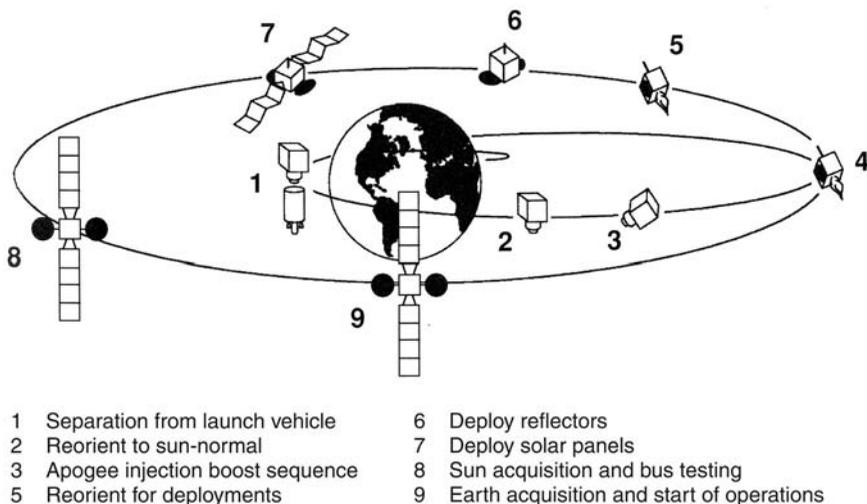
Figure 10.5   Deployment sequence for a typical three-axis satellite (transfer orbit mission summary).

verification that the payload has withstood the launch mission and provides the basic functions. On the other hand, it is typical to fully exercise all channels and types of communication and coverage before entering service. Such a process can be accomplished in a matter of days for a simple bent-pipe type of payload with a single coverage footprint. The only time-consuming aspect is verifying the full coverage of the spacecraft antenna system. Recall that the typical repeater has both an uplink and a downlink band, as well as two different polarizations in each. There likely will have to be tests to determine the cross-polarization performance of all beams. Depending on the demands of frequency coordination, it even might be necessary to verify levels of interference into other countries or regions.

Looking now at more advanced payloads, complexity increases rapidly for greater numbers of individual channels, multiple spot beams, and processing repeaters. As pointed out in Chapter 7, spot beams impose a unique difficulty in that a given testing Earth station may not be able to receive its own transmission directly. The simplest approach might be to move the spacecraft body, allowing each beam to migrate past the main TT&C site. Alternatively, measurement signals could be relayed by a distant station within the beam in question. One concept that was employed in the past was to make separate measurements at different points on the ground using a mobile vehicle containing the necessary test equipment and support personnel.

Consideration of normal operation of the satellite system during its lifetime is given in Chapter 11. Right now, we go into a discussion of how onboard fuel is utilized throughout the mission.

10.1.6   RCS Fuel Allocation

The operational stage starts when final orbit has been achieved. A GEO satellite requires frequent correction of orbit parameters to ensure that the desired longitude and low inclination are maintained. Otherwise, the satellite will not remain within the beamwidth of the largest class of ground antenna (e.g., ones lacking automatic tracking). That critical requirement and its consequences for fuel allocation are discussed here.

If the Earth could be represented by a sphere of uniform mass density (or at least composed of concentric spherical shells of uniform mass), the GEO satellite would stay put at this longitude indefinitely. The reality, indicated in Figure 10.6, is that the Earth is not uniform. To that is added the presence of gravitational pull from the sun and from the moon. As a result, the orbit distorts over time so that the satellite appears to move relative to a point on the Earth. First, the orbit becomes somewhat elliptical, a property called eccentricity, and the plane of the orbit inclines relative to that of the equator. Eccentricity is not particularly troublesome and is easily corrected along with the primary orbital adjustments. Inclination, on the other hand, makes the satellite appear to move perpendicular to the plane of the equator, with the period of motion being 24 hours. The bulges in the Earth's mass produce east-west drift along the orbit. With the satellite as shown in line with the bulges, the gravity is in balance and the satellite tends to stay put. Likewise, positions 90 degrees away in the orbit have equal pull from each bulge, and no force acts to move the satellite. Intermediary positions experience a maximum force
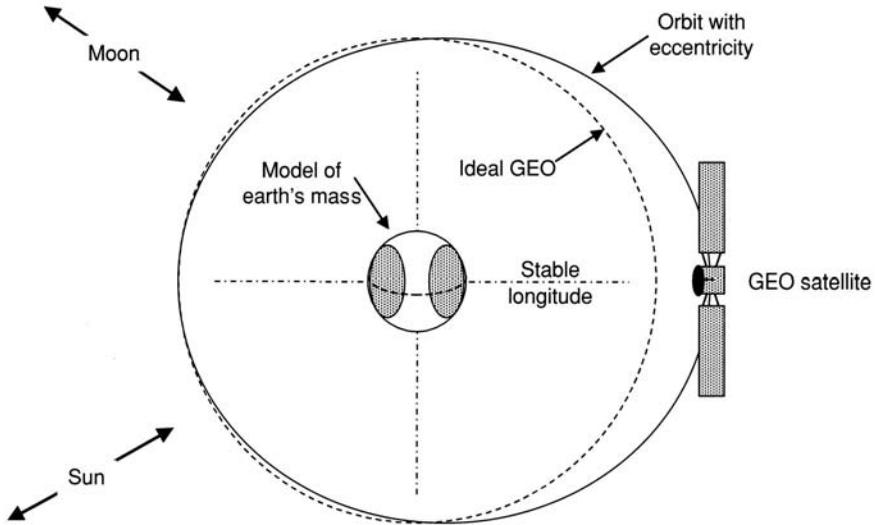
Figure 10.6   Variations in the geostationary orbit in the plane of the equator.

that tneds to drift the satellite toward an equilibrium longitude. Fuel usage for a satellite positioned at the worst longitudes is still a small fraction of the total budget, which is dominated by the required to remove inclination (north-south stationkeeping). The actual physical and mathematical relationships are much more complex than that; however, this simple model can be a way of visualizing the three forms of orbit distortion.

Most stationkeeping fuel of a GEO satellite is consumed to correct inclination. Properties of the orbit are shown in Figure 10.7, which has been exaggerated for clarity. Produced by gravity from the moon and, to some extent, from the sun, inclination increases by approximately 0.8 degree per year. For a given inclination angle, the satellite is below the equator at one point in time and then the same
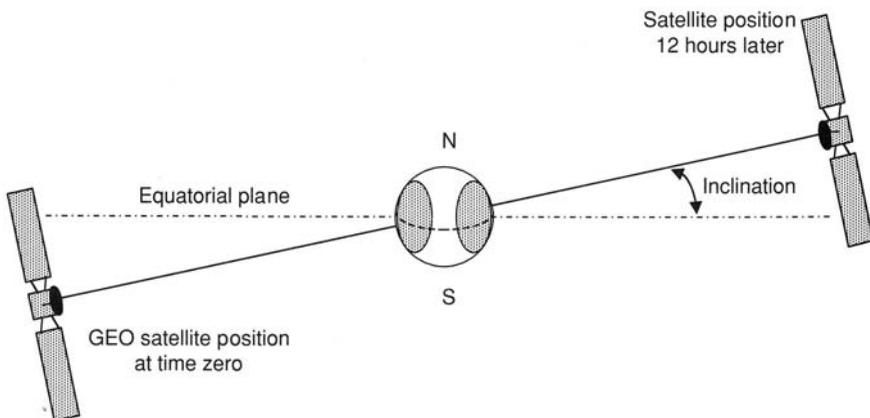
Figure 10.7   Relative position of a satellite in an inclined geostationary orbit as seen over a 24-hour period.

distance above the equator precisely 12 hours later. When viewed from the ground, a satellite appears to oscillate in that manner over a 24-hour period.

The reaction control system of the spacecraft is used to provide several of the preoperational velocity increments as well as those necessary for stationkeeping. Figure 10.8 presents an approximate allocation of liquid RCS fuel for an entire GEO satellite mission from transfer orbit to end of life. In this example, the RCS is not required to deliver the bulk of the velocity increments for perigee kick or apogee kick. One point of clarification is that east-west motion is in the plane of the orbit while north-south motion is perpendicular to it. From the Earth station's perspective, those directions usually are not aligned with the local vertical and horizontal because of the curvature of the orbital arc as seen from the particular point on the ground (except, of course, for an Earth station on the equator at the same longitude as the satellite).

Once on station, the orbit is corrected during stationkeeping, introducing velocity changes, which are numerically determined by the specific orbit parameters. However, the amount of fuel consumed to produce the necessary velocity increments is proportional to the average total mass of the satellite.

As is evident from Figure 10.8, approximately three-quarters of the fuel is required for north-south stationkeeping (correction of GEO inclination). East-west stationkeeping, while critical to every geostationary satellite, requires from zero to a maximum of 10% of total RCS fuel. Both usages are predictable over the satellite's lifetime, and little can be done to extend the mission once geostationary orbit has been reached (except by terminating north-south stationkeeping early). One exception is the use of a storage orbit to conserve fuel before the satellite is placed into service. That involves injecting the spacecraft into an inclined geosynchronous orbit so the inclination decreases naturally toward the equatorial plane (which is at the rate of approximately 0.8 degree per year). Thus, an initial "bias" of
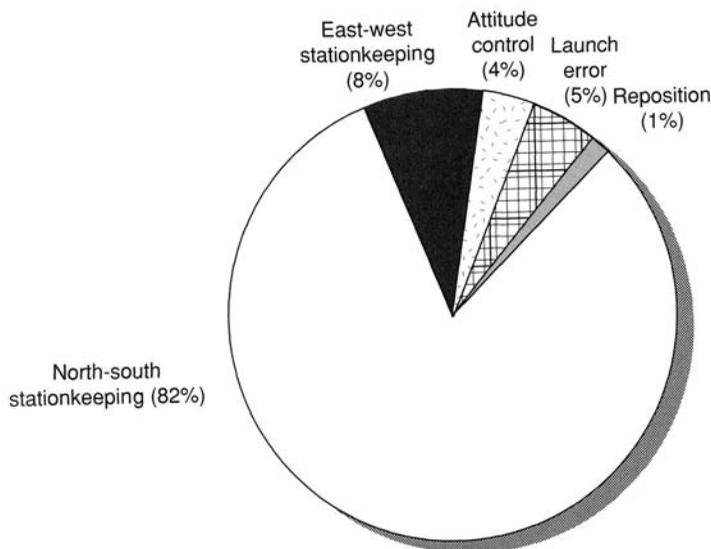
Figure 10.8   Allocation of liquid RCS fuel usage for a typical geostationary satellite over a 15-year period.

3 degrees provides about 3 years of storage potential. A satellite can be brought out of storage prematurely by consuming fuel to remove the remaining inclination.

The previous discussion relates to stationkeeping performed using liquid fuel such as monopropellant hydrazine or bipropellant MMH and nitrogen tetroxide. Many larger GEO satellites employ some form of electric propulsion to reduce the mass of fuel required for a 15- to 20-year mission. A reduction in fuel mass transfers directly into more mass for payload or electric power (some of which may be needed for thrusting). As discussed in Chapter 8, electric propulsion is an effective way to make this trade-off, although there have been some difficulties with short-ened thruster lifetime. This can be countered using redundant thrusters and having some liquid fuel available as a backup. The latter will come out of the hardware allocation of the spacecraft. Using electric propulsion, the physical fuel mass is quite small and the bigger impact is for the increased mass of the thrusters and other circuitry. It would not normally be required to thrust during eclipse, thus avoiding an increase in battery mass. On the other hand, some of the added power for the thrusters may result in a somewhat larger requirement for solar array mass. In general, the trade-off is very favorable in favor of electric propulsion, which is why these systems are incorporated in large satellites.

Most of the attention of SCC personnel is directed toward accuracy in orbit determination and fuel use during routine stationkeeping maneuvers. A small reserve for repositioning provides for moving the satellite to a new longitude if required for some reason. Recall that half that fuel would be used to start the drift and the outer half to stop it when the desired position is reached. A retirement budget allocation is used at end of life to accelerate the satellite into a higher drift orbit, where it will be turned off so it cannot interfere either physically or electromagnetically (e.g., cause RFI) with operating satellites.

A GEO satellite that nears its end of life in terms of north-south stationkeeping fuel can potentially have its useful life extended by operating the satellite in an inclined orbit. That amounts to nothing more than halting north-south stationkeep-ing operations while some fuel still remains. That fuel can be used to continue east-west stationkeeping. Because the amount of fuel needed per month is substan-tially less than for north-south, it can be projected that the life would be several years longer (depending on the fuel remaining and the particular orbit longitude). There is a tradeoff here because while the satellite likely will continue to operate (the power and other spacecraft subsystems allowing), the inclination will build up at the rate of 0.8 degree per year. That adds to the north-south pointing range as seen on the ground. Earth stations would either have to track the satellite (common for large antennas used for TV uplinks and teleport Earth stations) or have correspondingly wide beamwidths (typical for small DTH antennas and mobile user terminals).

10.2   Launch Technology and Systems

The LV situation is an ever changing scene, as new rocket configurations replace some older counterparts and new LV service providers appear in the marketplace. However, because of the complexities and risks associated with launching satellites,

it is often the rule that newer vehicles are merely modified versions of their predecessors. There are always exceptions when new approaches prove themselves in technical and operational terms, as was the case for the Space Shuttle, Pegasus, and Sea Launch. It is likely that future commercial launches still will rely on expendable rockets of improved design, many of which have familiar names like Ariane, Atlas, Delta, Long March, Proton, and Soyuz.

This section reviews the evolution of commercially available launch vehicles suitable for placing a spacecraft into GTO. As discussed in Section 10.1, that usually involves three stages: a first and second booster and a perigee kick stage. In some systems, the perigee kick stage is part of the LV, while in others the spacecraft provides its own integral perigee boost. The discussion begins with an overview of the technology and evolution of current LVs.

While the details of rocket technology are beyond the scope of this book, the basic alternatives are relatively easy to understand. A liquid-fueled first stage uses separate fuel and oxidizer supplies that burn when combined, producing the necessary thrust. The engine uses pumps to maintain constant fluid pressure, and temperature is controlled by piping the cooler liquids around the hotter parts of the engine. Typical fuel/oxidizer combinations include hydrazine/nitrogen tetroxide (i.e., bipropellant approach described in Chapter 8), kerosene/liquid oxygen (used by the Saturn 5 rocket of the Apollo program), and cryogenic liquid hydrogen/liquid oxygen (used in the main engines of the Space Shuttle).

Three kinds of "hydrazine" are utilized for spacecraft and LV propulsion:

1. Anhydrous hydrazine, N_2H_4, is used in monopropellant hydrazine systems in some satellites. It requires a catalyst to start decomposition, yielding an I_{sp} of around 230 seconds. N_2H_4 is somewhat unstable. Preheating N_2H_4 increases thrust and I_{sp}.
2. Monomethyl hydrazine (MMH), $(CH_3)NHNH_2$, is in bipropellant systems like most GEO spacecraft today. With N_2O_4 (nitrogen tetroxide) as an oxidizer, one can have I_{sp}s of around 300 seconds. It is more stable than anhydrous hydrazine; in fact, it can be added to anhydrous hydrazine at the 10% level to stabilize the latter fuel (it removes N_2H_4's monopropellant qualities).
3. (Unsymmetrical) dimethyl hydrazine (UDMH), $(CH_3)_2NNH_2$, is in bipropellant systems for launchers in both the West and the East with various oxidizers. It can be mixed 50-50 with anhydrous hydrazine, the resulting mixture is called Aerozine 50.

Cryogenic propellants must be kept at an extremely low temperature so they remain in liquid form and do not revert to a gaseous state. The tradeoff between high energy cryogenic propellant and propellants that can exist at room temperature is that the latter can be stored for long periods of time (i.e., in a missile silo or in space). Solid fuel rocket motors offer a convenient alternative to liquid because of the simplicity of design and integration and because they can be stored for years at a time. That is the same technology used in the AKM of some spacecraft. (The technical approach to solid motor design is presented in Chapter 8.) The two large rockets attached to the Space Shuttle are solid rocket motors (SRMs) filled with

highly combustible propellant material. Solids are attractive for that application because they offer a great deal of boost during critical parts of the launch sequence. Solid rocket motors burn rapidly and at a high temperature, requiring great precautions to be taken with their exit nozzles, cases, and attachments to the vehicle.

The second and third stages of the booster can take the form of solid or bipropellant liquid-fueled rockets. As explained in Section 10.1, the second stage is ignited after the LV is above the denser part of the atmosphere, propelling the vehicle into the parking orbit. Separation of the stages may involve the firing of special small rocket motors or explosives that sever bolts holding the stages together. Clearly, the sequence of ignition, burn and velocity control, ignition, and separation is vital to the achievement of mission success. Much of the attention and concern of launch operation personnel is focused on this critical sequence, which occurs automatically without ground control.

Preparation of the LV and the spacecraft is another vital area, one that may be underestimated. Many modern LVs are capable of launching two or more payloads at the same time. Even with a single payload, the preparations at the launch site take one or more months since the spacecraft must be checked and then properly integrated with the LV. The launch agency, rocket manufacturer, and spacecraft supplier work essentially around the clock during this period to assemble the LV, install the payload or payloads, check and fuel the system, and conduct various prelaunch tests and rehearsals. The facilities involved are extensive, and there are only a few qualified launch sites in existence. Also, the tracking sites required during each phase of the launch and transfer orbit all must be prepared and checked out prior to liftoff.

This discussion is necessarily brief and of a general nature. Section 10.3 reviews a number of specific LV systems. It is evident that each LV is a unique combination of rocket design, operation and fuel, which should be considered in detail when making a selection for a particular mission. That is because the success of the launch mission requires that each function and step occur properly or at least within acceptable limits. Failure of just one hardware or software component can be all it takes to preclude mission success. Ability to recover from serious launch failure generally is quite limited, due to the large thrust levels and velocity changes required from the various stages of the LV.

10.3   Typical Launch Vehicles

The LV industry became commercialized in the late 1980s following the withdrawal of NASA from this market. Arianespace of Europe was the first to offer commercial launches, followed by Lockheed Martin with Commercial Titan and McDonald Douglas with the venerable Delta II. Boeing joined in the 1990s by introducing a new launch concept with an ocean-going platform and Lockheed Martin joined forces with Russian rocket manufacturer Khrunichev with International Launch Services (ILS). The Convair division of General Dynamics produced several Atlas Centaur rockets for commercial and government launches.

Boeing acquired McDonald Douglas and Lockheed Martin acquired Convair, giving each company the respective rocket resources. Through a series of sales and

joint ventures, the market has rationalized itself into just three main providers of commercial launches: ILS (now an independent company), SeaLaunch (backed principally by Boeing), and Arianespace. Lockheed Martin and Boeing merged their government launch capabilities into a joint venture company called United Launch Alliance (ULA), which no longer supports commercial customers. However, the Delta and Atlas are offered to commercial customers directly through Boeing and Lockheed Martin, respectively.

China Great Wall Industrial Corp was active in the commercial market in the 1990s but has been restricted from use by U.S. customers. In addition to domestic missions, China has launched a satellite for Nigeria and appears to be prepared to provide launch services for other foreign entities. Both Japan and India manufacture and operate launch vehicles but are currently focused on domestic missions as well. Table 10.2 lists various systems and compares their nominal characteristics.

10.3.1   Ariane

The European commercial expendable LV, Ariane, was developed by the ESA and the French Centre National d'Etudes Spatiales (CNES) and is now manufactured, operated, and marketed by a European company called Arianespace. For many years the basic vehicle was the Ariane 4, which contained first and second stages that used hydrazine–nitrogen tetroxide and a cryogenic third stage using liquid hydrogen-liquid oxygen. Solid and/or liquid strap-ons could be combined with the first stage. The strategy of Ariane was and is to provide a dual-launch capability to offer a more cost-effective service to commercial customers. The dual-launch capability was proven in U.S. government programs; however, it was applied for the first time in commercial service for Ariane 2. Another important feature of the program is the use of the Guiana Launch Center at Kourou, French Guiana, located at a latitude close to the equator. That reduces the inclination of the transfer orbit, which offers the capability to save fuel or extend life. The Ariane rocket is a derivative of earlier European rocket programs from the French, British, and German governments. On the other hand, the concept of a commercial company, that

Table 10.2   Characteristics of Typical Commercial Satellite Launch Systems for GEO Missions (Representative Information as of January 2008)

Launch Vehicle	Provider	Launch Site	Mass to GTO, kg	Year First Launched
Ariane 5	Arianespace (France)	Kourou, French Guiana	6,920	1996
H-2A	Rocket System Corporation (Japan)	Tanegashima Space Center, Japan	4,000	1994
Proton M	International Launch Services	Baikunor Cosmodrome, Kazakhstan	5,500	1998
Zenit	Sea Launch Co. (U.S.)	Pacific Ocean or Baikunor	5,000	1985
Atlas	Lockheed Martin	Kennedy Space Center	4,300	1993

is, Arianespace, was totally new and has become the pattern for several others to follow.

The principal commercial vehicle today is the Ariane 5 (see Figure 10.9), developed to support the Hermes reusable "spaceplane." The Ariane success record is among the best in the industry, having improved over the decades through extensive design refinements and improvements in operating procedures. The Ariane 5 rocket improves launch mass and volume in allowing two large satellites to be launched simultaneously.

In the Ariane 5, the cryogenic first stage has two large solid-propellant strap-ons and is used for all missions. Above it, the upper composite stage can support Earth and sun-synchronous missions to LEO or, using an attach fitting, missions to GTO. This stage burns MMH and N_2O_4 and is pressurized with helium. Dual- or triple-payload launches to GTO are accomplished with a fitting called the *structure porteuse externe lancements triples Ariane* (Speltra).

Arianespace has begun to offer the Russian Soyuz rocket using Kourou site. It is planned to have the capability to put 2,700 kg into GTO.

10.3.2 Atlas

Having been around for three decades, the Atlas Centaur has played an important role in lifting medium payloads into GTO and other orbits. The liquid-fueled rocket

Figure 10.9 Ariane 5 rocket at the launch of Horizons 2 (*Source:* Arianespace.)

was developed and originally manufactured by the Convair division of General Dynamics. That organization was acquired by Lockheed Martin in 1994, and Atlas continues to be available through Lockheed Martin Commercial Launch Services. The Atlas first stage, originally a ballistic missile, employs a liquid kerosene-liquid oxygen system. The second stage, called the Centaur, was the first high-energy cryogenic liquid hydrogen–liquid oxygen engine developed in the United States. Delivery to GTO is provided by the Centaur stage; the Atlas (acting effectively as one and a half stages) performs a portion of the injection into parking orbit. All launches into GTO are conducted from NASA's Kennedy Space Center (KSC) at Cape Canaveral, Florida.

The current offering of LMCLS consists of two versions of the Atlas V rocket, manufactured by the United Launch Alliance. The larger of the two configurations, the Atlas V 500, had a 5-m diameter payload fairing and has up to five solid rocket booster strap-ons. It employs a Russian RD 180 engine with two thrust chambers burning RP-1 propellant and liquid oxygen. The second stage is an improved Centaur configuration. Multiple standard payload adaptors are available to meet various mission requirements.

10.3.3   Delta

The Delta rocket family is offered to commercial customers by Boeing Launch Services Inc. Referred to as the workhorse of NASA, the Delta rocket has established an impressive track record since the first U.S. launches of geosynchronous satellites. It dates back to the beginning of the U.S. space program, when NASA Goddard Space Flight Center contracted with the then Douglas Aircraft Company for 12 complete rockets. While the first Delta flight in May 1960 was a failure, its eventual record proved to be the best in the industry. The Delta II employs up to nine solid rocket strap-ons to the side of the first stage for added life. To propel the spacecraft from LEO into GTO, a solid rocket motor called the payload assist module (PAM) can be provided. All launches to GTO are conducted from NASA's KSC facility. Non-GEO and polar missions can be initiated from Vandenberg Air Force Base in California.

The Boeing Delta IV rocket system is produced by the United Launch Alliance and is capable of launching payloads to polar orbits, sun-synchronous orbits, geosynchronous and geosynchronous transfer orbits (GTO), and low Earth orbit (LEO). Delta IV vehicles are planned for NOAA and NASA missions in coming years and can launch payloads weighing from 4,300 kg (9,480 pounds) to 12,980 kg (28,620 pounds) to GTO, and can lift over 23,000 kg (50,000 pounds) to LEO. This makes it suitable for nearly all of the large GEO satellites currently in production. The vehicle combines a first stage consisting of the Pratt & Whitney Rocketdyne RS-68 main engine and the Pratt & Whitney RL10B-2 upper stage engine. There are also two sizes of expanded fuel and oxidizer tanks, illustrated to the right in Figure 10.10.

10.3.4   H-1 and H-2

The Japanese government supports a domestic program for launching geosynchro-nous satellites. The Japanese Aerospace Exploration Agency (JAXA) is the govern-
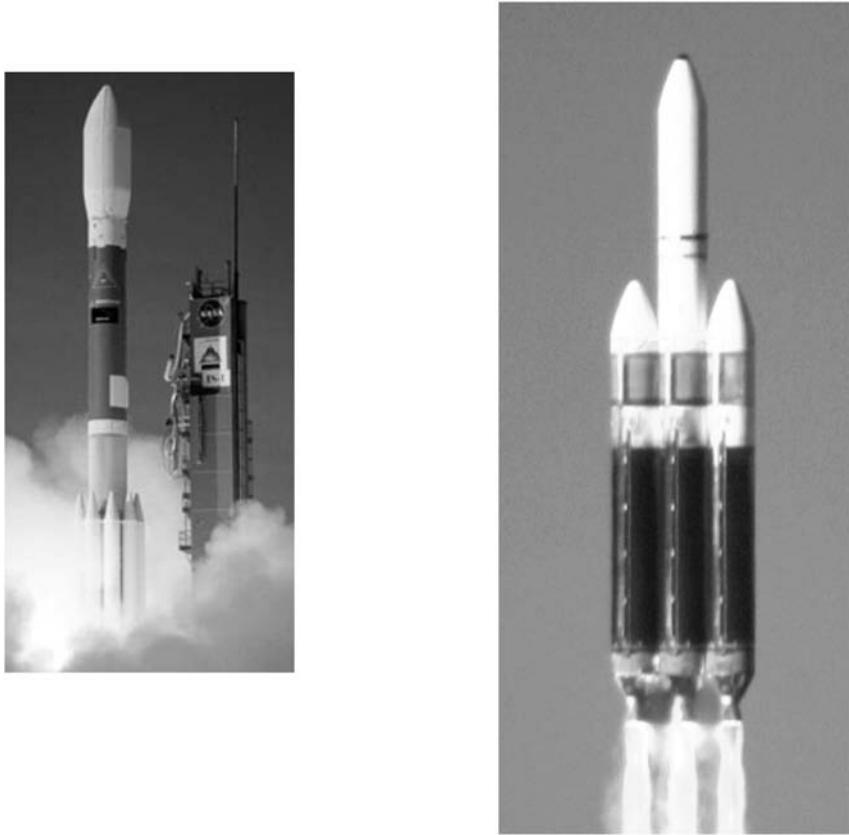
Figure 10.10 The Boeing Delta II and Delta IV Rockets. (*Source:* Boeing.)

ment agency in Japan responsible for launch and other space programs. The N-2 rocket demonstrated a perfect record for its nine launches. The first stage is a kerosene–liquid oxygen stage with solid strap-ons (similar to the Delta), while the second is a liquid hydrazine–nitrogen tetroxide rocket engine built in Japan. Perigee kick is provided by a Thiokol solid rocket motor. Launches take place on the Japanese island of Tanegashima. The upgraded H-1 rocket is built on the N-2/Delta design but with more Japanese technical and manufacturing content. Both the N-2 and H-1 ceased production at Mitsubishi Heavy Industries, to be replaced by the more powerful H-2.

The H-2A LV, the result of Japanese technology development, was built on the H-1 and designed to place larger payloads into LEO and GTO. The first stage employs cryogenic liquid hydrogen–liquid oxygen with a pair of rather large solid rocket motors strapped on the side, as shown in Figure 10.11. GTO is achieved by the cryogenic second stage through two burns of its engine. The first burn places the payload into a LEO parking orbit, and the second performs GTO injection.

10.3.5 Long March

With decades of experience in missile and LV development, the Chinese government has made available its Long March LV to the commercial market. The contracting

Figure 10.11   The H-11A launch vehicle. (*Source:* Japan Aerospace Exploration Agency.)

authority, China Great Wall Industrial Corporation (CGWC), has provided launch services for several satellite operators, both Chinese and foreign. Currently available rockets include the LM-3A, and LM-3B. Lift capability is in the low to middle range, making the Long March suitable for many commercial spacecraft.

In terms of rocket design, the first stage of the 3B employs four liquid engines using UDMH–N_2O_4 propellant. The 3B cryogenic third stage, using liquid hydrogen–liquid oxygen, can place payloads directly into GTO. The first foreign launch of the LM was for AsiaSat 1, an HS-376 satellite that was successfully placed into GTO in April 1990. Subsequently, the Chinese provided launches for larger satellites using the LM-2E and LM-3B series, the latter being the newest and more capable rocket. As discussed previously, the LM-3B includes a third stage of direct injection into GTO. All GEO missions are supported from Xichang because that site is relatively southern and allows overflight over relatively uninhabited Chinese terrain.

10.3.6   Proton

The Proton launch system, built by the Khrunichev State Research and Production Space Center, is the largest Russian launch vehicle in operational service. It has

been used in more than 325 launches and the booster has a reliability record of 96%. Proton launches all Russian geostationary and interplanetary missions under Khrunichev, establishing it as the principal workhorse of the Russian space program. International Launch Services (ILS), based in the United States, markets commercial missions on the Proton. ILS is a joint venture of Space Transport Inc. and Khrunichev. The ILS joint venture has launched over 40 Proton missions since 1996.

All Proton missions launch from the Baikonur Cosmodrome, which the Russian government leases from the Republic of Kazakhstan. The first test launch of the original two-stage Proton took place in July 1965, when it was used to launch the four Proton satellites for which the vehicle was named. Last flown in 1966, the two-stage Proton was succeeded by the three-stage Proton K and the four-stage Proton K/Block DM launch vehicles. The current version is the upgraded Proton M with the Breeze M upper stage. An Enhanced Proton with the capability to lift more than 6 metric tons was introduced in mid-2007.

The Proton first stage consists of a central assembly containing an adapter, oxidizer tank, tail section, and six lateral assemblies. All the lateral assemblies are similar in design. Each assembly includes a fore section, fuel tank, and tail section where the rocket engine is fixed. The first stage is powered by six RD-275 single-chamber cruise liquid propellant engines developed and manufactured by NPO Energomash. The propulsion unit develops 9,600 kN thrust at liftoff. The second, cylinder-shaped stage includes an adapter, propellant section, and tail section. The thruster facility of the second stage includes four similar sustainer liquid propellant engines: three RD-0210 engines and one RD-0211 engine—all built by KBKhA. The nose bay hosts six solid-propellant retro motors.

The rocket's third stage is a cylinder-shaped structure that includes an equipment bay, a section of fuel and oxidizer tanks, a sustainer engine fixed on the fuel tank, and an aft bay with a steering motor inside. The aft bay accommodates four solid-propellant retro motors. The propulsion unit of the RD-0212 third stage consists of an RD-0213 sustainer liquid propellant engine and a four-chamber RD-0214 steering engine. In terms of design and operating principle, the sustainer engine is similar to the RD-0211 engine of the second stage: in fact, the 0213 version is a modification of RD-0211. The space head (satellite, adapter, and fairing) is attached to the nose section of the third stage booster by the lower ring of the upper stage (structural interface). The flight of the first and second stages is controlled by gimballing main engines, while the third stage is controlled by gimballing four chambers of the steering thruster.

In an effort to enhance performance and reliability of missions to GTO, Khrunichev has developed, tested, and commissioned the Breeze M upper (fourth) stage. The compact design of the upper stage includes a core section surrounded by a toroidal auxiliary section of propellant tanks, which is jettisoned in flight following depletion. The 14D30 liquid propellant sustainer engine is mounted in a recess provided in the fuel tank of the core section. The engine can relight many times during flight. Low thrust liquid propellant engines use the same propellant as the main engine. These engines also control orientation and stabilization during unpowered flight and serve to settle the propellant in tanks as required for restarting the sustainer engine.

Installed in the equipment bay on top of the core section, the inertial control system is intended to control the Breeze M and its onboard systems during flight. The Breeze M upper stage accommodates the power supply system and avionics used for telemetry data transmission and trajectory measurements.

10.3.7   Zenit

Sea Launch Company was formed by Boeing Commercial Space Company in 1995 in partnership with Kvaerner of Norway, RSC Energia of Russia, and NPO-Yuzhnoye of Ukraine. This commercial venture provides launch services to GTO from a converted oil drilling platform that is towed to the equator in the Pacific Ocean. Kvaerner modified a drilling rig to use as the launch platform, constructed the assumbly and command ship, and now manages all maritime operations. The home port is in Long Beach, California, in close proximity to the first customer, Boeing Satellite Systems.

The Zenit-3SL liquid propellant rocket comprises the first two stages of Sea Launch. Produced by Yuzhnoye and first flown in 1985 from the Baikunor Cosmodrome in Kazakhstan, Zenit uses liquid oxygen and kerosene in both stages, the first containing four thrust chambers and gimbaled nozzles and the second containing a single main engine and another aft-mounted engine for three-axis control. The two-stage system is horizontally integrated, self-erecting, and self-fueling, which facilitates prelaunch operations onboard the ship and towed platform.

The third and final stage, used to propel the payload from LEO to GTO, is a modified version of the Block DM produced by RSC Energia of Moscow. The Block DM is a restartable upper stage using a liquid oxygen and kerosene main engine. A single gimbaled nozzle provides directional control. The vehicle is itself a spacecraft that provides its own three-axis control using a separate reaction control system consisting of five side-mounted engines.

The ship was designed and built in Glasgow, Scotland, in 1996 exclusively for this function. Besides providing the facilities necessary to process the LV and payload and to control the launch, the ship tracks the initial ascent of the LV and is the principle means of transportation from Long Beach to the aquatic launch site. The rocket is assembled and fueled on the ship, then transferred to the platform with a crane. Following launch, the rocket is tracked by another down-range ship and the Russian ground tracking stations.

The Zenit rocket is also offered for launch from the Baikunor Space Center in Kazakhastan through a facility known as Land Launch. Using existing Zenit infrastructure at Baikunor, the Land Launch system is based on a modified version of this proven Sea Launch vehicle. Land Launch's Zenit-3SLB vehicle addresses the launch needs of commercial satellites weighing up to three-and-a-half metric tones (approximately 7,700 pounds). The two-stage Zenit-2SLB is also available for lifting payloads up to 13 metric tons to inclined low Earth orbits. Sea Launch provides management functions and subcontracts all hardware and services to Moscow-based Space International Services, Ltd.

10.4   Launch Interfaces

From Section 10.3, it should be clear that there is a wide variety of LV arrangements and capabilities. That is reflected to some extent in the interface with the spacecraft to be launched. Many modern spacecraft designs incorporate features that permit the use of different LVs, allowing the satellite operator to obtain the best possible arrangement for launch and to deal with delays. Beyond the physical interfaces between the spacecraft and the LV are the management interfaces, which require a continuing process of coordination among the manufacturer of the spacecraft, the purchaser, and the launch service providers.

10.4.1   Physical Launch Interfaces

A communications spacecraft that is to be launched must be designed and tested in such a way that it can be properly attached to the LV. In addition, the spacecraft must withstand the expected acceleration, shock, and vibration of rocket flight. An important consideration is that the spacecraft fit within the dimensions of the LV. The electrical connections between spacecraft and LV, which must be severed at time of separation, are primarily for the purpose of activating separation and the spacecraft itself. In general, the spacecraft is self-powered by its internal batteries during ascent.

Figure 10.12 presents in simplified form the physical interface for two typical expendable LVs. The fairing is the outer shell of the LV that contains the spacecraft, constraining its height and width. In the illustration at the left, the spacecraft is shown attached to the perigee kick motor with an adapter ring. The use of an adapter ring allows the spacecraft and LV manufacturers to build their respective attachment systems ahead of time without having to bring the two parts together prior to integration at the launch site. Even with that technique, integration tests are still performed on the first sample of any new combination of spacecraft and LV.
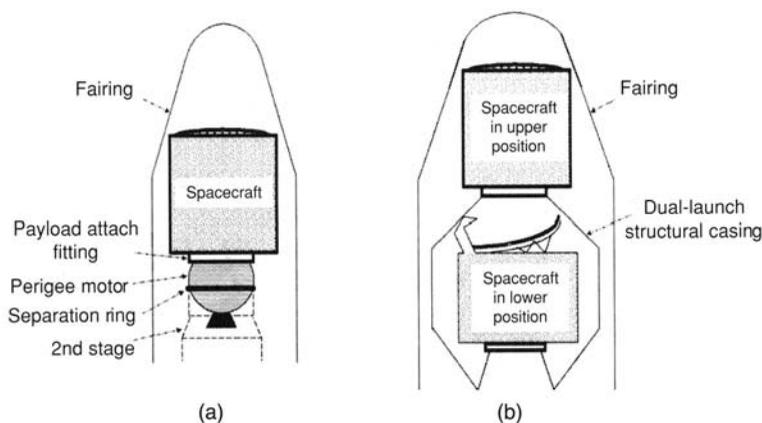
Figure 10.12   Payload integration arrangements for the (a) Delta and (b) Ariane expendable launch vehicles.

The interfaces in the Ariane are compatible with the dual launch feature. In Figure 10.12(b), one spacecraft is mounted on top of the casing, within which the second (and possibly third) spacecraft is contained. Once in transfer orbit, the casing frees each spacecraft for subsequent injection into GEO by their respective apogee engines. Arianespace uses an enlarged 4.5-m diameter fairing, which makes possible the launch of large spacecraft. The trend in other expendable LVs is now toward such enlarged fairings because of the flexibility afforded to the spacecraft buyer.

As discussed earlier in this chapter and illustrated in Figure 10.1, the spacecraft requires two major orbit changes to reach GEO from the parking orbit. The second change, at the apogee, usually is provided by a rocket engine internal to the spacecraft, such as a solid AKM or a high-thrust liquid engine. Perigee kick can be provided either by the third stage of the LV or by a separate rocket engine attached to the spacecraft. The spacecraft buyer does not need to be concerned with the perigee kick stage if its function is provided by the LV agency (as is the case with Ariane, Atlas-Centaur, Proton, and Zenit).

10.4.2   Management Interfaces

While not all-inclusive, this section provides a brief overview of the plans and activities carried out by the organizations involved with the launch of a commercial communications satellite. The first step in arranging a commercial LV is to enter into a launch services agreement (LSA) with a launch agency such as Arianespace, Boeing, Lockheed Martin, or ILS. Because the LSA is a legal document, it usually is prepared and negotiated like any other contract. In addition to laying out the schedule for launch and payments to the launch agency, the LSA delineates major responsibilities, including those shared by the buyer of the services, which could be either the satellite manufacturer or the intended operator of the satellite. There is a technical appendix to provide specifications for the rocket and its interfaces with the payload. One of the most important specifications is the maximum payload weight the LV is capable of placing into parking orbit or GTO, as appropriate for the particular LV.

While the technical appendix to the LSA covers physical interfaces and requirements, it still may be necessary and desirable to prepare a document after the launch agency is under contract to precisely define every aspect of the integration of the payload with the LV. Often, most of the information is provided by the spacecraft manufacturer since that organization is the most knowledgeable about the physical and electrical characteristics and constraints of the payload. However, all parties have a great deal at stake in the launch process and therefore need to take an active interest in the preparation of the management plan. From a contractual standpoint, the requirements of the plan need to stay within the limits of the LSA so the buyer is not subject to increased cost and to ensure that the capabilities of the LV are not exceeded.

While the technical specifications in the documents and agreements cover the physical interfaces, there is a vital need for a time schedule for all the activities leading to liftoff. The manufacture of the spacecraft and the LV are the responsibilities of the respective contractors, but delivery of the hardware from the factories

must be consistent with the overall project schedule. Prior to the start of launch site activities, compatibility tests and rehearsals usually are carried out to verify that things will fit together and that the launch itself will go smoothly. Those activities require management attention by the three key parties to the agreements.

A detailed schedule is generated for the activities at the launch site. The spacecraft contractor working at the site performs a series of checks and tests to prepare the spacecraft for integration with the LV. Meanwhile, the launch agency and the LV manufacturer assemble the LV and begin the process of payload integration. Most Western expendable LVs are integrated with their payloads vertically on the launch pads; Russian and Chinese LVs are integrated horizontally.

The program plan must deal with the availability of support services and facilities at the launch site since the spacecraft contractor usually will not have its own facility nearby. The launch agency that operates the site may provide those services as part of the cost of the launch itself; otherwise, a separate agreement must be entered into with whatever organization happens to have the ability to provide that support. Critical activities at the site include payload mating, electrical system checks, fueling of the spacecraft RCS, and the attachment of any external rocket motors.

While many of the responsibilities can be delegated to the spacecraft manufacturer, the ultimate responsibility for the mission direction lies with the buyer of the launch service, often the prospective satellite operator. The buyer's mission director, working for the launch services buyer, is positioned at the launch control center and is consulted by the launch agency and its launch team. The launch director works for the launch agency and commands the team that executes the actual launch. The launch director, however, will not agree to launch without the consent of the mission director.

Following liftoff and after the spacecraft separates from the LV, all control reverts to the buyer and the subcontractors. This is the transfer orbit phase (described in detail at the beginning of this chapter). As part of the program plan and preparation, the mission direction and TT&C support around the world should have been tested by rehearsals and simulations well ahead of the actual mission. It is also a good idea to have contingency plans to deal with the unexpected. For example, a TT&C site in a remote part of the world may experience an equipment failure just prior to the time of need. Contingency plans can be incorporated into the mission analysis performed by the orbital dynamics group so that tracking and commanding can be accomplished in a number of different ways. In any case, it is important to have the full operational and analytical capability available at all times during the initial phases of the mission.

10.5   Risk Management in Launch and Operation

Placement of communications satellites into service has some risk inherent to it. The odds are that a given launch will be successful, but it is not a surety. Therefore, satellite operators must plan for all reasonable contingencies so their customers can continue to obtain reliable communications services. We use the term *risk management* to include those steps taken during the design, manufacture, launch,

and initial operation of the spacecraft that make up the space segment. Risk management must be embedded in the operations of all program participants (spacecraft manufacturer, satellite operator, launch vehicle manufacturer, and launch service provider) and should not be considered a peripheral task. It would not be sufficient simply to cover the direct financial risk with an insurance policy, because that alone cannot ensure the service will be available (only that the cost of replacement is covered).

An important first step is to make sure that all spacecraft and, in the case of non-GEO systems, the constellation are properly designed from a system and reliability standpoint. As a rule, a spacecraft that comes from an established product line from one of the leading manufacturers probably is more inherently reliable than one that represents a unique design. This can be overcome through extensive design analyses and testing to uncover potential reliability shortcomings, but it comes at considerable extra expense. Careful consideration of redundancy and likely failure modes must be made early in the design of the spacecraft. The customer should maintain oversight during manufacture and test phases of the program. It is common for customer personnel to be located in the spacecraft manufacturer's factory to attend progress meetings, design reviews and to witness major subsystem and system tests. That instills in everyone the importance placed on reliability and risk management.

In summary, the risks that apply to commercial space systems relate to cost, schedule completion, launch, and on-orbit operation. We are particularly concerned here with launch and on-orbit risk considerations. This section provides some guidelines and examples of how risk management is conducted in the commercial satellite industry. We focus on the front end of the project, which ends when the satellite has successfully reached its operating orbit and entered service. From that point, there still are considerations of how satellites may need to be replaced due to normal or premature end of service. Our discussion is by no means exhaustive and should not be applied directly in cookbook fashion. Particular situations will differ widely, so a unique risk management strategy must be considered carefully and subsequently implemented. Help for doing that can be obtained from the relevant spacecraft manufacturer, insurance underwriters, and experienced industry consultants.

10.5.1　Launch Insurance

Commercial communications satellites typically are insured against loss due to launch failure. The launch is the part of the mission of highest risk to a specific satellite and therefore poses the greatest potential for financial loss to the satellite operator. Of course, loss of the satellite at any point in its life is a serious concern, particularly when high-valued services aleady are being rendered from the space segment. For the established LVs and launch service providers, the likelihood of successfully reaching the prescribed orbit and becoming operational is in the range of 90–95%. Stated another way, there is between a 1-in-10 and a 1-in-20 chance of launch failure. Launch insurance includes coverage of all steps, plus some period of initial operation, usually 6 months.

Satellite projects that depend on commercial sources of funding generally must use launch insurance to satisfy the demands of investors and banking institutions. In certain cases, some large organizations have self-insured by taking account of the odds and holding sufficient reserves, either in terms of extra spacecraft or a financial pool of money. The question remains as to the amount of coverage to purchase and the likely premium cost. By practice, owners of GEO spacecraft typically insure the total value of the spacecraft and the launch service, as well as the cost of the insurance itself. The value assumed should be based on the current replacement price, which can be called out in the original contract as an option. Insurance premiums are in the range of 12–15% during normal times but can increase to 20% or more after one or more launch failures.

Launch insurance is purchased through a broker or an agent, who actually obtains the coverage on the commercial high-risk insurance market. In the past, Lloyds of London was famous for being the house that places this type of coverage, but it largely has been replaced by a multiplicity of specialists in the United States, Europe, and Asia. It is the job of the broker to work with the insured to fix the coverage requirements and desired costs. From that point, the broker goes out into the market and assembles a syndicate of insurers to take on the risk. The typical insurance policy assumes risk of loss at the time of intentional ignition of the LV and terminates 6 to 12 months later. Readers should be aware that the precise definition of intentional ignition depends on the particular spacecraft and LV design and should be specified before going to the insurance market.

Many insurance brokers and underwriters have technical experts on staff to review the reliability history and performance of the LV and the spacecraft. They may have come from the same manufacturers that provide the hardware and services, so they tend to know what questions to ask and what to look for. Insurance rates can depend on the particular design and manufacturer and fluctuate depending on recent experience. The insured party can place the policy well in advance but may not be able to lock the rate until some months ahead of launch. All those matters require careful consideration because of the large sums of money involved.

An approach taken by some operators is to contract with the spacecraft manufacturer for delivery in orbit (DIO). In that arrangement, the manufacturer is responsible for procuring the LV, managing all aspects of its use, and possibly purchasing the insurance as well. On the one hand, that simplifies things for the operator but, on the other hand, takes certain matters away from direct control. It is a tradeoff that needs to be considered; it can be advantageous for some (but not all) satellite operators. Some spacecraft manufacturers (and operators, as well) avail themselves of good deals for launch vehicles through the facility of the long-term agreement (LTA) with the launch services provider. That assures them of a supply at an agreed price per launch in exchange for guaranteeing a minimum buy quantity. For the DIO type of contract, the LTA can reduce the risk of having another launch available to get a replacement up.

The other, more common approach is to purchase the spacecraft for delivery on the ground (DOG), thereby placing responsibility for obtaining launch service and insurance on the operator's shoulders. The buyer must have adequate technical and administrative resources to manage the launch services provider and insurance broker. Operators that do this generally have a need for many launches over time

and so can justify the expense and the complexity. There are, of course, hybrid approaches in which the DIO contract does not include the launch insurance.

Satellite operators may purchase another type of coverage, called life insurance, for the period of actual operation. The cost is more nominal, being in the range of 1–2% of the insured value per year. At the beginning of a satellite's life, the covered value is essentially the same as for the launch insurance. That coverage can be reduced as the satellite ages, to account for depreciation. At the satellite's end of life, life insurance is theoretically unnecessary because a new satellite normally is made available through a new contract.

10.5.2 Backup and Replacement Satellites

We have spoken about a satellite launch as a discrete event taken out of a general context. In reality, the satellite operator has to consider not just one but many satellites. A system that depends entirely on the successful launch of one spacecraft is a system that is taking a decided gamble. However, commercial business and government operations as well need to view their position as that of absorbing risk so that the ultimate users are not themselves having to gamble on a successful launch. (In spite of that, using organizations may find it prudent to arrange for alternative capacity from one or another satellite system.) The fundamental way to improve the probability of success is through diversity, which simply means the satellite operator needs to have multiple ways to deliver service besides the spacecraft involved in the specific launch of the moment. That is easier for established operators with many working satellites and more difficult (but not impossible) for newcomers to the business.

Diversity can represent a high barrier in financial and operational terms and is a discriminator in the marketplace. There obviously are many approaches and options for providing diversity, and there are price tags to go with them. The most basic package is to purchase, along with the spacecraft, either a second backup spacecraft or major portions of one. Then, in the event of launch failure, the second spacecraft can be readied (or completed, as the case may be) for a second launch. The LV also must be made available for the second try, which is something that can be accommodated in the LSA. If the first launch is successful, the backup can be placed in storage until it is needed to expand capacity or deal with an on-orbit failure.

The use of the second and subsequent spacecraft for expansion also ties in with providing on-orbit diversity. That is the most common approach to both developing a growing business and reducing risk for the operator and its customers. As a general practice, operators tend to utilize all their on-orbit resources to provide revenue-producing service. However, there could be excess reserved capacity for use for restoration of primary services in the event of launch or on-orbit failure. Then the operator can use the insurance proceeds to initiate construction of a replacement on a more methodical (and leisurely) basis. The objective is to better manage risk so that operations and finances are evened out over the long term.

This section has been only a brief introduction to the subject of risk management. No one can determine the particular risk environment and remedies better than the satellite operator itself. Individual situations differ greatly and depend on

many factors, some of which are under the operator's control and others that are not. For example, the approach taken by an established operator with many functioning satellites already in service is much different from that of a new operator just starting out. Also, a different approach must be taken by a private company that has investors or public stockholders, who must be protected from financial downside losses, in constrast to a government organization that may be more concerned about the risk of not being able to service a vital need. Another important point is that all LVs are not the same because their reliability records differ, both in absolute terms and over time. Also, what seems to be a simple change in a rocket configuration (such as adding a different stage or changing some design feature) may have a significant impact on launch success. Those issues demand the greatest attention of all levels of management. It is well worth the time, effort, and resources to find and maintain the most effective risk management strategy.

Satellite Operations and Organization

There is a saying that a chain is no stronger than its weakest link. That is certainly true when it comes to ground control operations of a satellite communications system. This chapter is a discussion of the overall operation of the ground segment, utilizing both technical and human resources. Vital to the ground segment is a dependable satellite control center (discussed in Chapter 9) that provides appropriate and adequate contact with all working satellites, allowing a trained and experienced staff of professionals to perform the tasks necessary to achieve overall performance requirements. In our experience, that requires considerable investment in quality hardware, software, and human capital. What good satellite operation strives for is zero defects in terms of all satellite control functions. In practice, problems occur and mistakes, unavoidably, are made.

Satisfactory performance of the satellite control system depends on how well the operator deals with contingencies. We previously dealt with the physical hardware of the space and ground segments. Those elements must be designed and implemented to provide a working system. Keeping the system in operation and delivering services is the challenge of the soft side of the business—the assignment and management of the staff of professionals who attend to the immediate and long-term needs of the hardware. To that is added the users and the customers who access the space segment; they are partners in the important process of maintaining reliable and effective services. We provide this review to give the reader an appreciation of the requirements and issues that impinge on the quality and reliability of an operation. This chapter should not be taken as a comprehensive set of recommendations but rather as a brief introduction to the topic and a starting point for more thorough planning and development efforts.

11.1 The Satellite Control System

An overall satellite control system can take many forms, corresponding to the requirements of the satellite constellation, whether GEO or non-GEO. GEO satellite operation principles are well developed, having been initiated for the INTELSAT system and later employed in domestic satellite systems, first in the western hemisphere and later in the rest of the world. The approach was reviewed in Chapter 9 for typical TT&C ground facilities comprising one or more TT&C ground stations and an SCC. Figure 11.1 is an overview of that type of arrangement.

The equipment installed in the TT&C stations and the SCC (illustrated in Figure 9.19) satisfies the basic requirements of satellite monitoring through the
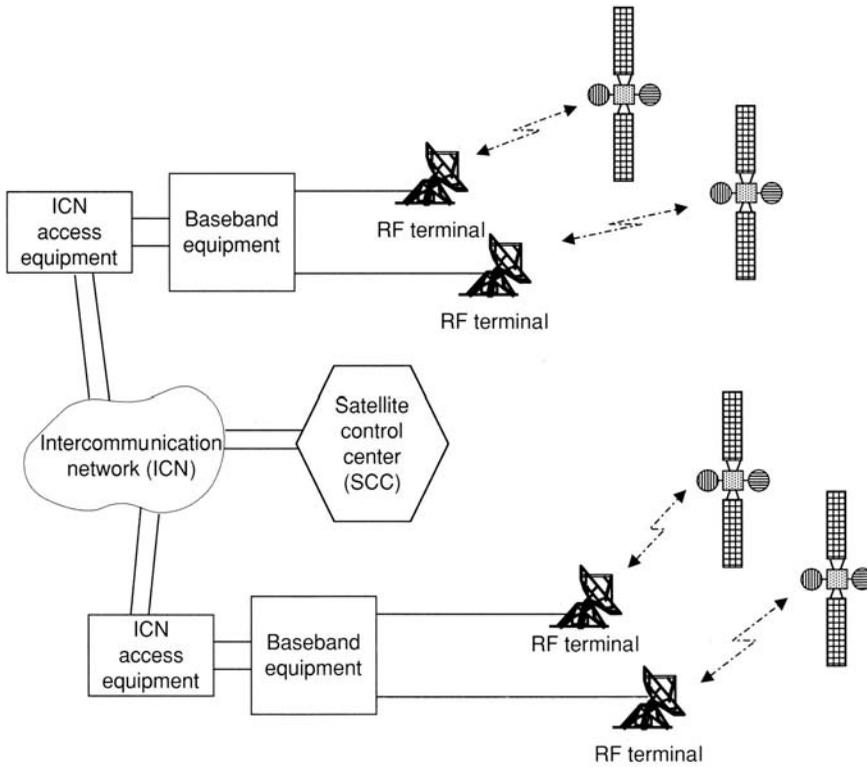
Figure 11.1 A typical TT&C ground segment, using two TT&C Earth stations and a centralized SCC.

telemetry downlink, control through the command uplink, and orbit correction, ranging and orbit determination functions. Those are the most basic technical requirements of satellite control and must be dependably performed over the satellite mission and lifetime of the system. TT&C system equipment and software may be purchased along with the spacecraft themselves and tested prior to and just preceding launch. Over time, the equipment must be properly maintained and upgraded when necessary to expand capacity or improve the quality of operation. Likewise, software is debugged and enhanced throughout the operative lifetime. The difficulty of making such changes should not be underestimated, because of the problems associated with working on a "live" system in a way that will not disrupt the ongoing operation and the ground control staff.

The investments in facilities and personnel have risen with the quantity of satellites under the control of a particular operator. This has encouraged the largest operators like Intelsat and SES to become efficient, thereby reducing the cost of operating one particular satellite. As a result, a new satellite operator with only a few satellites could find it attractive to outsource the satellite operation activity to one of the larger, more efficient satellite operators. Those who have gone this particular route have found it helps improve their focus on the business they are in and obtain the experience of a company that has operated satellites for a long time. This would require that the outsourcing satellite owner develop a detailed

service-level agreement (SLA) with the large operator to address the necessary satellite operation services, actions to be taken under contingencies, and appropriate penalties and rewards that apply.

11.1.1   GEO Satellite Control

The first step in gaining an understanding of GEO satellite control is to identify the control requirements. They can be summarized as follows for principal aspects of satellite operation with the exception of the monitoring of payload services, which is discussed in Section 11.3.3.

- *Support the initial get-to-station activities following launch*. The spacecraft manufacturer usually is responsible for carrying out this phase of the mission but might need to use the TT&C ground stations and SCC of the satellite operator in either a primary or a supporting role. Operations personnel often benefit from being part of the transfer-orbit team, since they ultimately will take control and responsibility. Also, it is an excellent time to verify and validate the function and performance of new or modified hardware and the software elements of the ground system.
- *Conduct or actively monitor initial on-orbit testing*. This can follow the same line of thinking as for the get-to-station activities. Since there is a heavy emphasis on the long-life capability of the satellite following testing, the satellite operator takes a stronger and possibly leading role in the conduct of these functions. Handover of the satellite from the spacecraft manufacturer to the satellite operator occurs when both parties agree that the system meets contractual requirements. If that is not satisfied, it still might be possible to commence service, with the manufacturer retaining responsibility to complete any open action-items (not unlike the "punch list" common in the purchase of a newly constructed house).
- *Maintenance of the satellite in its assigned orbit position with sufficient accuracy to support the ground network*. The common practice is to maintain orbit position to within ±0.1 degree north-south and east-west (or, stated another way, within a square of 0.2 degree on each side and centered on the assigned location). That generally will meet agreements to control adjacent satellite interference and is within the technical capability of TT&C ground and spacecraft systems. If multiple satellites are to coexist at the same orbit slot (something that is common and can now be viewed as a requirement), accuracy must be improved by a factor of 2 at least. Individual colocated satellites can be precisely located from the ground and their position "choreographed" to preclude collision (there is only a small probability of collision in the absence of such active avoidance, but satellite operators prefer not to have to worry about that potential for loss). During all these activities and over the life of the satellite, fuel usage must be carefully controlled and budgeted so life expectancy is achieved.
- *Monitoring of satellite health data*. The SCC is provided with facilities for receiving, displaying, recording, and analyzing spacecraft health data provided by the onboard telemetry system. The practice in commercial GEO

operations is to perform those functions continuously on a 24-hour basis and without interruption. The reason is that GEO satellites are also expected to provide uninterrupted service to users, so their operators, of necessity, feel compelled to know what is going on with their in-orbit assets. Any difficulty with service can then be traced back to telemetry data and analysis performed by computers and spacecraft technical experts. Also, the data permit a long-term evaluation of satellite performance and the identification of potential failure modes before they occur.

- *Control of spacecraft subsystems*. Modern GEO spacecraft contain sophisticated onboard computers, which make their operation nearly autonomous. While that is a desirable objective in the long term, there are many circumstances when human intervention is required to ensure a safe operation and proper service performance. Most stationkeeping actions are initiated and conducted from the ground, using the SCC hardware and software systems. Satellite operations often become more active and even stressful during expect eclipse seasons. Battery discharge and recharge must be actively monitored at the least. Response to anomalies and faults onboard the satellite almost always will be through human intervention via the command link. For those reasons, the SCC must be ready at any time of night or day to take control, either by plan or in the event of an unexpected contingency situation. Routine changes in the configuration of the payload or bus will also involve the SCC.
- *Operation and maintenance of the TT&C Earth station and SCC system*. The previous requirements relate to the proper operation of the satellites in GEO. In this paragraph, we are concerned with the like performance of the ground system itself. The ground facilities are just as essential to reliable service provision as any of the bus or payload elements. For example, if the TT&C antenna were to move off the satellite due to a failure of its tracking system, both the command and telemetry links would be cut off. The satellite would then have to function on its own, a situation that would have been considered in its design and that ordinarily would not produce a loss of service. However, the number of possible situations that the system could encounter cannot be anticipated in practice, so the best philosophy is to take the steps to ensure as close to 100% availability of telemetry reception and command transmission as is technically and humanly possible. An approach that improves on that situation is to have a fully functioning backup TT&C station (see Figure 11.1).
- *Ensure that satellites can be replaced or relocated as demanded by the service plan*. Satellite replacement is often an important responsibility of the satellite operator, since users prefer to view their on-orbit repeater as a permanent feature of the network. Just as with launch risk management, there are complexities in retiring old satellites and providing a smooth transition to the new on-orbit resources. For example, a new satellite may have to be configured and tested in an unoccupied orbit slot, then drifted into place before transferring traffic to it.
- *Develop and maintain contingency plans*. The satellite operator and user need to recognize the true meaning of Murphy's Law: "Anything that can

possibly go wrong, will (go wrong)." (It also has been said that Murphy was an optimist.) Engineering and maintenance must identify any weak links and failure modes and create backup and replacement strategies. At a high level, the satellite operator needs a set of plans that identify how the organization would react to a partial or full satellite failure. Any major user of GEO satellite capacity, such as a TV network, should have alternate or diverse resources to maintain transmission in the event of a failure. Through attention and commitment to continuous service, organizations may stay ahead of Murphy.

The requirements and considerations described here are based on experience and should provide some guidelines for GEO operations in general. For a specific GEO satellite and system, operations managers should conduct their own evaluation and produce an operations plan appropriate to their particular needs. Many of the issues concerning human resources for the operation are covered in Section 11.4.

11.1.2 Non-GEO Satellite Control

Non-GEO systems employ centralized satellite control with TT&C stations at critical points around the world that transfer the data to an SCC located at or near the general headquarters of the satellite operator. (INTELSAT has used this concept since it came into being.) The non-GEO approach has unique features, particularly that service availability is dependent on having enough satellites to afford adequate view by users around the globe. The following list summarizes some of the key ground support requirements for a non-GEO constellation, based on the requirements of a typical overall system [1].

- *Establishment of the constellation.* The satellite orbits are determined primarily by the launch vehicles, with limited ability of the spacecraft propulsion system to make the kinds of significant orbit adjustments common in GEO systems.
- *Maintenance of an adequate satellite complement.* A given satellite could be removed or replaced at any time without significant impact, provided coverage remains available from the overall constellation. Also, there is likely to be a need to maintain a continuous flow of replacement launches over the system lifetime, something that is much less frequent for the typical GEO system.
- *Assurance of continuous service through coverage.* The ability of the constellation to provide continuous coverage to users depends on the architecture of the system. An approach based on the bent-pipe type of repeater and without the benefit of intersatellite links must rely on simultaneous linkage between remote user and hub station (or between pairs of remote users for a mesh architecture). For such a system, sufficient hardware must be maintained in orbit to ensure that at least one working satellite has simultaneous view. Diversity operation, provided by Globalstar, is based on at least two satellites. The situation for a network like Iridium, with intersatellite links and without a requirement for national gateways, is that each user

have a clear, unobstructed view of one satellite (and not necessarily the same satellite). The MEO-type constellation reduces some of that difficulty because each satellite covers a much wider piece of the Earth's surface and remains in view for much longer periods than do the LEO satellites.

- *Non-GEO satellite monitoring and control philosophy.* As introduced in Chapter 9, the operator of a non-GEO constellation might not have a continuous link from the SCC to every satellite. (As will be discussed in Section 11.2, the system using intersatellite links like Iridium may allow the SCC that 100% availability.) Without continuous availability, telemetry could be collected once per orbit using a few well-located TT&C stations [2]. Control of the satellite could take place as needed, either simultaneously with monitoring or during a subsequent period of contact. There is the obvious impact on spacecraft design that the system must operate autonomously and record relevant data between downloads to visible TT&C ground stations. That general approach of onboard data storage helps reduce the number of TT&C stations. Also, because the satellites presumably are of the same design, the system need only keep track of unique spacecraft signatures and hardware status through a well-structured database.

- *Simplified on-orbit control.* Considering the preceding factors, the bus system of a non-GEO satellite might be simpler than its GEO counterpart. That is in keeping with a desire to reduce the parts count, cost, and complexities involved with satellite operation. One would hope that such measures do not reduce reliability or hamper proper identification and correction of faults (detected or otherwise).

- *Replenishment of satellites.* Operators of LEO constellations have built their strategies on expecting 5–8 years of satellite life in orbit. One limitation on lifetime is a need to upgrade and improve the design of the initial constellation. Production and launch schedules involve 1 to 2 years to achieve the full complement of satellites. Considering that there will be launch or on-orbit failures, a LEO operator must continue to construct and launch spacecraft throughout the operating lifetime of the system.

The obvious and subtle differences among non-GEO constellations and networks make it extremely hard to generalize on the requirements for TT&C ground operations. Each system must develop its own unique operation approach. However, the guidelines presented here should give the reader a basic idea of the factors involved in this type of effort.

11.2   Intercommunication Networks

The majority of satellite operations depend on the reliable transfer of telemetry and command data between the actual TT&C Earth stations that are in direct contact with the orbiting satellites and the SCC, where the full range of technical and management functions are performed. Those links, which can employ both terrestrial and satellite networks, must satisfy the data throughput and quality-of-service (QoS) objectives of the overall satellite control system. Included in QoS

objectives are such measures as service availability (in percentage of time), error rate performance, delay or latency (as well as the variation over time), lost packet ratio, and call setup delay (for connection-oriented services like ISDN). There has been renewed interest in QoS of commercially available communication services, since they affect how well the network meets (on a continuous basis) the overall needs of the system and the users. Modern fiber optic systems and digital satellite communication exist on a worldwide basis, so it is no longer a major challenge. The difficulty now posed is how to interconnect the elements of the network into a reliable fabric that is manageable by the satellite operator under all expected conditions. To that end, we provide a brief introduction to the subject.

11.2.1   Backbone Communications

The fundamental matter for the intercommunication network (ICN) is to connect the various sites with sufficient bandwidth to meet the needs of the SCC. That is the purpose of the backbone portion, suggested in Figure 11.2. Locations usually are determined by the basic design of the TT&C system, considering satellite locations and orbits, the requirements for maintaining continuity of service, and the needed support for the personnel who work throughout the system. All those factors have to be taken into account, which often requires several iterations during the development period as well as after the system begins service. That is a natural consequence of the learning process that goes on in the organization charged with managing the satellites and the services they provide. Also, the particular mix of
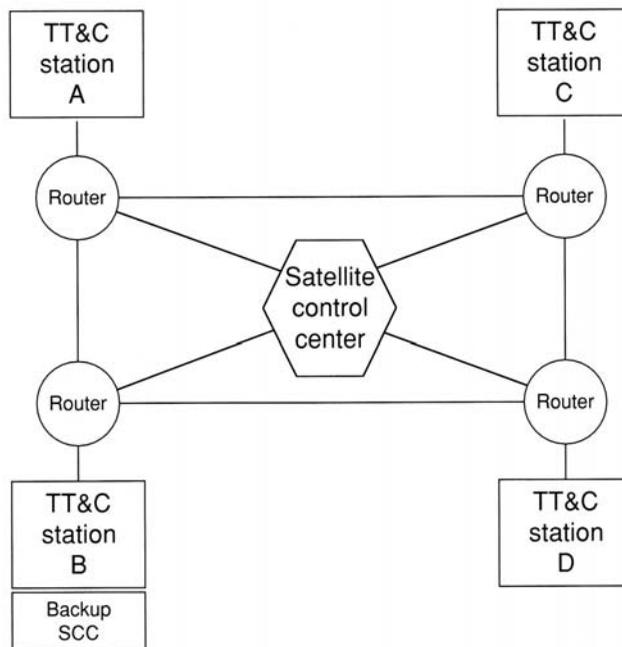
Figure 11.2   A fully redundant communication network for service to a TT&C system composed of four TT&C Earth stations and an SCC. Links between locations may be provided using leased fiber circuits, microwave, or independent satellite services.

telecommunications technologies will change over time—what was once state of the art could, within even a few years, become obsolete and possibly taken off the market! However, this is probably not one area where public Internet service can suffice. Even when an appropriate technology is located and vetted for use, it may not be available at all locations or with a common feature set. That is, after all, one important reason why satellite communications itself is of benefit to telecommunications users.

The designer of the backbone, therefore, must determine which telecommunications systems and services are available for application in the ICN. The ideal solution for some is to select and contract with a single service provider, likely to be an international carrier like AT&T, Verizon Business, British Telecom, Cable & Wireless, France Telecom, KDD, or the like. In reality, whichever carrier is selected must work to pull the resources of multiple foreign service providers together to provide the backbone solution. Satellite operators in many cases prefer to carry out that task themselves, using their own qualified telecommunication staff.

The technologies that are most attractive for the backbone are no different from the hottest products and services on the world market. They include user-owned equipment such as IP routers, intelligent multiplexers, and ATM switches for WANs. Interconnecting those elements are backbone links over terrestrial fiber links, undersea fiber cables, and satellite links from GEO satellites, provided that those satellites are separate from those being controlled.

Access to the backbone is another concern, because that often is the cause of service and reliability problems. The communication requirements generally can justify installing fiber between the SCC and the public network point of presence (POP). Where that is not available as a service, the satellite operator may have to pay for construction or even install and operate the equipment themselves. Similar considerations apply to the remote TT&C stations, wherever they may be located. As we have stated, that becomes troublesome where facilities are located in regions without good terrestrial communications. Satellite links can be employed, provided one does not rely on the same satellite being controlled (to do so could increase the difficulty of responding to a satellite problem that affects that link as well). Backup service through Inmarsat or Iridium might be a good insurance policy.

11.2.2 Alternate Routing for High Reliability

It is our general philosophy that a reliable telecommunications network can be achieved only through redundancy and the provision of alternate means of transferring the needed TT&C data. The general problems with achieving that in practice are cost and local availabilty. Many satellite operators find that the budget they can allocate to their ICN is fairly small, suffering from the problem that it derives no direct revenue. (Consider what would happen to revenue, however, if the network were to have a hard and lasting failure.) How does an organization justify the rather high expense and management challenge of developing the needed reliability?

In our experience, the telecommunication manager for a satellite operator must create a vision for the organization and develop a strategy for meeting the need. This person must become very familiar with the business and technical strategy of the satellite operator and be sure to achieve as close a match as possible. Through an

interactive process of listening carefully to other satellite operations and engineering personnel, communicating the options, and working out the solutions, the telecommunications manager becomes the solutioneer for the whole organization. He or she learns to recognize the specific needs of each site and department, and responds proactively to calls for help. That methodology works whether we are talking about the design phase (when the ICN as well as the TT&C system itself are still on paper) or through implementation and during the operating life of the satellites.

The power of diverse routing has been demonstrated time and again. Primary links cannot be depended on for 100% availability on their own. An alternate path could follow a reverse direction around a ring, as illustrated in Figure 11.2, or might be connected on demand. The latter could be available in some regions, in which case it can be employed for a reasonable cost. An important feature of routers is that they may automatically engage the working path. Installing dedicated alternate paths, on the other hand, almost always is costly. A rule of thumb is that if an alternate path already exists or might be included at little or no direct cost, it should be maintained in service. It would be embarrassing, to say the least, if a desperately needed alternate path had been removed from service only because someone thought it had no current use. That may seem far fetched, but it is all too common in practice. Also, the best way to ensure that a redundant path is working is to pass a portion of the ICN data over it at all times.

11.2.3 Network Management

The ICN is a distributed facility, with potentially many remote locations. As a result, it, too, requires an effective system to monitor performance and respond to problems when they arise. Fortunately, there is a good selection of specialized hardware and software on the market to greatly simplify the task. Underlying that is an open communication protocol, called the Simple Network Management Protocol (SNMP), developed for the Internet and already applied in a majority of large corporate WANs and intranets. SNMP provides the communications services that deliver network management data from remote communications devices (the routers, switches, and servers in the network) to a centralized network management station (NMS).

There are a number of specialized NMS solutions on the market that support a wide variety of proprietary telecommunications and computer networking products, including older designs no longer offered for sale. That may be appropriate when an existing system must be upgraded for centralized control without the benefit of SNMP. Also, NASA already uses its own system for satellite control and network management, which could impose a constraint on commercial users as well. Some specialist software companies offer NMS platforms that support SNMP as well as a number of proprietary network management approaches, particularly in the computer and telecommunications areas. Such tools that focus on Earth station resources are referred to as monitor and control systems.

11.3 Network Operations

Sections 11.1 and 11.2 dealt with background activities that facilitate the reliable operation of the satellites in orbit. From the users' perspective, those functions

should be invisible and, for all intents and purposes, perfect. They more likely will interact with the network operations element of the space segment, which is established to coordinate the actual communications services. Users make contact with network operations to establish connections, obtain assistance and information during routine activities, and seek assistance when difficulties arise. The last function is perhaps the most critical because network operations is the one place where users can have their problems addressed and is the first line of defense for the satellite operator when exercising control over the ground segment.

From a general perspective, the design and management of the network operations function needs the highest level of attention from the satellite operator. Every satellite system will, at one time or another, depend heavily on that activity to bring new services into being and address critical problems that are being experienced by current users. That cannot be done if network operations is treated as an additional duty by otherwise overworked satellite control personnel or inexperienced staff who lack the training and technical resources they need.

11.3.1 Standard GEO Transponder Services

Network operations for conventional GEO satellites with bent-pipe transponders are the most developed and, in many ways, the most straightforward with which to deal. That is because every transmission to each satellite can, in principle, be monitored at one or more locations on the ground. An area-coverage type of beam is the most convenient because, wherever the satellite operator is located, it is possible to monitor the full downlink spectrum from the GEO satellites under its control. We consider later the case of monitoring satellites that are outside the local reach of the operator, such as GEO satellites in another region or non-GEO satellites during the majority of time they are out of view.

Network operations in this environment can be routine but quite busy during periods of heavy customer traffic (e.g., prior to the news hour or during a major event such as the U.S. presidential campaign, the Olympic Games, or the World Cup). The major functions that one would expect to address are reviewed in the following tabulation.

- *Development of requirements for the network operations facilities*. This is where the manager of network operators works with a technical and marketing/sales staff (who are involved with defining and coordinating the actual services) to produce the technical requirements for the facilities that support the users. Besides the basic monitoring equipment, another necessary facility is a computerized database that contains contact information for users as well as scheduling information for daily transponder activity. The best way to develop a service strategy is for these players to work as a team. The people who market and coordinate the services work with the network control personnel to resolve the specifics of how users will make contact and ultimately are served as quickly and as professionally as possible.
- *Qualification of user Earth stations for access to the space segment*. In historical terms, the procedures for technically qualifying Earth stations for their first access were created for the INTELSAT system. The requirements

usually are attached to the satellite service contract with users and should be as effective as the force of law. User Earth stations must be willing to take direction from network operations, both during testing and later when services are being rendered. The procedures cover every important aspect of Earth station performance and service capability. Over time, the procedures to qualify new Earth stations have been streamlined; however, the critical need is that any new Earth station meet the minimum technical requirements and therefore not introduce difficulties into the current satellite as well as its neighbors. An example of a typical set of qualification requirements is provided in Table 11.1.

The extent of Earth station qualification validation depends on the degree of harm that a maladjusted station could cause. Hence, large Earth stations (such as those used for DTH uplinks and major hubs) are thoroughly tested. The trend for small Earth stations is to try to validate their performance on a network basis. For a VSAT, the antenna pattern and HPA power likely would be measured on one unit (a process called type acceptance). But frequencies, timing and protocol, and control-associated parameters would not need to be verified for each terminal individually over the satellites. For an Imarsat mobile laptop terminal, there is a type acceptance (which may presume certain tests at the manufacturer) so the user can take it out of the box and begin communicating in a half hour. The same can be said of hand-held MSS terminals.

At the extreme are RO terminals, as for DTH receive systems. For those, the validation process verifies compatibility with conditional access systems and assurance of consumer standards. It is common for the operator to specify every detail of the design down to the chip level to assure service consistency.

- *Monitoring of all transmissions to and from the satellites.* This is accomplished with a combination of automated monitoring and test equipment along with manual systems that can be used by network operations personnel, who are available on a 24-hour basis. The automated equipment would be

Table 11.1   Typical Qualification Test Requirements to Verify Earth Station's Technical Performance

Antenna pattern, sidelobe levels
Cross-polarization isolation
EIRP
G/T
Carrier frequency
Modulation bandwidth
Burst timing stability (if needed)
BER performance (if needed)
Channel amplitude response
Channel group delay response (if needed)
Threshold performance
Frequency stability
Operator control
Contact information

connected to sufficient downlink antennas to view all the operating transponders on the satellites and deliver the data to computers and video screens in the network operations center. A photograph of a typical facility is provided in Figure 11.3. Special software can measure and record carrier power levels, frequencies, and bandwidths and thus track the actual occupancy over time. Those data can be compared with what is recorded in the database of authorized services and users and can be an input for billing, if necessary. Operators can review that information and act on irregularities that indicate that either the Earth station(s) or the satellite transponder(s) are not working within prescribed limits. Ideally, this approach allows the satellite operator to identify and respond to problems before customers call in. In addition, monitoring software can tie in with the commercial management of services, in particular billing and customer account maintenance.

- *Telephone, fax, e-mail and other facilities to permit direct and effective communication with users who have access to the space segment.* There is no substitute for reliable direct communication between users and the network operations function of the satellite operator. With all the potential ways to do that, experience has shown that the telephone is usually the most effective communication tool. (Some systems require that all uplinks use a standard satellite telephone integrated into the operator's network.) The service agreement contract should require that all users provide a working fixed line or mobile telephone number where network operations can directly contact the uplink. That is more complicated in the case of VSAT networks, where individual terminals are usually unmanned. In this case, the VSAT network operator provides the control function and point of contact with network operations. Written or record communication is needed to provide specific instructions to users, particularly when RF interference or other unidentified problems arise. That often will amount to a broadcast message to all users and potentially other satellite operators as well. Network operations could then dispatch a repair person, if needed.

Figure 11.3 Automated test facilities for GEO satellite transponder management.

- *Good and effective internal communication with satellite control and sales/
  marketing.* Network operations is in direct contact with users, who are
  actual customers or their Earth station operators. Therefore, they should be
  treated with care. That can be done most effectively through a good working
  relationship among the other key activities of the satellite operator: satellite
  control, where spacecraft problems often are identified first, and sales/
  marketing, where user services are put into place and customer relationships
  maintained. Teamwork and cooperation among those internal organizations
  are vital to the success of the business and then the provision of reliable
  communication services, as well. Frequent formal and informal contact is
  highly recommended. In instances where those functions are physically sepa-
  rated, there is a need for effective use of the ICN and face-to-face meetings,
  when appropriate. There is another opportunity to involve an important
  outside constituency: the users or customers. A technique that has worked
  is to conduct user conferences and seminars and to create a satellite user
  group (SUG), which is similar to the user groups that associate themselves
  with particular types of computers or networks.

This brief summary of the requirements and considerations for the network
operation function is not exhaustive and should not be applied directly in specific
situations. The best approach is for the managers of the particular satellite operator
to develop their own network operators' organization and technical strategy and
then to implement that strategy in the particular context at hand. The size and
composition of network operations will change over time as communications traffic
grows and services evolve.

11.3.2   User Network Monitor and Control

The overall network operations function of the satellite operator may include the
direct control of Earth stations in the user network. That, of course, is a very
different kind of activity from qualifying Earth stations, monitoring transponder
signals, and dealing with RFI problems (although each of those can and does affect
the overall network). A user network of Earth stations requires the type of network
management discussed in Section 11.2.3. The difference in this case is that, rather
than dealing with the ICN (which is, by definition, internal to the activity of the
satellite operator), the management of a user satellite network involves the custom-
ers and their respective activities. The range of possible needs and problems
increases, sometimes exponentially, with the number of networks under manage-
ment responsibility.

A classic example of such network management relates to a hub-based VSAT
network serving one or more corporate customers. In the shared hub environment,
the network operator monitors and controls a quantity of remote locations (some-
times numbering in the thousands), a backhaul circuit to the customer's host
computer, and a set of end-to-end communications applications as required for
the particular customer. If the satellite operator is fulfilling the network manage-
ment role, there likely will have to be a separate function with the appropriate
expertise made available. As the demands of customers increase, the size of the

organization and associated resources must increase correspondingly. The point here is that what develops is a different kind of activity from that described in Section 11.3.1. On the other hand, there will need to be continuous coordination and cooperation among those activities so the Earth station network can be installed, activated, operated properly, and expanded as required.

Integrated network management for an advanced non-GEO network represents a challenge not previously undertaken during the development of GEO satellites [3]. The two different architectures, represented by the bent-pipe/ground hub approach of GlobalStar and the routing/intersatellite relay approach of Iridium, have different requirements for user network management. As shown in Figure 11.4, the bent-pipe/ground hub approach allows user network management information to flow when the satellite is in view of both the user terminal and a gateway Earth station (or specialized control and monitoring station, if applicable). Information can then pass from user terminal over the satellite, through the gateway, and on to the network operating center (NOC) through an associated network such as the ICN. Using two-way communication, the NOC can send control messages in the other direction to remote users.

Normal handover procedures from satellite to satellite would be required to maintain continuity of management communication (which is how subscriber telephone and data connections are maintained as well). The overall network can therefore control user terminals and user services when gateways are available to provide the point of relay. If a satellite is in view but it cannot simultaneously connect to a gateway, then no management information can be delivered to the NOC (this may be academic since no calls can be established either). An independent
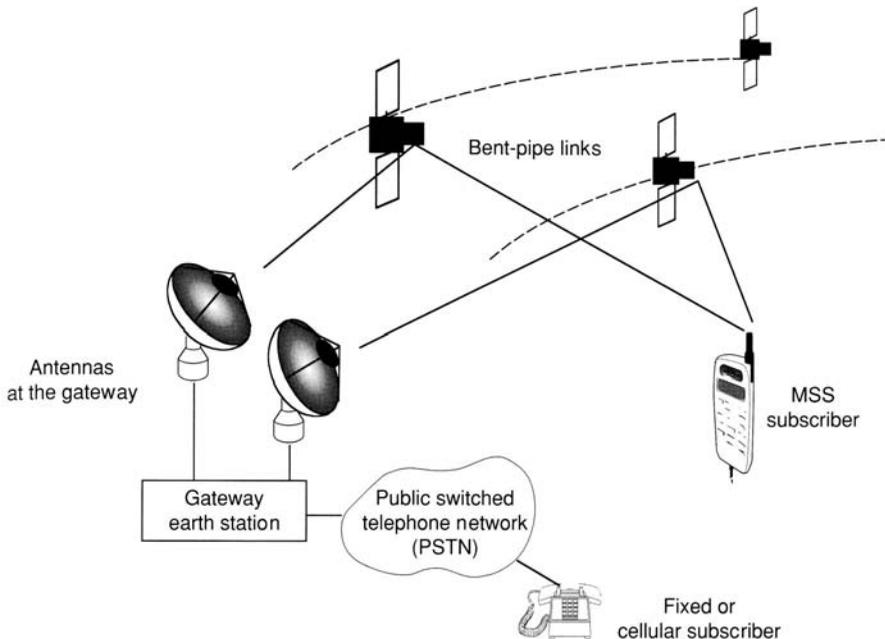
Figure 11.4   User network management for a non-GEO bent-pipe satellite constellation (based on GlobalStar).

terrestrial or satellite network is required between the gateway relay points and the NOC, since there are no intersatellite relays.

The routing/intersatellite relay, indicated in Figure 11.5, provides continuous paths for management data independent of gateways and terrestrial networks. As with subscriber calls, the management data may be exchanged directly between user and NOC on a unified basis. The messages can pass over any available links so as to leave maximum bandwidth available for revenue-bearing calls. Because of the routing nature of this architecture, the same basic procedures and protocols can be used for all classes of communication, whether associated with network management or the relaying of user traffic information. The network is robust in that, with an excess of satellites and links between them, a communication path can be established under nearly any circumstance. The one exception is when the user terminal or gateway is blocked from view of any satellite.

11.3.3   Payload Configuration Management

Requirements and procedures for management of a communications payload heavily depend on its particular design and complexity. The simplest is the basic bent-pipe repeater with service to a simple area-coverage footprint. Repeaters of this type are configured for service at beginning-of-life and may never require a configuration change. The exception is when an anomaly or failure is detected, requiring the replacement of a bad active component by an available spare. Transponders that contain passive or active gain control also can require adjustment by ground command, to maintain proper loading or in response to a request from users or network control personnel. Such changes possibly occur only once per month and can be implemented by manual SCC commanding.
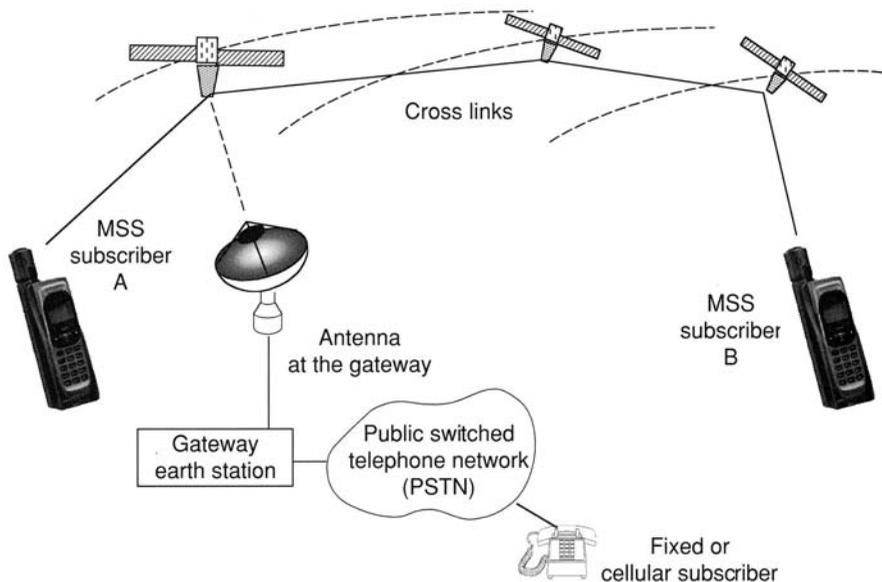
Figure 11.5   User network management for a routing/intersatellite relay non-GEO digital processing satellite constellation (based on Iridium).

As payload complexity increases, the requirements for ground-commanded changes increase as well. On the extreme is the digital processing repeater with the capability to route narrowband IF channels or baseband packets. The particular routing pattern or channelization can be preloaded through the command uplink or as a separate processor-control link, checked for accuracy and then activated to provide a particular service arrangement or pattern. That can be changed as frequently as dictated by the network, in periods as short as one second or longer than one month—it all depends on the processor design and the dynamics of the traffic.

Another type of payload is one that can establish connections on an individual call basis (similar to dialup telephone service). The request for service comes directly from the user and is relayed over the control link to the payload processor. There is no requirement for human intervention at the control center; rather, the system operates autonomously like a telephone exchange or cellular telephone network. The network control system on the ground should manage the end-to-end connection and perform such functions as collecting call records, tracking problems, and billing for services.

As payloads become more complex, management of onboard communication equipment and the actual links themselves become the most complex task of satellite operation. Traffic in such a system is dynamic and can produce interesting and unexpected effects, depending on the particular distribution and timing of user demand in local areas and throughout the entire network. Examples of such networks have existed for many years on the ground, the U.S. long-distance telephone network operated by AT&T being a good example. For this case, AT&T maintains a major network management center in Bedminster, New Jersey, that can track telephone traffic demand at every node and on every link in the network.

A major satellite operator like Intelsat has not had to go that far for its system because its primary role is to provide transponder bandwidth to users. Individual satellites are monitored through TT&C Earth stations that are strategically located to view every satellite. Traffic is monitored using automated test equipment that then transfers the information to the headquarters. Communications controllers can observe the overall downlink performance of each satellite and coordinate corrective action wherever it is needed.

The approaches to managing user terminals in non-GEO systems were reviewed in Section 11.3.2. The first global mobile satellite system to go into service, Iridium, demands substantial resources and capabilities to carry out the needed payload management function. Each payload acts as a channel router for information packets that contain digitized telephone traffic and control information. The integrated system of satellites, control stations, and network control places large demands on network management computers and software.

11.4   Human Resources for Satellite Operations

The human side of satellite system management is of equal importance to the technology. Additionally, it poses difficult challenges to the new operator yet offers opportunities for people and societies wanting to advance. While developing the

required technologies for digital processing payloads and new launch systems is the domain of Ph.D.-level experts, the operation of the resulting satellite systems and networks often can be delegated to qualified technicians with specialized training, both formally and on the job. In fact, the highly skilled technologists who develop the system often are not best suited to operate it on a long-term basis. That is because those experts find more challenge from creating the system than from keeping it in working order. Good operations people, on the other hand, are motivated by meeting the day-to-day operating needs of the system and seeing that users are satisfied. Of course, there is an ongoing need for technologists to support the maintenance and upgrade of the system, and operations people should have a strong input into the design of the system to enhance operational performance and efficiency after services are cut over.

In spite of the fact that the critical hardware and software elements are produced in highly industrialized countries, virtually every country and region will have candidates who can satisfy the educational and background requirements suitable for operations and management staffing purposes. The key to success on the human resources side of the equation is to select the proper staff and provide them with quality training and support during system development.

The initial months and years of service are a period when bugs are worked out of the systems and the people become proficient. In the words of a new satellite system manager in a developing country, "First, the contractor puts the system in and operates it for a trial period—after that, it's our turn." Experience has shown that the staff of new satellite operators are well up to the task, provided the right steps are taken along the way. The guidelines in this section are offered with that in mind.

The structure of the satellite operator organization must follow the technical architecture and resulting demands of the hardware and software. In addition, a satellite operator is usually a business organization to address marketing, sales and administrative functions as well. We assume in this discussion that the business side has been adequately taken care of, so we concentrate on the technical side of the organization design.

Figure 11.6 is a typical organization chart of a medium-size GEO satellite operator that services one country or region. The following list summarizes the
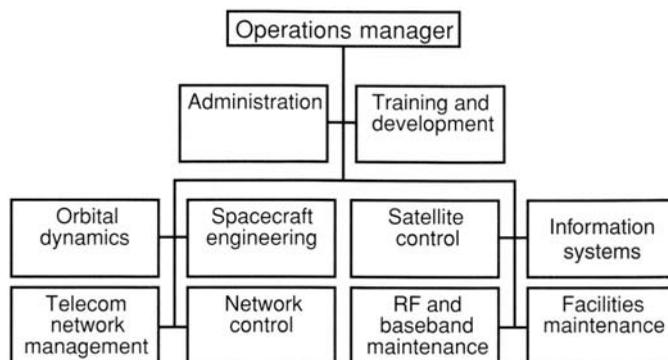
Figure 11.6 The technical organization of a typical GEO satellite operating company.

major organizational functions and how they relate to one another. We have identified a total of 10 different activities, which generally fall along lines of particular functions. That generally eases the problem of identifying experienced talent or obtaining appropriate training. In any particular organization, some of the functions are combined in a common department, based on the particular skill sets available from staff.

1. *Operations management.* Any technical operations structure demands experienced and qualified general management. In the case of a satellite operation organization, it becomes somewhat difficult to identify a senior leadership team with the requisite background. The operations manager must have experience with an organization of this type and be familiar with the local style of doing business. The leadership team will also have the responsibility of building the organization, typically from scratch. The operations manager should work closely with staff, conducting daily meetings to review routine occurrences from the night before and to discuss upcoming events of general importance. Special problems must be addressed quickly and efficiently to prevent their escalation into a system failure.

2. *Satellite control.* This function was outlined at the beginning of the chapter and represents 24-hour-per-day controller activity that ensures safe operation of all operating satellites. The controllers are usually organized into shifts, each under the direction of a shift supervisor. While on duty, a controller spends much of the time behind a computer workstation console that displays all vital facts for one or more satellites (depending on the control and management philosophy). The supervisor generates daily plans and shift assignments and monitors total performance to meet quality and reliability objectives. Unusual data and suspected anomalies must be reported immediately, in accordance with well-crafted procedures for their evaluation and resolution.

3. *Spacecraft engineering.* Spacecraft engineers are to operating satellites what physicians are to their living human patients. Through basic technical education in electronics, spacecraft systems (as reviewed in Chapters 6, 7, and 8), ground TT&C facilities, and the principles of satellite control, spacecraft engineers are in a position to better learn the particular characteristics and needs of the satellites under their charge. They are called on to review routine spacecraft data obtained through telemetry processing and to identify long-term trends that might indicate incipient trouble. They are available to assist the controllers on duty who detect potential or actual problems with the operating satellites. To aid in that type of investigation, spacecraft engineers must have access to the spacecraft design information from the manufacturer or be able to make contact with technical specialists at the manufacturer or consultants, as appropriate.

4. *Orbital dynamics.* Also called flight dynamics or maneuver planning, this is the rather technical function that determines the precise orbit parameters for each satellite and produces the needed orbit correction maneuvers for GEO stationkeeping or LEO orbit maintenance. The object of those plans

is to satisfy the position control requirements with satisfactory error and to extend fuel lifetime to the maximum extent feasible using the available technical and human resources.

5. *Network operations*. This function was reviewed in detail in Section 11.3. As with satellite control, network control is a 24-hour-a-day job, because user services generally are available at all times. In a startup situation, this function might be combined with one of the others; however, as activity builds, it becomes a requirement that a dedicated staff be selected, trained, and managed. That ensures that the needs of users, many of whom are paying customers, are addressed quickly by staff who can focus on the demand of achieving a high degree of customer satisfaction. (Complex networks, particularly those using processing satellites, should start with a dedicated network operations staff.)

6. *RF and baseband electronics maintenance*. Since all satellite monitoring and control are exercised through the TT&C system, the ground equipment used for that function must be maintained in full working condition at all times. The maintenance technicians who do that are familiar with the RF equipment for the uplink and downlink with the satellites and the baseband equipment that provides the modulation/demodulation and encoding/decoding functions. The type of training and experience base may be sufficiently different between the two areas to warrant separate teams. Also, in cases where the RF equipment is part of a remote TT&C Earth station or gateway, the separation of the specialties also will be required. Repair of particular failed subsystems usually is effected by replacing printed circuit boards, modules, or entire units, as appropriate for the specific piece of equipment. Often, the TT&C ground maintenance engineers work hand in hand with the other maintenance experts discussed in the following functions.

7. *Information systems and software maintenance*. Both the TT&C station and the SCC rely heavily on computer systems, networks and software to carry out their many functions. There is often a variety of computing machines in each of those facilities, ranging from PCs on desktops and servers in equipment rooms. In some organizations, all the computing and network resources are put under one manager and his or her group. That promotes standardization and the possibility for efficiencies in buying and maintaining equipment and services. On the other hand, some organizations prefer to segregate computing and software according to the activities of the particular major group. If that is the case, multiple computer hardware and software maintenance activities will be spread around the satellite operator's organization. There has been a trend since the 1990s to outsource the generic aspects of computing; however, the specialized needs of satellite control often dictate that some software expertise be maintained in-house.

8. *Telecommunications network management*. The telecommunications requirements of the satellite operator are satisfied by the ICN, discussed in Section 11.2. The size and sophistication of the ICN reflect the complex-

ity of the space segment and service networks being managed. In the case where the satellite operator is able to perform the mission using a single combined TT&C station and SCC, the ICN is reduced to providing good access to the public telephone network. That is still a minimum requirement because of the need to interface with customers and other outside entities. As more sites are added, the network and its management become a considerable challenge. Systems of this type demand that a separate telecommunications network management function be established and supported throughout the operating life of the system. Personnel for this area would have received specialized training in telecommunications and networks. Technology has been changing rapidly in this field, so it is important that the manager and certain of the lead engineers and technicians continue their educational and research activities. A detailed discussion of network management can be found in Section 11.2.3.

9. *Facilities and mechanical equipment maintenance*. Just as a communications payload needs a properly working bus, an Earth station or SCC is just as dependent on support functions that allow the critical electronic equipment to continue to work properly. These systems include the prime electrical power, derived from the commercial power grid of the country where the facilities are located; backup power in the form of the UPS; heating, ventilation, and air conditioning (HVAC); internal telecommunications (telephone and LAN); water; gas (if required); and physical housing in buildings. Maintenance of those facilities must be performed on a timely basis and staffed to respond immediately to problems. This is one of those areas where, when things are going normally, no one notices nor cares. But when an element of the facility fails to perform its function, everyone's patience is put to its limit. A simple thing like a heavy downpour of rain or a snow-clogged access road can put the system at risk, so the right human resources with the right tools need to be available, either on site or on call.

10. *Training and development*. The final organizational capability deals with what is the fundamental purpose of this book, that is, to develop and enhance the capability of professionals who are either new to the field of satellite communications or who need to update and extend their knowledge and skill level. Any well-run satellite operation needs to have a permanent training and development function that has the role of providing the training systems and materials and ensuring that they are used properly. Satellite control personnel, maintenance engineers, and other experts must be trained and retrained over time. That can be done in-house with trainers who are on staff, combined with training provided by the various vendors that supply the hardware and software. For a new system, it is common for the satellite or ground station provider to conduct training courses and on-the-job training both prior to handover and for several months after the first satellite is launched. From that time forward, it would be an internal responsibility of the satellite operator to make sure training and development needs are satisfied.

Our final word on human resources is simply to say that the most important activity in bringing a new system into operation is the proper selection and training of staff. That is not as simple as it sounds because there are many considerations in finding the right people and making sure they are prepared to do their jobs. In any organization, the structure is never perfect, so staff will always find that there are overlaps and gaps. That is normal and to be expected. Every organization must locate and retain people who are willing and able to pick up the slack and make sure the needs of the system are met. Choosing high-quality people is extremely important; providing them with the right training and tools is just as crucial.

References

[1]   Howard, R. E., "Low Cost Little LEO Constellation," *AIAA 16th Internat. Communications Satellite Systems Conf.*, Washington, D.C., February 25–29, 1996, Collection of Technical Papers, American Institute of Aeronautics and Astronautics, New York, paper no. 96-1050, 1996.

[2]   Smith, D., "Operations Innovations for the 48-Satellite Globalstar Constellation," *AIAA 16th Internat. Communications Satellite Systems Conf.*, Washington, D.C., February 25–29, 1996, Collection of Technical Papers, American Institute of Aeronautics and Astronautics, New York, paper no. 96-1051, 1996.

[3]   Aidarous, S., and T. Plevyak, *Telecommunications Network Management-Technologies and Implementations*, IEEE Series on Network Management, New York: IEEE, 1997.

Satellite Systems Engineering and Economics

Our final chapter brings together the elements of a satellite communications system in a manner that it is designed to meet its technical and financial objectives. We have chosen to divide this into two major efforts: (1) systems engineering, which is the discipline that starts with the overall requirements that the system must address along with the selection of major elements that meet those requirements, and (2) engineering economics, to evaluate the costs of the elements and how the system funds those costs through budgets and revenues, as appropriate. The end result is a system that can be constructed and operated throughout its lifetime, returning capital and, if appropriate, profit, to its owners. We conclude with our approach to creating the future of satellite communications.

12.1   Satellite Systems Engineering Principles

The general discipline of systems engineering has existed a long time, beginning during World War II and growing through major technology programs associated with the Cold War. It took a commercial direction when space technology became the basis of the satellite communications industry. Thus, people who were schooled in systems engineering on government programs (nuclear subs, ICBMs, and large aircraft) found it very useful in the development of larger commercial projects like INTELSAT IV, DIRECTV, and Iridium. The following is a general statement that introduces the objectives of systems engineering [1]:

> An interdisciplinary process to ensure that customer and stakeholder's needs are satisfied in a high quality, trustworthy, cost efficient and schedule compliant manner throughout a system's entire life cycle.

The basic premise is that you start with the customer's requirements and develop the system from there. This is different from a thinking process that starts with some particular technology or existing system and tries to figure out who can use it and for what purpose. Of course, you need a technology base and resources in space and ground segments to put a satellite system together. The difference here is that the customer requirement comes before the selection of specific technology or system.

The steps in the total systems engineering process have been defined for us by the International Council of Systems Engineering (INCOSE), a professional society

for systems engineers that addresses some of the larger challenges in bringing a system into existence. The essence if the INCOSE approach to systems engineering is to follow a structured process from beginning to end. The elements are shown in Figure 12.1 and reviewed here.

The process starts with the all-important requirement; and it ends with the delivery of some output in terms of a system or a result from having the system. There is potentially another element to be inserted after *Investigate alternatives*, and that new element is *Decompose into elements*. This is where the systems engineer relies on the specific knowledge of how the system can be put together from a combination of existing technology and possibly with the introduction of as-yet unproven technology.

We can start with the first step in the process: *Stating the problem*; this is something that is probably obvious but deserves special mention. It is driven by a business or other operational requirement, typically to establish some kind of communications capability. Thus, it would describe the key drivers for the requirement:

- Who requires the communications (the market, the people, the organizations);
- Where are the points of delivery of the communications (locations, such as physical places on the ground or platforms that are in motion on the ground, on water, or in the air);
- What the nature of the communications are (voice, video, files, or other forms of data).

This should be as abstract and detailed as possible. From here, we go into the "What" of the system, namely the alternative approaches to solving the problem or satisfying the requirements. This is where we look at satellite orbits, network architecture (star versus mesh), and transmission standards to be employed. For most data communications applications, we will employ the Internet Protocol and its family of protocol layers. Satellite links would likely use techniques that adequately address the approach taken.

One of the above steps is generally subsumed within others, namely, Decompose into elements. Bringing this out makes clear that the systems engineer needs to get
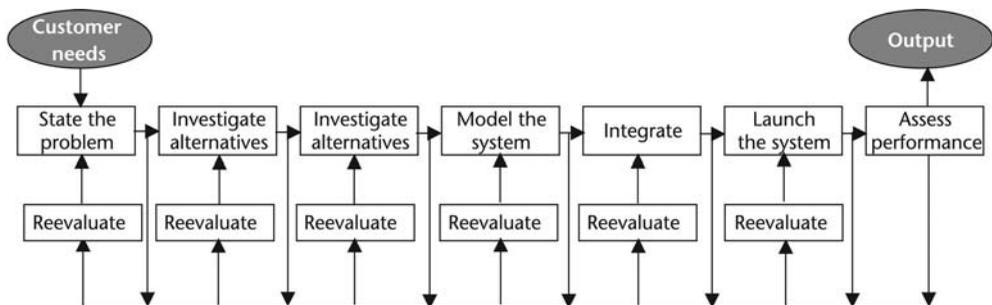
Figure 12.1   The SIMILAR process for systems engineering. (*Source:* International Council of Systems Engineering.)

down into the details of the technology and components of the system as they impact the objectives cited above (e.g., high quality, trustworthy, cost-efficient, and schedule-compliant).

Say that you are assigned the task of putting in place a satellite communications system for several Pacific islands. The process above starts with a problem statement, in as much detail as you can construct and collect from the customer or end user. At this point, the alternatives in the next step are not as yet known. From the problem flows the requirements in sufficient detail to permit us to architect an initial approach, based on GEO or non-GEO satellites and an associate ground segment. We have not as yet selected technology for this, nor would we know the best providers of hardware, software, and services. Anyone who has managed this kind of effort knows that it is both difficult and a key success factor for the result. We may save time by skipping these steps, but that saving will be overwhelmed by time and money spent later correcting problems and misunderstandings that could have been avoided.

The rest of the steps proceed logically through the engineering process familiar to all of us engaged in technology undertakings. A key attribute of systems engineering is that we evaluate and reevaluate performance at each step of the way, until the system is ready for service (and after that to assure that performance is maintained).

After decomposition, much of the effort is taken over by subsystem engineers and other technologists who can define and develop the functional elements. It would be wonderful if these elements were produced exactly as defined so that the system performs exactly as intended. Many things happen along the way that can derail the final product. For example, the performance of a subsystem may fall short of the specification. Even if every element meets its requirements, the final integration of the system can uncover issues that were not identified during the early functional decomposition stage. Both of these problems can be redressed through the modeling process. In it, subsystem performance is simulated as is the performance of the overall system. As test and other data comes in from the subsystems, the system model is updated and run again to provide predications of performance and operation. When the actual hardware and software are delivered and integrated, the performance ought to fall in line with the models. If not, the same models can be adjusted and solutions tested for their appropriateness.

Once the system is launched, performance is reassessed to see that it has not changed during final integration and test. A key example of this is the end-to-end testing that goes on after a satellite is launched and put into service. Such testing is common in the telecommunications industry to characterize the throughput, time delays, and quality of service. The same applies to a satellite communications system; the difference is usually the greater distances involved and the difficulty of coordination across those distances. After service initiation, it is important to continue to assess performance to identify any degradation due to a variety of causes. Power amplifiers in satellites and Earth stations are notorious for causing such degradation; another is radio frequency interference that could appear and even increase in level as neighboring satellites and Earth station networks expand. Performance may also be affected as traffic on the satellite increases due to an

increase in demand. A role for systems engineering is to track key performance parameters and assess the changes cited above.

Figure 12.1 indicates return or feedback links after every step in the process. This reflects the learning that goes on throughout a typical development program. Both time and resources need to be made available to permit an adequate amount of these returns.

In the next two sections, we address the broad category of satellite communications systems—fixed and mobile. This subdivision is consistent with the two primary services as defined by the ITU. The broadcasting satellite, while separate from an ITU perspective, can be contained within the fixed portion because the communication is generally to fixed points on the ground. Mobile satellite services, on the other hand, are delivered to users and platforms in motion. Generally speaking, this has a greater challenge and impact on the selection of technology. Within FSS and MSS, we might describe broadband applications like video and high-speed data, versus narrowband applications like voice and simple messaging. This creates a four-by-four matrix of possibilities that can be described in a reasonably straightforward manner. Table 12.1 contains an example of such a matrix (this is not intended as a solution or recommendation).

12.1.1 Fixed Satellite Systems and Services

The FSS portions of the spectrum were allocated by the ITU for stationary earth stations and user terminals. As discussed previously, it falls within the C-, X-, Ku-, and Ka-bands, and offers substantial bandwidth for a full range of communications applications. It was also discussed that the network can take either of two topologies: the star, with its hub station that controls the network and provides a gateway to the terrestrial infrastructure, and the mesh, which allows remote sites to communicate directly with each other on a point-to-point basis.

Star networks are effective when the same data is broadcast over a coverage region and is the basis of the DTH business throughout the world (whether FSS or BSS). Mesh networks can be combined with the star, but there are technical challenges in making the service consistent and cost-effective. This is because every station must have the technical capability to communicate directly with any other. In a star, stations would only communicate through the hub.

A trend over the past decade has been to make broadband applications available to mobile platforms. This places special demands on the user terminal, which generally employs some kind of reflector (dish) antenna with a diameter of one

Table 12.1   Example of Service Evaluation Matrix for FSS and MSS Applications

	Narrowband Applications	*Broadband Applications*
Fixed satellite service	Can be provided into fixed antennas; potentially efficient in terms of bandwidth	Wide bandwidth available, but requires dish-type antennas requiring precise aiming or tracking
Mobile satellite service	Limited bandwidth available, suitable for voice and medium speed data; nontracking antennas permitted	Can be provided at high cost; tracking antennas may be required to provide adequate link performance

meter or greater (smaller diameters are used for some services, such as DTH and broadband data for the home). Greater data rates may demand larger antennas and higher transmit power from the ground. In spite of this, fixed satellite service has been used on board aircraft, ships, and moving vehicles. This may not have been the intent of the ITU is allocating spectrum for the FSS, but there are viable techniques for offering broadband applications to users on the move.

12.1.2   Mobile Satellite Systems and Services

The MSS portion of the spectrum is allocated in the L- and S-bands, where the available bandwidth is typically 30 MHz or less. This makes it difficult to achieve the same broadband applications that are common to the FSS. Historically, the MSS has been limited to narrowband applications like telephone, fax, and low-speed messaging. With the launch of large GEO MSS satellites like Inmarsat IV and Thuraya, bandwidth has been increased through multiple spot beams and improved modulation and forward error correction.

The architecture of MSS is generally based on the star topology since there is usually a hub facility involved in serving users. It behaves like the cellular base station and mobile switching center. Only the Iridium LEO system can provide a true hubless mesh network service that allows all users to access the satellites without touching the ground at intermediary points (although most connections are through a gateway).

At the time of this writing, a number of U.S. companies were developing GEO MSS satellites with broadband capabilities that rival that of the low end VSAT. The FCC accepted applications from three companies to provide ancillary terrestrial component (ATC) using satellite allocations at L- and S-bands. The approach here is to use the spectrum to deliver a terrestrial wireless service akin to 3G cellular or WiMAX fixed wireless service. There would be two or more GEO satellites that provide an overlay to allow communications with the same terminal devices when out of range of terrestrial base stations. Again, this would use the same frequency band on the satellite and terrestrial networks.

L- and S-band MSS spectrum is also used for mobile broadcasting principally of music and talk radio. However, the same system can deliver a video service to mobile users. As a result, there is a blurring of distinctions among the services and platforms, as well as broadband versus narrowband. The systems engineer who starts with a fresh sheet of paper will need to follow the process outlined at the beginning of this chapter; the best architecture in terms of FSS versus MSS and broadband versus narrowband may indeed be a surprise.

12.2   Satellite System Economic Principles

The successful implementation and operation of a system or network depend as much on economics as on technology. Once the development risks have been identified and addressed properly, the economics of satellite communication can be defined using the basic principles of capital budgeting and service development. Geostationary satellite systems are evaluated like other industrial fields, including

public telecommunications services. DTH and VSAT networks employing bent-pipe repeater satellites are well understood, so it is possible to make accurate predictions of what it will cost to implement and operate a system.

Unknown factors in system economics often have to do with new technology that must be developed and introduced for commercial service. Another set of unknowns involves the expected amount of traffic the system can carry and the prices users are willing to pay. Furthermore, competition, such as exists in the United States, is an extremely strong factor that can render even the best system implementation concept uneconomic.

A satellite system in its entirety is costly to implement, but the technology provides versatility and flexibility that are potentially greater than that of any other telecommunications technology. The cost of space segment implementation ranges from approximately $250 million for one conventional bent-pipe satellite to more than $10 billion for a complete DTH system. A medium-power GEO communications satellite delivers any combination of voice, data, and video services between a potentially large set of Earth stations located within the footprint. A comparable terrestrial network must be implemented on a wide-scale basis, and rapid relocation is virtually impossible. A non-GEO constellation, on the other hand, promotes the concept of ubiquitous global telecommunications service, independent of the terrestrial infrastructure. For some payload designs, communications satellites are attractive from an economic standpoint because the particular mix of services can be altered rapidly and the value of the investment is not lost due to changing demand. A satellite that has been launched will normally see its use increase over time even to the end of orbital life, when it must be retired from service.

The communications Earth stations employed to deliver services cover a wide range of applications and costs. Large Earth stations used as broadcast centers and telecommunications gateways have long economic lives, since it is possible to expand, modify, or even replace the specific equipment used to provide new capabilities. Low-cost end-user terminals, particularly DTH receivers, VSATs, and MSS mobile phones might have relatively short lives of 2 to 5 years. Part of the reason for that short economic life is the trend toward obsolescence as newer products are introduced. The advent of microelectronics has greatly increased the versatility of ground equipment at the same time that unit costs have decreased. Also, higher power satellites permit a significant reduction in ground antenna size and RF electronics costs. Those mitigating factors encourage the replacement of functional but obsolete Earth station equipment.

The approach that follows is meant as a framework for understanding the underlying economics of a satellite system. It should be possible to use this framework to create a specific economic model of any system so the detailed costs can be determined. As a rule, the economics of the system go hand in hand with the functions it is required to perform. Typical studies of system economics consider the recovery of the original investment by including the time value of money and computing equivalent annual payments. The investment costs are embodied in the block diagrams of the elements and for the system as a whole. The annual costs usually are more difficult to assess, involving labor and other expenses for managing, marketing, and operating the system, and financial matters such as the cost of money and taxes. In large, complex networks, costs and therefore revenues must

be split among the space segment operator, the Earth station operator, and the ultimate service provider or distributor. That can be exceedingly difficult to assess in advance of actually creating the required business structure. Economic studies are still useful to size the business and to attempt to convince investors (internal or external) that a satellite system is an attractive investment.

12.3   System Development Methodology

The basic approach for evaluating system economics is to view the system as an investment. When in operation, the system produces services that can be sold to customers. That generates revenues that, hopefully, will recover all costs and ultimately yield profit to reward the investors. Figure 12.2 presents a simplified flow chart that identifies the various stages of system implementation and operation. Financial inputs (shown as arrows on the right) are required for each aspect of system evolution, that is, investment to create the system and service revenues from users (e.g., customers, subscribers). The objective is to end up with an operational system that is self-sustaining in an economic sense. This same principle applies to other investments in the telecommunications fields, including cellular telephone, cable TV, and undersea fiber optic cables. The unique aspect of satellite communication is how we address the risk associated with launching and operating spacecraft in orbit. Risks to the ground segment exist as well, but there is a perception that the risks are addressed in the routine course of running any terrestrial business.
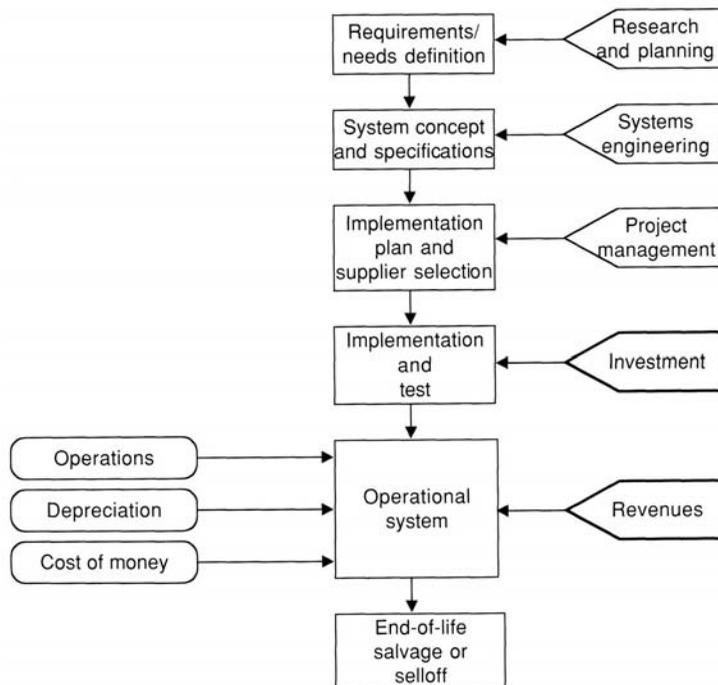
Figure 12.2   Flow chart illustrating the economics of implementation and operation of a commercial communications satellite system.

Beginning at the top of Figure 12.2, the first step in the planning process is to determine the service requirements or needs the system is to satisfy. Research into the needs of potential users is performed over some period of time either by investigating historical patterns of telecommunications usage (satellite and terrestrial) or by conducting surveys among the prospective users. Such a project represents a one-time cost, being the first fixed expense in system development. It can be considered part of the investment in the system or it can come under the normal operating budget of the private company or government agency doing the research. Developing requirements can be costly because data must be collected for each market segment and geographic region to be served. The most costly is primary research, in which one actually goes out into the field and gathers data directly from the targeted user population. Secondary research is considerably cheaper and quicker to conduct because it amounts to assembling already available data from published sources. It includes reports from consulting or market research companies, as well as data published by government and international bodies like the FCC, the National Association of Broadcasters, the European Commission, the World Bank, or the ITU. Most market research studies begin with secondary research and move on to primary research only when initial feasibility has been established and there are some special (and extremely important) questions to be answered.

With the business requirements more or less quantified, the next step is to use the systems engineering process (Section 12.1) to determine the system concept. This is a major planning project and is best performed by an organization that is very familiar with the technology and economics of satellite communications. Such organizations include large manufacturers of spacecraft, experienced satellite operators, and certain specialist consulting firms with experience on these types of projects. The final product of such an effort is a system design document that lists performance specifications for the major elements of the system. For example, the system concept for an advanced GEO satellite network would identify the geographic coverage requirements, its frequency band(s), repeater design (analog or digital), communication capacity, and other specific characteristics. Part of the design would be to determine, through appropriate tradeoff studies, the modulation and multiple access methods to be employed for each service. The locations and service capabilities of Earth stations also would be identified.

One of the most difficult aspects of determining the system concept is to know what technologies are available and appropriate. That information comes from the resources of the system designer or can be learned through exposure to the marketplace. It is common to issue a request for information to relevant developers and manufacturers so they can suggest design methodologies and solutions. Besides the technical aspects of that effort, the system designer needs to prepare estimates of the investment and operating costs. A thorough systems engineering study includes tradeoff analyses of various alternatives, yielding an approach that is nearly optimum for the particular set of user requirements. Another important aspect of the process is the consideration of how the system can be expanded or modified after it has gone into operation. That can have a decided effect on the approach taken for the initial system design.

The design is not finished, however, until the requirements/needs definition is revisited to verify the original assumptions. Done properly, the top two boxes in Figure 12.2 represent an iterative process, seeking the best technical and business solution. One result could be that one or more of the original services is discarded in favor of another more viable market opportunity. Likewise, a technology solution, which appeared attractive on first examination might have an excessive business risk associated with it.

Perhaps the most dynamic phase of the process is system implementation, during which most of the investment is made. Before construction begins, the functional requirements from the system concept are turned into detailed hardware and software specifications. With respect to major Earth stations, locations around the service area are firmly established with consideration given to the footprint of the satellite antenna. Contractors and manufacturers are selected and schedules set for construction of all the elements of the system. The system operator can perform the role of system implementation manager, but in many instances another experienced system integrator is selected. Depending on the types of facilities involved and whether a spacecraft must be constructed, system implementation can take anywhere from one to five years. Essentially all the expenses incurred during implementation become the investment in the system itself and are recoverable mainly through collecting revenues during its useful life. Alternatively, presale of satellite and Earth station capacity to one or more large users can provide a needed financial boost to offset some of the capital and initial operating costs.

After the system is implemented, the major space and ground elements are tested to verify performance and functionality. Then the operational phase begins and lasts as long as useful services can be provided. The economic input to the operating system usually is in the form of service revenues from users, as indicated at the right of Figure 12.2. Revenues can be collected monthly, yearly, or in a lump sum before service is even rendered. Examples are the monthly telephone bill that most of us pay and the up-front payment for a transponder on a condominium satellite. The system must be operated and maintained by ground personnel over its useful life, indicated at the left of the figure. Operations expenses cover the labor resources discussed in Chapter 11. Depreciation is a charge against revenue to allow for the fixed lifetime of the space segment and ground assets. For example, a 15-year GEO satellite depreciates at the rate of 1/15th its initial investment cost per year (in actuality, a much shorter economic lifetime is assumed to allow the operator to mitigate taxes). The third charge, finance, is simply the cost of money, amounting to the effective interest that might be paid back to lenders. On top of that is the profit that would result when all costs are paid, delivering value for the owners/investors.

Additional investment would be made over the life of the system to add new users and to provide expanded services, in terms of both quantity and type. Any proposal for new facilities would go through the entire process described above, although on a smaller scale. Repair of existing facilities associated with current services comes under the expense of operating the system and would not be a new investment. When the system or an element has reached the end of its useful life, due to wearout or to obsolescence, it is disposed of. If it can be salvaged or sold,

some additional recovery of investment money is possible. Otherwise, it is disposed of in a manner that cannot interfere with continued operations or future activity.

A business plan is a document that delineates all the information discussed in this chapter to this point. The following is a brief list of typical major sections of a business plan that might be used to describe a proposed satellite communication project:

- Executive summary;
- Basic service proposition (services to be offered);
- Satellite system description;
- User terminal equipment design, manufacture, and distribution;
- Spectrum, orbit assignment, and regulatory matters;
- Expected usage patterns and service take-up (market research);
- Competitive analysis;
- Marketing and sales plan;
- Financial model;
- Source of investment funds;
- Financial results and sensitivity to critical parameters;
- Risks and their management;
- Organization structure and implementation plan.

The breadth of the list indicates the wide range of the type of information that a good business plan must contain. Of course, there would be additional descriptive information to help potential investors understand every aspect of the system and the business. Most important is how investors' money will be protected, increased, and returned. Good business plans are not easy to assemble. They consist of credible and consistent market research, development plans, financial analyses, and a clear business structure. Often, the people behind the business are the ones who are in a position to assure potential investors that the proposed services can be delivered in a profitable way. To sum it up, the definition of a good business plan is one that can be financed.

This has been a brief and generic discussion of the evolution and economics of a system used to provide telecommunications services. More information is provided in our other work [2]. The size of this type of planning effort should not be underestimated; it is the responsibility of the system developer to do adequate homework. In the rest of this chapter, the particular aspects of satellite communications are discussed using that framework.

12.4 Space Segment Economics

The economics of implementing and operating the space segment of a satellite system are dominated by the initial cost of manufacturing the spacecraft, providing the launch into orbit, and insuring the combined value against failure. To that is added a much smaller, but significant, amount to implement the required TT&C ground facilities. Figure 12.3 reviews the main elements of a space segment, includ-
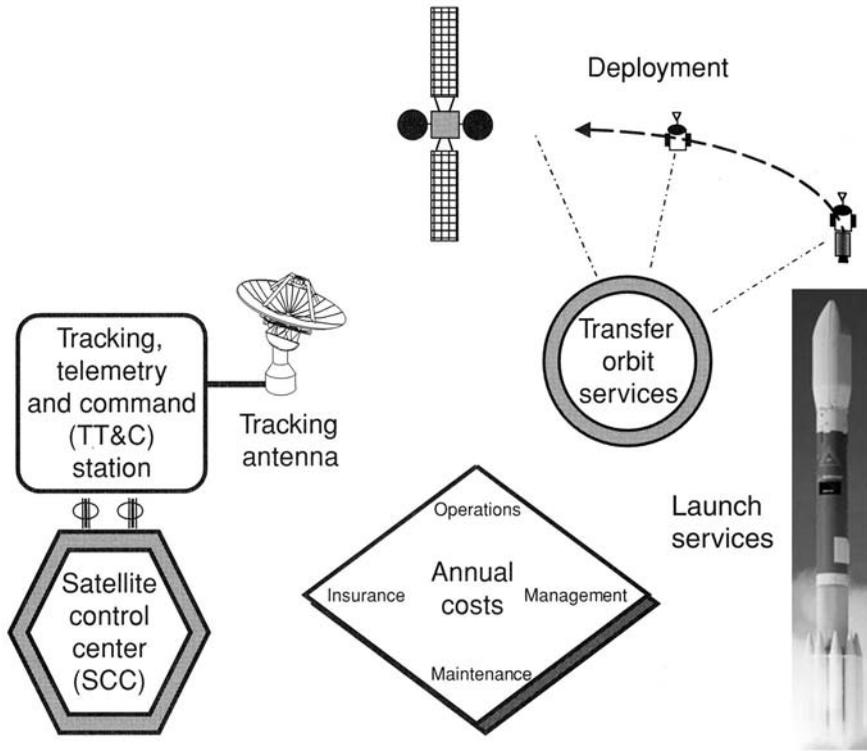
Figure 12.3   Major elements of the investment and operations expense of the space segment of a commercial satellite system.

ing initial investment as well as operating costs. A spacecraft manufacturer typically enters into a fixed-price contract to design, manufacture, and deliver the required quantity of communications satellites. A separate contract usually is required with the launch agency, which in turn pays for the LV hardware and then provides the integration and launch direction services at the launch site. Some customers have included the LV services along with the spacecraft in what is called a delivery-in-orbit contract.

Following injection into the desired orbit, either the satellite operator or the spacecraft manufacturer takes over tracking and control of the satellite, performing the services indicated by the circle in Figure 12.3. Non-GEO spacecraft may be designed to configure themselves when released from the LV, thus reducing the demand for ground-based tracking. LV and tracking services normally are considered part of the investment cost of the system. Prior to the start of revenue-producing service, a GEO satellite is drifted into its final position, configured for operation and tested to ensure that the launch environment has not damaged any subsystems. An ever increasing element of space segment cost is the launch insurance, which provides coverage for a possible launch or injection failure. Premiums for such insurance have become extremely costly, running 12–20% of the cost of the spacecraft, LV, and insurance. That is roughly equivalent to holding a financial reserve based on one failure in five to one in eight attempts, which is conservative in comparison to long-term experience. However, insurance has become nearly essen-

tial because of financial commitments to other parties, including public and private stockholders.

The TT&C ground facilities (shown at the left of Figure 12.3) are part of the space segment investment, to be continuously maintained during the life of the satellites. Due to the close tie to the particular satellite design, it is a common practice to procure most of the electronic equipment and software from the spacecraft manufacturer. More recently, some specialist technology companies have developed generic satellite control equipment and software to provide an element of competition into what has been a very closed market. In either case, the manufacturer must demonstrate prior to launch that all the ground and space components will work together properly.

The annual costs associated with operating the space segment, indicated in the diamond-shaped element in Figure 12.3, include expenses for the personnel who perform the technical and administrative functions at the satellite control center and the TT&C station. The cost of training the technical staff is high, and many personnel are required to be available around the clock. Fortunately, only a limited number of facilities are needed to operate a space segment consisting of several communications satellites. A global constellation also can be managed with limited TT&C ground assets, since a reliable intercommunication network allows centralization of technical and management functions.

The expenses of maintaining the ground facilities are usually fairly small due to their compact nature. The people who perform that task (introduced in Chapter 11) tend to be quite resourceful and versatile, which is probably the result of having to perform a wide variety of tasks on critically operating equipment. Nevertheless, an ongoing relationship with the spacecraft and ground equipment suppliers is essential to be able to maintain and expand the system over time. Perhaps the most difficult task in a commercial environment is the marketing of the services on the system. That can involve a large staff of sales engineers and service representatives, organized in teams to cover different user groups, applications, regions, and facilities.

Another annual expense is that of on-orbit insurance. This type of coverage, introduced in the late 1970s, allowed satellite operators to address part of the risk of loss of a working satellite. Later, purchasers of transponders needed this insurance coverage because it was demanded by the banks that provided the funds necessary to purchase the transponders in the first place. It could be argued that on-orbit insurance (also called life insurance) provides little benefit because there have been relatively few significant commercial satellite losses once successful placement into service has been achieved. On the other hand, the financial loss is great in the event of a total failure of a satellite, which could justify the coverage. On-orbit insurance works like depreciation, paying an amount that approximately equals the remaining value of the satellite. A failure that occurs in the first year (but after the launch insurance coverage has run out) would result in payment of nearly the full replacement cost of the satellite (less inflation, unless that had been considered). The payment for failure at the end of life is, by definition, zero. The insured value in the intervening years is found on the line connecting those end points. The policy may also cover a partial failure if the satellite were to lose, say, 50% of its capability.

The relative magnitude of each major cost element for a two-satellite system is shown in the pie chart in Figure 12.4. Annual costs have been considered by taking the simple sum of such costs over an eight-year period, which is one-half the actual life of an assumed space segment. That takes account of the time value of money (money spent several years in the future is worth less than money spent in the current year). The most costly single element is for spacecraft, representing approximately one-third the total. The sum of the cost of launch and launch insurance is comparable. Investment in ground facilities for the TT&C station and SCC is relatively small (2%). Operating expense over the 8-year period takes another relatively large chunk, approximately one-third the total expense.

12.4.1   Space Segment Investment Cost Elements

As discussed at the beginning of this chapter, the investment cost elements of the space segment are purchased prior to the initiation of operation and are intended to last through the end of life of the satellites. In larger GEO systems and constellations of non-GEO satellites, there is a need for methodical replenishment of spacecraft when they reach end of life. That places on the satellite operator the burden of maintaining a new flow of investment in spacecraft and LV services for as long as the business can be sustained. Replacement GEO spacecraft usually are quite different from their predecessors due to enhancements in coverage, power level, and even frequency bands. The elements to be included in a basic GEO space segment are indicated in Figure 12.4 by the shaded areas and are discussed next.

12.4.1.1   Spacecraft

A communications spacecraft is designed and built as an integrated system, as presented in Chapters 6 through 8. On delivery to the launch site, it is capable of
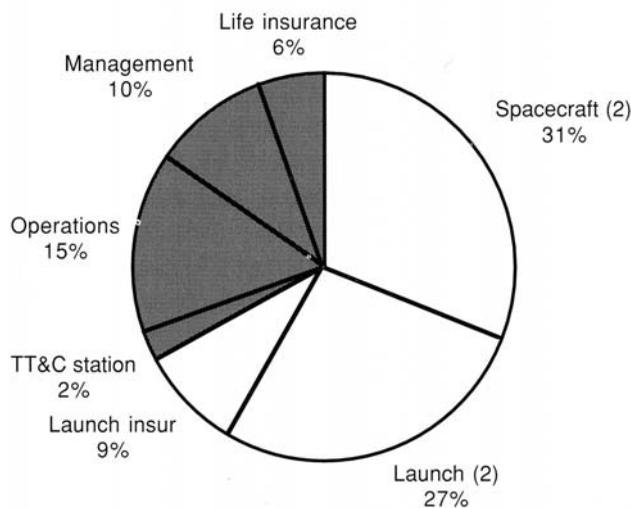
Figure 12.4 Allocation of the cost of a space segment composed of two satellites, including the initial investment and 8 years of operations support.

being placed into orbit and operated for its useful life without direct servicing. Any repairs or adjustments must be made by ground command and with test and redundancy capabilities that are built in. Once a satellite is placed into orbit, it generally cannot be removed from service or stored for later use. There is an exception for a GEO satellite that has been placed into an inclined orbit. If the inclination is positive, meaning it will decrease to zero at a rate of 0.8 degree per year, the operating life can be delayed. Once a satellite is actually placed into service, its cost is completely "sunk," meaning it cannot be recovered except by selling services to users or by selling the entire satellite to another operator. The latter applies only to GEO satellites because non-GEO designs are highly customized for the particular constellation. Provided that the footprint and frequency plans are suitable, a GEO satellite can be repointed or moved to a new longitude to provide service in a different manner or to a different region.

It should be evident from these discussions that a communications spacecraft is a complex and therefore costly item. As reviewed in Chapter 2, the historical trend has been from small spacecraft with few transponders to larger spacecraft with many transponders, multiple frequency bands of operation, and digital processing. The power levels of the individual transponders have increased over the years. Even ignoring inflation, the cost of a spacecraft has continued to increase in line with spacecraft weight and complexity. Communications payload weight increases because the quantity of repeater electronics and the number of the antennas (as well as their size) have increased. A good illustration of this is the Anik F2 satellite, shown in Figure 12.5. Providing more payload power expands the quantity of solar cells and batteries; that, along with the enlarged propulsion system and fuel supply, increases the mass of the spacecraft bus. The weight of the structure
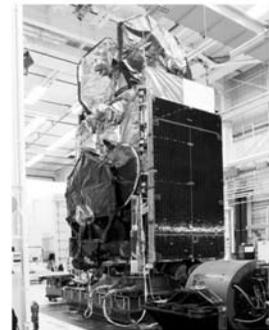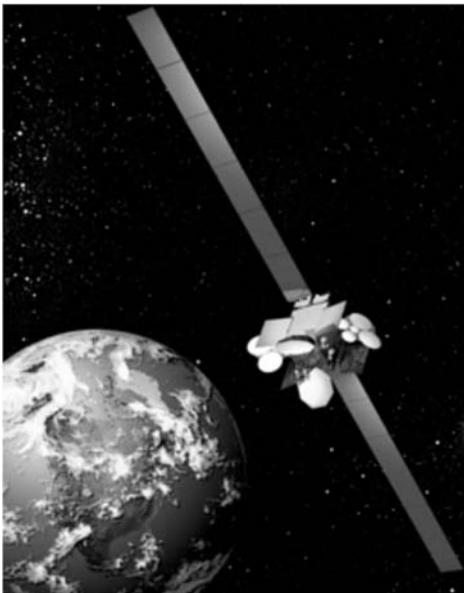
Figure 12.5 The Telesat Anik F2 satellite provides C-, Ku-, and Ka-band payloads with a prime power system capable of 17 kW (http://www.telesat.ca/satellites).

also has to increase to provide the volume and physical strength to support the expanded payload and bus.

As a rule, it is possible to view the cost of a spacecraft as being directly related to its mass; the heavier the spacecraft, the more it will cost. Other factors have a direct bearing on cost as well, including the number of components, the complexity of the payload, and the maturity of the design. The relationship is not linear. Because one needs the same project team size for a big satellite as a small one and has only slightly more design documentation and meetings, there is an economy of scale (i.e., a decrease in the relative cost in terms of some monetary unit such as dollars per kilogram) as one moves toward larger spacecraft with more communications capability. Components of a spacecraft such as battery cells, solar cells, and electronic amplifiers can be manufactured with modern automation techniques, which tends to reduce cost. However, larger elements of the spacecraft typically are custom made by highly skilled technicians, whose time expended tends to increase as complexity increases. That aspect is difficult to change because all work must be of the highest quality to ensure proper operation during orbital lifetime.

An area in which automation has played an important role is in the testing of subsystems and the entire spacecraft. Digital computers programmed to control modern test equipment can perform the testing during each phase of manufacture and at the launch site. That increases the accuracy and the amount of data and can decrease the number of personnel on test teams. The primary benefit of using this form of automation, however, is the improvement in reliability that comes from being able to spot potential problems before launch. Once on orbit, this database becomes very useful during initial testing to see if there has been any detrimental change after launch and to aid in troubleshooting during life of the satellite.

12.4.1.2   Launch and Transfer Stage

As evidenced by Figure 12.4, the fixed cost of placing a GEO spacecraft in orbit is comparable to the cost of the spacecraft itself. Non-GEO systems have the advantage that multiple spacecraft can be deployed in a given launch event. That reduces the launch cost per spacecraft by the same factor (although there is a requirement to launch substantially more satellites to complete a particular constellation). A satellite operator does not purchase LV hardware but rather purchases the services of a launching agency that is then responsible for the manufacture and performance of the rocket system. Whether the transfer stage is part of the launch service depends on the particular LV and mission, as discussed in Chapter 10. An LV such as Ariane 5 or Zenit injects the spacecraft into geostationary transfer orbit (GTO), from which motor firings are required to reach geosynchronous orbit. Direct injection into geosynchronous orbit is available with the Proton rocket, eliminating both the transfer stage and the apogee kick motor.

Some additional expenses associated with the launch mission involve the support services at the launch site and, if appropriate, during the transfer and drift orbit phases. The spacecraft contractor can bundle the launch site (launch prep) expenses into the price of the spacecraft as may the launch agency. Transfer orbit services are fixed and independent of the particular satellite design and can be

purchased from another satellite operator. The spacecraft contractor normally directs the timing of apogee boost, drift orbit, and initial checkout activities and turns the tested satellite over to the satellite operator. However, the communications payload is tested by the satellite operator to determine the fitness for service. The costs of all those services, while essential, are relatively low compared to the investment for the spacecraft and the LV.

12.4.1.3 Launch Insurance

Launch insurance has been an important element of satellite system implementation, particularly when the system is operated as a commercial business and requires outside financing. The general trend toward higher insurance premiums for high-risk kinds of coverage (not just for satellites) also has contributed to the cost. The risks associated with launch have always been high, but a good commercial launch record between 1970 and 1980 made the risks appear to be diminishing. Consequently, launch insurance premiums during that decade were low enough to encourage satellite operators to insure not only their investment but expected revenues as well. That removed most of the business risk from the satellite operator and placed it in the hands of the major insurance underwriters, notably Lloyds of London. With Lloyds now largely out of the market, satellite operators have discovered a new breed of launch insurance underwriters who better understand the nature of the game.

A recent trend is for satellite operators to receive partial claims when a satellite can still provide a useful fraction of its design capability. This is one reason why on-orbit life insurance rates have gone up.

12.4.1.4 TT&C Ground Facilities

As discussed at the beginning of this chapter, the investment cost of TT&C ground facilities is tied primarily to the equipment block diagram and facilities layout. To that must be added the fixed cost of providing the needed telecommunications equipment associated with the ICN. It is a fairly straightforward matter to identify all the equipment necessary to perform the TT&C functions as well as those associated with system management. The buildings are designed so that the equipment is properly housed and environmentally controlled. It is also important that human factors be taken into account, since operations are conducted around the clock throughout the life of the satellites.

The relative cost of implementing the ground facilities (indicated in Figure 12.3) is small in comparison with that of any other element. Designing and implementing a proper ground environment, therefore, is a wise measure, even considering the extra cost in doing the job right. Placement of equipment for convenient access and the provision of well-designed operator consoles and support software tend to reduce personnel costs. It also results in more reliable operation of the satellites themselves—the proper resolution of problems depends on the capability of the people who staff the SCC and TT&C station.

Particular attention should be focused on the computer system used to monitor and control the satellites, perform orbital and engineering analyses, and operate

other parts of the TT&C facilities. Having excess computing power and backup computer equipment almost always will prove to be of significant value during the life of the system.

The personnel and the equipment within the facility require an internal telecommunications network to allow the technical staff to speak to each other and to exchange data over a LAN. In modern facilities, those are part of a cabling system (preferably fiber based) and an IFL between buildings and antennas. The reasons for using fiber rather than coaxial cable or twisted-pair cable are: (1) it virtually eliminates the possibility of conducted mutual interference; (2) it provides ample bandwidth for expansion of services and modifications for additions; and (3) it is more secure because of the greater difficulty to tap the cable. Another area is in the monitoring of the transponders and signals within them. That can be important in providing services to users, who rely on the SCC to control access to the communications payload and to resolve interference problems as quickly as possible. Costs of transponder monitoring systems vary widely due to the disparity in capabilities, including whether the receiving antennas (one per satellite to be monitored) are local or remote. Another cost factor is the degree of automation of the testing functions, something that can improve the efficiency of aligning new Earth stations.

The location cost of TT&C facilities can be substantial, depending on the particular country and closeness to urban areas. Real estate costs can be very high in attractive, built-up areas. For that reason, some satellite operators tend to place their ground operations in remote areas. That also reduces the problems associated with RF radiation and interference. On the other hand, the people who work at the site may or may not find rural living acceptable. There are issues concerning adequate utility services, access roads, and conveniently located communities where employees can live or, at least, find a place where they can eat. In extreme cases, the satellite operator provides bus or van transportation to help the employees deal with remoteness.

12.4.2   Annual Space Segment Costs

This section reviews the main annual costs of operating the space segment. In addition, the financing of the investment in the space segment usually results in an annual expense to the satellite operator for interest, depreciation, and return on investment. Because of the wide variety of ways for financing a system and the different circumstances that satellite operators are under, specific financing strategies and costs are not covered.

12.4.2.1   Satellite Operations

The performance of satellite control and network management is by the team at the TT&C station and the SCC. As discussed in Chapter 11, those personnel include engineers, computer scientists, and technicians, many of whom hold advanced degrees. Specific training in specialties such as spacecraft engineering, orbital dynamics, and communications monitoring often is provided by the spacecraft manufacturer because the knowledge relates to the particular satellite and TT&C system design. The operations personnel who perform the ongoing tasks at the

spacecraft and communications consoles need to be on duty 24 hours a day, seven days a week. As shown in Figure 12.4, the expense for that labor is substantial, amounting to approximately 15% of the total system cost, assuming an 8-year period. Because the quality and experience of the staff are vital to the successful operation of the space segment and the quality of services rendered, it would not be wise to economize in that area. On the other hand, the quantity of staff does not increase linearly with the number of satellites, yielding another potential economy of scale.

12.4.2.2 Equipment Operations and Maintenance

The ground equipment and facilities, which allow the operations personnel to perform their tasks, must be maintained properly during the useful lifetime. Those expenses, while significant, are relatively small in comparison to the annual operations costs reviewed in Section 12.2.1. Allowance is made for the utilities and maintenance on the buildings housing the SCC and TT&C stations. The electronic and computer equipment usually is fairly costly to maintain due to its complexity. There is a tradeoff between hiring a full-time staff to perform electronic maintenance versus contracting for such services with the original suppliers of the equipment or other maintenance specialists. Even if outside services are used, a skeleton crew of electronic technicians is still needed to respond to emergency situations at the SCC and TT&C station.

12.4.2.3 In-Orbit Insurance

The need for in-orbit insurance coverage depends very much on the circumstances of the particular satellite operator. For example, life insurance proceeds would compensate the satellite operator for lost revenues in the event of a loss of satellite capability. In the absence of such a need and if the satellite operator has sufficient on-orbit capacity to deal with any contingency, then life insurance might not be required. The significant amount shown in Figure 12.4 is based on the assumption that the insurance is to cover the cost of replacing the satellites in orbit (i.e., including the cost of the LVs), reduced by the factor of remaining life.

12.4.2.4 Management and Marketing

The nonoperations functions of management and marketing might be viewed as unnecessary overhead. From an economic standpoint, however, a space segment cannot become self-supporting unless those critical functions are provided. Management organizes the various personnel resources of the system and arranges for hiring, training, and administrative support. In the case of a commercial system, management is responsible for operating the system at a profit in the face of internal technical and administrative problems and external difficulties with suppliers, customers, and competitors.

The role of marketing is to provide a continuing flow of new customers and revenues, without which the transponders onboard the satellites probably would remain empty. In that context, the sales function (i.e., locating prospective custom-

ers and executing service agreements with them) is part of marketing. GEO satellite systems have, over the years, built their businesses on meeting the demand of a limited market. It is not uncommon for a satellite operator to obtain more than half its annual revenue from just two or three customers. Building a business further involves introducing new and often innovative ways to divide the space segment, making it more obtainable by a larger customer pool. Going from three customers to 50 to 100 means accepting lesser amounts of revenue per customer. In DTH networks, the customer quantities are measured in millions, so the means for selling to and servicing such a large customer community represent a completely different paradigm. Companies usually address that situation by forming a new company and populating it with experts in the new class of service and consumer mass marketing. That requires a different economic model of the project, along the lines of a cable TV or cellular business plan.

New applications for the space segment should be examined continually by the marketing staff so services can be adapted to changing user needs. Marketing and engineering have a joint responsibility for estimating the future loading of the current space segment and for the preparation of plans for replacement satellites. Another related task is preparing the applications for new orbit slots and working with the regulatory authority to protect existing slots, through the process of frequency coordination. That could involve a proactive role in domestic regulatory affairs and participation in ITU WRCs. For a regional or international system, staff become involved with the challenge of gaining access to different national markets and building relationships that lead to new business.

12.5   Earth Station Economics

The ideal way to evaluate the economics of an Earth station is to identify each major element and relate its cost to technical performance. Such a detailed approach gives assurance of realism in determining network cost even if the implementing organization purchases complete stations from a contractor. Purchase of stations on a turn-key basis means that the contractor designs, installs, and tests the complete station and simply turns the door keys over to the buyer. That approach can be taken on an entire network, a single station, or one or more major elements of a station. An alternative approach that puts more control in the hands of the system operator is to build up an internal staff of engineers and installation people who can procure the individual elements and then perform the integration themselves. There is a cross-over, above which the fixed cost of maintaining such a group can be justified because of the more stable workload placed on them. Below that cross-over, there simply is not enough work to keep the group sufficiently busy with useful technical assignments and project work.

A proper network design analysis and optimization must consider the relationship between the performance of the Earth stations and that of the space segment. This brings into play the requirements of the microwave links to and from the satellite and the technical parameters of modulation and multiple access, reviewed in Chapters 4 and 5, respectively. For an existing GEO satellite, the price of using the space segment could be determined by contacting the marketing staff of the

satellite operator, or, if available, a broker. The analysis, therefore, is simplified because the satellite technical parameters and associated costs are givens. The case in which a new satellite is to be designed as part of the exercise is beneficial because the characteristics of the satellite can vary along with those of the ground stations. As discussed later in this chapter, there are useful tradeoffs that can help define an optimum space segment design that provides the desired capability for the least cost.

12.5.1 Earth Station Investment Cost Elements

The major elements of a communications station were reviewed from a technical standpoint in Chapter 9 and are illustrated in block diagram form in Figure 12.6. The RF terminal provides the actual uplink and downlink paths to the satellite, while the baseband equipment arranges the information (video, voice, audio, and data channels) for efficient access. Interfacing of services with users is by way of the terrestrial interface and tail (if required). Indicated at the bottom of the figure are categories of expenses for monitor and control (e.g., network management), power and utilities, as well as annual expenses for operations and maintenance. Economic aspects of each of those sections are reviewed next.

12.5.1.1 RF Terminal

The two essential characteristics of the RF terminal are the transmit EIRP and the receive figure of merit (G/T). Besides being the leading parameters in the uplink and downlink, respectively, EIRP and G/T have a direct impact on the investment
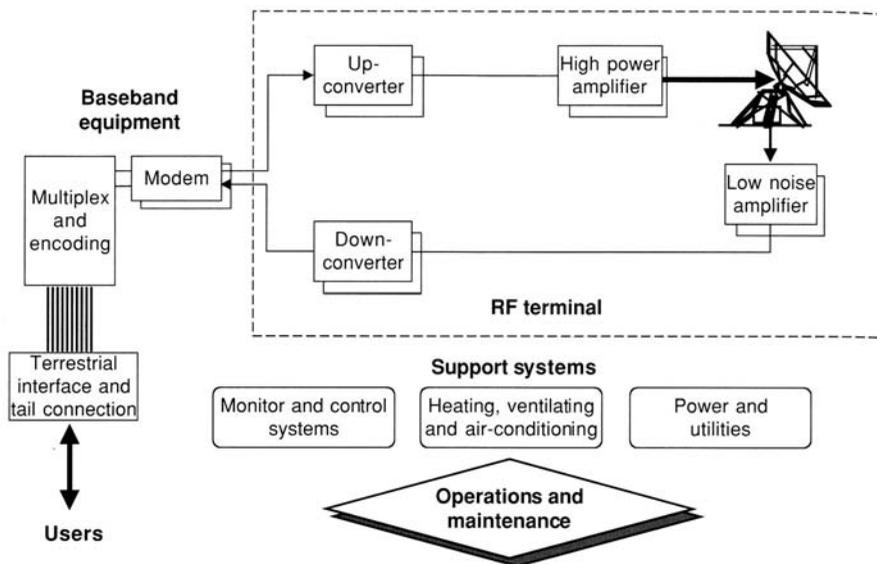
Figure 12.6 Typical Earth station configuration that identifies the major elements of investment and operations expense. Not shown is the physical building or other shelter arrangement which likewise contribute to investment and operations expense.

cost of the Earth station. In Figure 9.11, it is shown that the power output of the HPA and the gain of the transmitting antenna combine with each other to produce the specified value of EIRP. Figure 9.4 presents receive G/T as a function of the antenna size and LNA noise temperature.

The relationship of performance to cost of those elements is highly nonlinear and should be considered on a case-by-case basis using the most recent information from equipment suppliers. For example, the investment cost of an antenna increases at an increasing rate with diameter. Doubling the diameter from 3-m to 6-m increases cost not by a factor of 2 but by more like a factor of 6. A Ku-band power increase from 1W to 4W (within the typical range for low-cost SSPAs) may increase cost only by a factor of 2. But increasing from 10W to 20W could increase cost by a factor of 4. The use of a small antenna (with a relatively low value of gain) could demand a relatively high HPA power output. On the other hand, increasing the size of the antenna permits the use of a lower powered HPA at reduced cost. There typically is an optimum point where the sum of HPA cost and antenna cost is minimum. If that point occurs under the condition of a relatively large antenna, there will be the benefit of improved receive gain and G/T as well. A nonoptimum combination with a small antenna may be forced by the need to minimize the physical size of the installation due to local zoning restrictions or the constraints of a vehicle or aircraft. The upconverter and downconverter are essentially fixed cost items not dependent on the RF performance of the station (except when high-frequency accuracy or tracking ability are required).

A major factor in the cost of the RF terminal is the quantity of redundant electronic equipment and the degree of automation with which it can be controlled. In terms of overall network economics, the ultimate cost of outage of a critical service could far outweigh the initial cost of providing sufficient redundancy. Ignoring the antenna, a full set of redundancy (i.e., 100% or two-for-one backup) more than doubles the investment cost in electronics. That usually is justified in a major Earth station serving as a video uplink or hub in a network. A VSAT or TVRO would have a single string of RF components (without redundancy) under the assumption that failures are less likely with such low-power solid state electronics, coupled with the likelihood that alternative means of communication and/or rapid field service would be available in case of electronic failure.

12.5.1.2 Baseband Equipment

The modems, codecs, and compression and multiplex equipment that make up the baseband section of an Earth station often represent a significant portion of the investment and operation expense. Modern stations employ digital processing, notably TDMA and packet data, providing significant flexibility in the manner in which the station is used. On the other hand, digital technology is constantly evolving, and few systems have been standardized (a notable exception is the DVB standard, used for DTH television broadcasting). The result is that annual expenses for baseband equipment are second only to RF. Trained maintenance personnel must be close by to address technical problems when they occur. An Earth station may be designed in such a way that the RF terminal can be retained even if the baseband equipment is changed out. Also, baseband equipment, like computing

servers and other digital machines, tends to be sensitive to environmental conditions like high temperature and humidity. In VSATs or TVROs, on the other hand, part of the baseband portion may even be integrated into the RF terminal.

12.5.1.3   Terrestrial Interface and Tail

The cost of the tail is driven by the bandwidth characteristics and the distance covered. Other important factors include the type of technology (e.g., fiber optic or microwave radio), accessibility of installation sites, and cost of land or right-of-way. Minimum cost goes along with tying the station directly to the user, which is the concept behind the use of VSATs for private networks, TVROs for DTH service, and portable terminals of all types. On the other hand, a major Earth station can serve several users or user locations, requiring that local access or tail facilities be included. If an extended distance is involved, the choice must be made between renting the tail link from the local telephone company or installing a private link or bypass. An example of this type of trade off is presented in Figure 9.16.

12.5.1.4   Monitor and Control

The traditional method of monitoring network performance is to rely on users to complain when they experience problems. While that simple approach may have worked in the early days of satellite communications, the expectations of today's users cannot be met without continuous monitoring and extensive maintenance capabilities in the hands of the satellite operator or network service organization.

Having an automated system of monitor and control reduces or eliminates the need for on-site maintenance or operations personnel. That clearly is preferred in a large network where service delays and labor costs could be high. Such a specialized network management capability raises the initial investment in the network. If not incorporated in the beginning, an automated capability is extremely costly to add because the electronic equipment at remote sites might not have been designed for remote control and monitoring. Personnel at the centralized control site need to be trained in network management. They should be able to dispatch resources to expedite corrective actions once problems have been uncovered and isolated to a particular location or piece of equipment.

12.5.2   Annual Costs for Earth Station Networks

Two areas of annual expenses for Earth stations are shown in Figure 12.6. *Power and utilities* refers to the cost of operating the building that houses the Earth station electronics and personnel (if appropriate). That particular expense is significant in the case of a major Earth station. The other annual cost activity is *operations and maintenance* (O&M). Reliability of solid state electronics continues to improve even with the trend toward greater complexity. Incorporating the M&C capabilities often provides the means to detect and deal with problems before they affect the network services.

While the cost of O&M for a single station may be manageable, the issue of supporting a nationwide network of hundreds or even thousands of remote sites, major uplinks, hubs, and so on, can be particularly challenging. For that reason, the services of an existing maintenance organization are attractive, provided the service personnel can be trained to maintain the type of electronic equipment employed. Providing sufficient redundant units at the remote sites could eliminate the need for widespread service capability, relying instead on returning failed units to the manufacturer or a centralized depot. The capital cost of spares would need to be considered.

12.5.3   Teleport Earth Stations

Earth stations to uplink video or transmit point-to-point digital wideband links can be rented on a full-time or part-time basis. That basically is the concept of a teleport, which is a telecommunications facility intended to serve multiple occupants with Earth station antennas and associated indoor electronics. Many teleports are owned and operated by existing satellite communications service providers and international telephone carriers, often designed to provide public network services. A teleport can be managed as an independent business; alternatively, the teleport is an industrial development that sells or rents floor space in a building to operators that find the arrangement attractive. An example of the former is shown in Figure 12.7. As many readers can tell, there is no precise definition for a teleport, yet the term and business arrangements have persisted for many years.

A teleport service represents the "buy" alternative of the make-or-buy decision. Basically, a user of satellite capacity has the choice to construct and manage a dedicated Earth station or to arrange for capacity from a teleport owner/operator. The option of using a teleport may or may not exist—it depends on whether there is an available teleport that is suitable to the purpose. A user, such as a cable TV network, may decide to locate its studio and uplink in a particular city with an
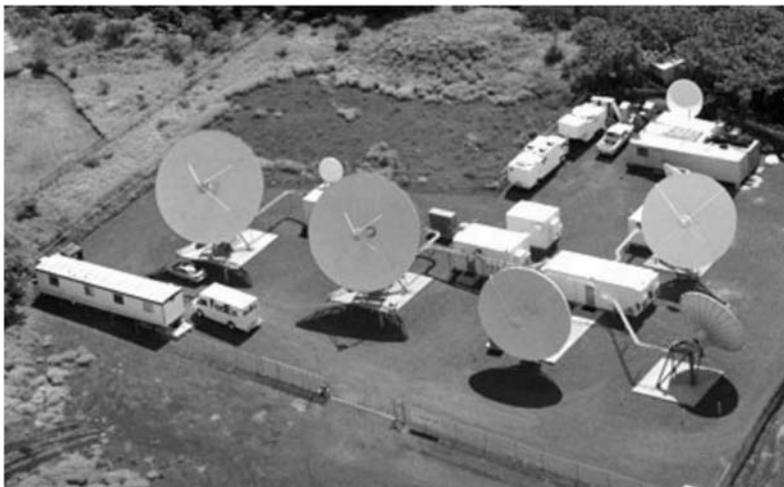
Figure 12.7   An example of a teleport serving multiple customers. (*Source:* Courtesy of Hawaii Pacific Teleport.)

excellent teleport. Evaluating the economics of the choices is a complex matter because the user must consider every significant cost, both initial and annual. The ideal situation is one in which the teleport operator has most or all of the necessary equipment and is willing to make an attractive deal at or near the marginal cost of serving the particular user.

12.6   Analysis of Network Economics

The overall economics of a satellite network can be analyzed by building as complete a financial model as possible and evaluating the costs as a function of various performance parameters. That can use the techniques of system analysis and financial modeling, aided by computerized tools. It is particularly convenient to develop the model on a personal computer using conventional spreadsheet software. The model can be as simple or as complicated as desired. Once the model has been entered into the computer, it is relatively simple to change some aspect of the network and observe the results. Sensitivity studies are useful to identify the most critical parameters. More sophisticated models can include technical parameters, such as EIRP or G/T of the Earth stations, transforming them into element costs.

A basic financial model uses simple arithmetic to estimate system investment cost and annual expense for a specific configuration. More complex modeling incorporates mathematical search and optimization algorithms that look at many possible network arrangements for satisfying a variety of traffic requirements. A particular network design then is selected for detailed analysis and specification, as outlined in Table 12.2. The steps involve gathering traffic data, formulating network models, and evaluating alternatives for network implementation.

12.6.1   Determining Traffic Requirements

One of the most difficult tasks facing the system designer is determining the expected usage (and ultimately the revenue) of the project before it goes into operation, which is the first step in Figure 12.2 and Table 12.2. The term *traffic* actually refers to the quantity of information flow and its distribution over time. For example, telephone traffic is expressed in call attempts and call durations of telephone conversations, measured in minutes. Existing telephone switching equipment can measure telephone traffic in a terrestrial network currently providing service, yielding the accurate traffic data needed to design a new satellite network. A totally new telephone network via satellite may not have such benefit, so estimates must be made through other research techniques (not, one would hope, including guess work). In years past, telephone traffic followed repeatable patterns and could be estimated using a set of statistical measures developed in the early part of the twentieth century. Later, telephone networks would deliver fax messages and be used to connect to the Internet on a dialup basis. That greatly complicated the problem of designing a telephone network with precision. What usually is done is to design using conventional traffic engineering principles and then to provide significant extra capacity to deal with peak loading situations.

Table 12.2 Outline of an Analysis Procedure for Determining the Economic Performance of a Satellite Network

I. Determine traffic requirements
   A. Collect current usage data (applications, data rates)
   B. Map the network
   C. Determine bandwidth requirements
      a. Video (full motion, teleconferencing)
      b. Voice (voice quality, ISDN vs. VoIP)
      c. Data (TCP/IP, ATM, file transfer, streaming)
II. Model Earth station elements
   A. Determine station categories
      a. Major Earth station
      b. VSAT
      c. TVRO
   B. Formulate Earth station cost models
III. Evaluate the total network
   A. Configure Earth stations
      a. Install traffic capabilities
      b. Determine factors such as redundancy
   B. Perform tradeoffs for network
      a. Sum total costs
      b. Test sensitivity
      c. Consider space segment
      d. Impact of traffic growth

For video distribution, the usage usually is predetermined by the TV programmer and can be specified as a requirement that is fixed for a period of time. A cable TV programming network operator would specify a quantity of full-time video distribution channels that provide a group of related programming formats. An example is a common network that must be delayed for different time zones and adapted to deliver some local advertising. Other types of video traffic are variable in nature, such as backhaul for news and sports events (discussed in Chapter 3). Those requirements could have a seasonal element (e.g., baseball and football) as well as a component that is entirely random (e.g., breaking news events and visits by heads of state). Another significant variable has been introduced as a result of digital compression and multiplexing of several video channels on one carrier. Compression standards, particularly MPEG 2 and 4, allow for compression and variable bit rate encoding; therefore, more video channels can be multiplexed on to one carrier.

Data communications cannot be treated like classical telephone traffic because there is wide variability of source information rates and their distribution over time (which is all too common in modern computer networking). Standard protocols, notably TCP/IP and ATM, have produced some uniformity in network data traffic, but patterns are still difficult to characterize in real-world situations. In the absence of real information, the best method for evaluating data traffic loading is to use computer simulation tools like Opnet and Simulink currently popular in terrestrial network design.

An example of the expected flow of traffic between the geographical nodes of a network is plotted in Figure 12.8. Primary locations such as corporate and division headquarters have large traffic requirements, justifying the use of a mesh of major
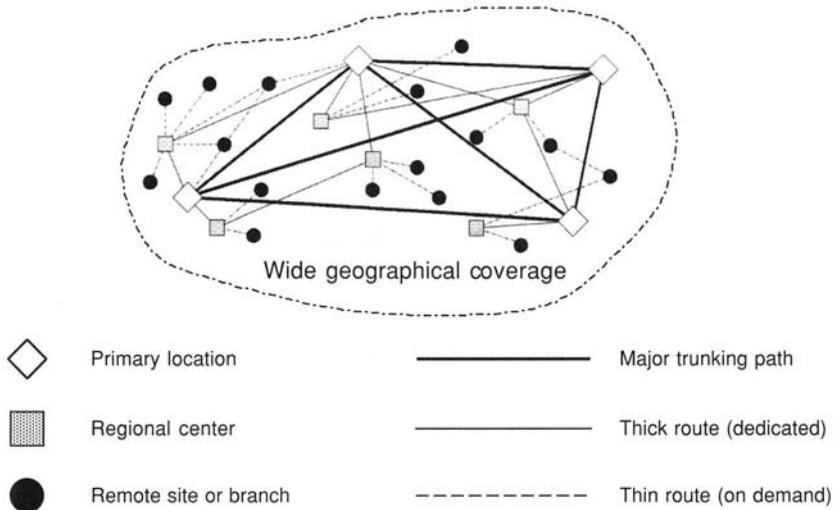
Figure 12.8 Interconnection of user locations in a private wide area communications network based on a terrestrial implementation.

trunking paths (shown as diamonds connected by heavy lines). Regional centers, indicated by squares, have significant requirements for bandwidth connections to the primary locations. The smallest predictable traffic requirements are between remote sites such as branch offices and customer locations. Those thin routes, which could number in the thousands, pass data and voice traffic to the regional centers and possibly the primary locations. Using terrestrial network technology, each location on the map is a node that can independently generate and accept traffic. For reasons of efficiency, it may be desirable to aggregate closely spaced nodes through a process called bundling. Many alternative arrangements of bundled nodes could satisfy the traffic requirements, but that must be analyzed with the overall economic model to yield the network configuration that is truly the most economical to implement. The disadvantage of bundling is that it increases the length and the quantity of local access lines and, in the case of a satellite network, the tails, since many user locations are some distance from the closest node.

The last step in determining the traffic requirements is to convert traffic flows into peak communications demand of the satellite system. This is converted into carrier bandwidths through the process of sizing. Flexible multiple access methods, like TDMA and CDMA, are adaptive in nature. In any case, there will be a specific quantity of interface ports to deliver the required video, voice, and/or data to the terrestrial interface. Devices that respond to the time-varying flow can be installed but there still is an upper limit to port capabilities. The model should allow the temporary selection of channel capacities and the testing of those assumptions to see if the projected traffic can be supported. Insufficient channel capacity between two nodes causes blockage, meaning that subsequent attempts to use the link will receive a busy signal or experience congestion. In a terrestrial version of the network, traffic from a remote site can reach a primary location by passing through a regional center node. That can pose a problem for a satellite network since such connections might involve a double hop. (The exception is for a LEO satellite system, in which

the delay of each satellite hop is considerably less than the GEO delay of 260 ms.) Direct point-to-point connections are possible over the satellite if the demand assignment feature is included in the network architecture. Otherwise, a quantity of terrestrial circuits should be retained as part of the backbone to prevent a double hop when a circuit connection must pass through another node.

12.6.2 Laying Out the Network

A priority in adapting satellite communications to an overall network design is the selection of various categories of stations and their capabilities. As suggested in Table 12.2, there might be three different classes of Earth stations in the network. A site would receive the type of station and carrier capacity that match the local traffic needs. Figure 12.9 indicates placement of three classes of Earth stations for the network previously discussed. In this first cut, the traffic has not been bundled, meaning there is one Earth station for each node in the network. Subsequent study of the network would consider reducing the number of major Earth stations by providing terrestrial tail circuits to extend service points or using a smaller class of Earth station. In countries or regions without a well-developed terrestrial network, bundling probably is not a feasible strategy for saving on investment costs. Rather, every site gets its own Earth station of a class to meet the local traffic needs.

Section 12.3 reviewed the main elements in the Earth station from an economic standpoint. Figures 12.10, 12.11, and 12.12 present the relative investment and annual costs for three examples: a major or hub Earth station, a VSAT, and a TVRO, respectively. The annual expenses for operations and maintenance and utilities have been assumed for a 5-year time period. Relative contributions are indicated as a percentage of the total station cost. The major station costs are
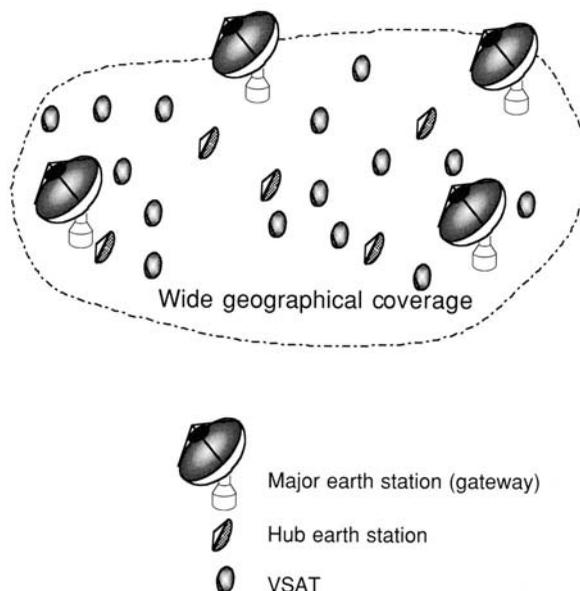
Figure 12.9 Configuration of a private satellite network to satisfy requirements for interconnectivity between points in a wide geographical region.
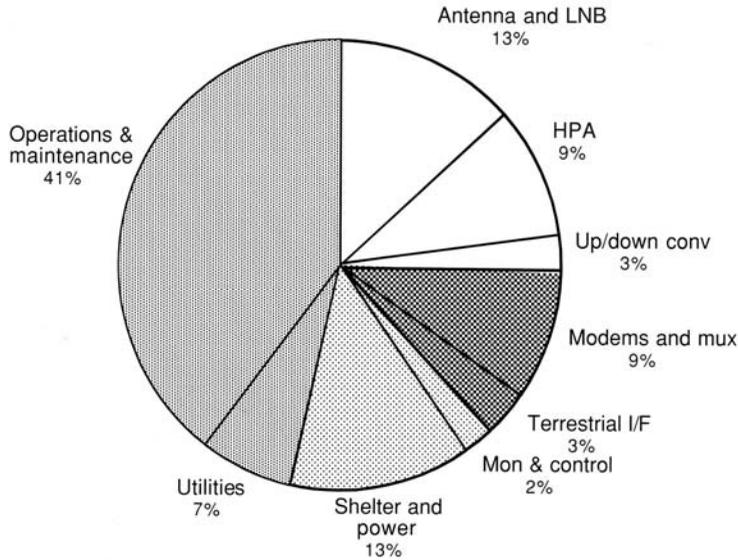
Figure 12.10   Allocation of investment and operating expenses among elements in a major or hub Earth station.
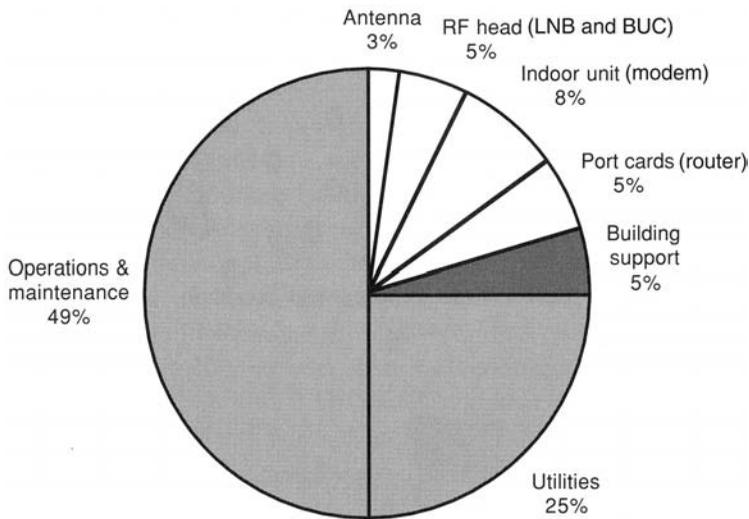
Figure 12.11   Allocation of investment and operating expenses among elements in a VSAT for interactive two-way communication.

dominated by the investment in electronics and facilities, with annual expenses being reasonably contained. A VSAT, on the other hand, requires more attention to its annual expenses, mainly because the electronics are not particularly costly and placement in an existing building is assumed. Many organizations convert VSAT investment cost to annual cost by leasing or renting the equipment. That is a good strategy when a using organization does not want to invest in equipment that they feel might become obsolete in a few years. Note that this breakdown does not include space segment costs.
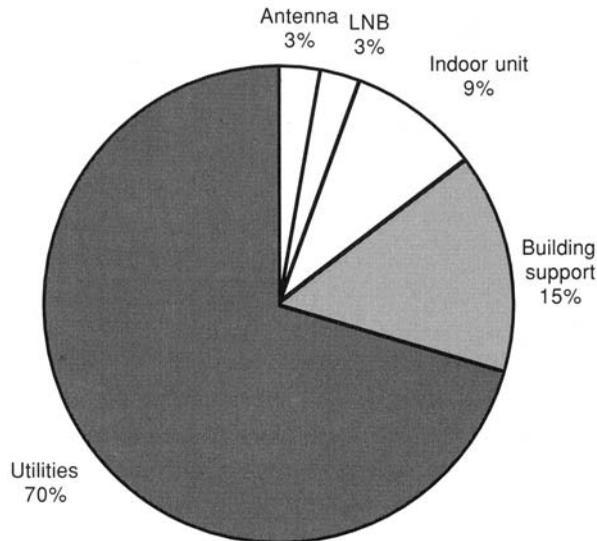
Figure 12.12   Allocation of investment and operating expenses among elements in a TVRO terminal for DTH service.

A TVRO station appears to represent an extreme case, in which the cost of power for the electronics over a 5-year period is comparable in cost to the initial expense for the antenna and LNB. The type of equipment used in digital DTH networks is simple to operate and maintain. Expenses for operations and maintenance are small because of the simplicity of the design of the TVRO and because repairs typically are accomplished by sending the failed unit back to the factory (or, more likely, discarding the unit and buying a new model). To compare the three stations with each other, the relative magnitude of total cost (i.e., the size of the individual pies) approximates 10,000 to 10 to 1 for the major station, VSAT, and TVRO, respectively.

12.6.3   Total Network Evaluation

In the last phase of the study, the constituents of the network traffic arrangement, Earth station deployment, and element cost are brought together for an integrated analysis. The stations are configured as nodes in the network, incorporating the channel capacities determined in the first phase. As indicated in Table 12.2, it is important that redundancy be included where appropriate. That can be accomplished by providing backup equipment in the individual Earth station or by adding alternative communication paths in the network.

The PC spreadsheet economic model can automatically sum up costs for the network and compute useful indices such as cost per transaction or per minute of service. In a commercial system, it is important to determine the projected profit (or loss) that would result under various assumptions for system traffic loading and pricing to users. Various tests of sensitivity to key parameters can be made after the accuracy of the model has been established. It is common in investment analysis to evaluate the internal rate of return (IRR) for a nominal case and then

to examine the change in IRR for different values of important parameters. Examples of such parameters include:

- Quantity of subscribers;
- Service usage per subscriber (minutes per month);
- Take-up rate (rate of increase in usage of the service) and churn (rate of customer turnover);
- Charge per month and per increment of usage;
- Market segments and penetration rates in each segment;
- Increase of 10% in the total investment in space or ground;
- Decrease in price per unit due to competition;
- Debt-to-equity ratio;
- Interest rate applied to debt;
- Depreciation method.

12.6.4 Optimizing the Space and Ground Segments

As mentioned previously, it is desirable to bring the technical and economic aspects of the space segment into the analysis. One of the more interesting tradeoffs is the determination of the optimum Earth station G/T. The results of such a tradeoff for TVRO service using the Ku-band BSS are presented in Figure 12.13, where the underlying assumptions are indicated at the lower right. The cost of a certain quantity of satellite bandwidth usually depends on the amount of RF power (EIRP) delivered in that bandwidth. In this example, it is assumed that the total weight and cost of the satellite are fixed but that the power per transponder and number
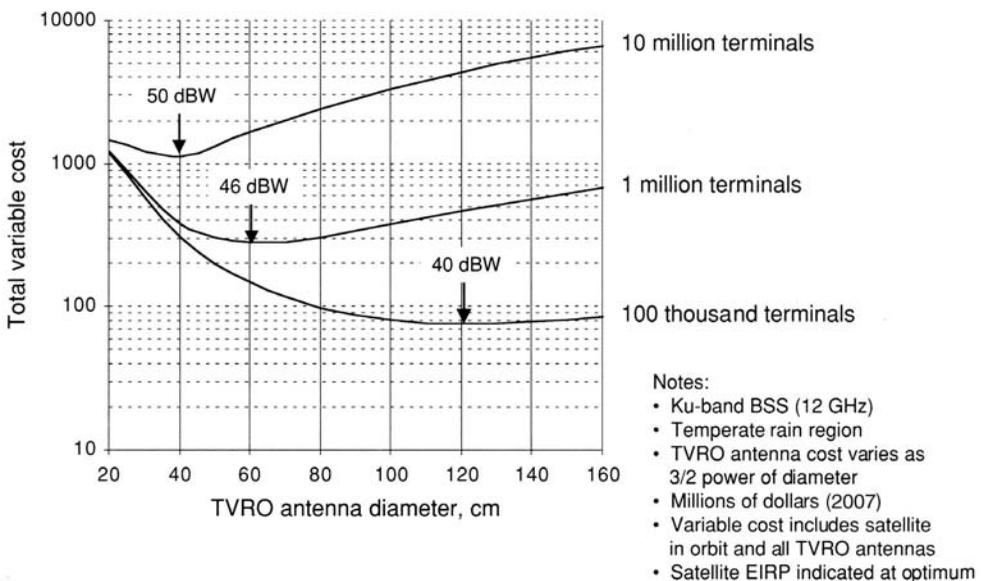
Figure 12.13 Determination of the optimum antenna size for a DTH TVRO terminal in the Ku-band BSS.

of transponders are varied (i.e., doubling of TWT output power halves the total number of transponders on the satellite). The benefit of more power is that the diameter of the receiving Earth station antenna can be reduced, saving ground total investment cost. That cost could be allocated either to the users (who must purchase their own DTH receivers) or retained by the network operator (who provides the DTH receiver free with the service). For a fixed system noise temperature such as 100K, the Earth station G/T is determined entirely by antenna gain. Consequently, the tradeoff is presented in terms of receiving dish size.

There will be a combination of satellite power and antenna diameter that minimizes the sum of the costs of those two elements for all stations in the network. It usually turns out that the more Earth stations there are in the network, the lower is the optimum antenna size (corresponding to a higher value of satellite EIRP). In terms of a nominal design point for the satellite, the number of transponders is 16 and their individual power is 200W, corresponding to a typical BSS satellite design with a per-transponder EIRP of 50 dBW. Investment costs are estimated for the satellite in orbit and for the receiving dish without the indoor settop unit because only the RF element costs are affected by the tradeoff of dish size and satellite EIRP.

The curves in Figure 12.13 indicate that in the broadcasting application using a very high powered satellite transponder, the minimum cost occurs at approximately 1m for a system of 100,000 terminals, while the corresponding minimum for 1,000,000 antennas is approximately half that diameter. To allow the diameter to be decreased to around 40 cm, the satellite EIRP should increase from approximately 50 dBW, as indicated along the X-axis of Figure 12.13. The BSS approach with relatively wide spacing between satellites is favored in those results because of the small dish size used to achieve optimum. Minimum cost conditions in the figure are within flat bottoms of the respective curves, demonstrating that the sensitivity of the optimum to the selected antenna size is not particularly strong.

This particular tradeoff and optimization study has been simplified to some extent, and the results should not be taken as conclusive. However, the power of the technique is evident. To be able to perform studies of this type, it is essential that accurate performance and cost data be used and that the model be properly constructed. Planning of this type requires a considerable investment in time and talent, beginning with a basic model and extending and expanding it as new data are gathered. Furthermore, the results should be checked against other hypotheses and available real-world data. Review by systems engineers and operations experts will provide greater confidence in the models and what they are telling their developers.

12.7   Satellite Communications: Instant Infrastructure

The systems and services to contribute to the business side of satellite communications have grabbed the attention of the financial community [3]. Satellites at any altitude have the power to deliver applications over the widest area, creating an instant infrastructure with a predetermined economic life. The established FSS, BSS, and MSS applications derive excellent revenues for established operators. The

challenge for a new entrant is to develop a new niche, gain the spectrum and orbit resources, provide for market entry, and create and manage the instant infrastructure for maximum advantage.

The companies that will be successful are those that:

- Properly play the space game. They obtain the right orbit slots (GEO) or orbit resources (non-GEO) and frequency assignments. From there, the satellites must have high capability and utilize proven technology to reduce risk.
- Lead the ground game. They get the right partnerships with equipment suppliers, service providers, and distributors. To that is added the key aspect of market access and local licensing.

Success is not a given—the project must have adequate funding to implement the system and deliver services during the initial lean years. Furthermore, having satellites in orbit is not enough. The ground partnerships need money, too. Making all that happen in a coordinated way makes the industry unique.

12.7.1   Satellites Versus Fiber Optics

The versatility of satellites should allow them to fill important and vital needs even as fiber optic networks extend to the home. Likewise, mobile and fixed wireless systems cover much of the developed land areas, potentially reducing the need for satellites. At the same time that satellite applications are evolving in response to competition from terrestrial systems, the technology base also is undergoing evolutionary change. There is every reason to expect that the capability and versatility of satellites and Earth stations will continue to improve significantly over the coming years. That provides a technology push all over again, allowing system designers and operators to approach new applications and markets with powerful hardware and software capabilities. Lower costs from microminiaturization in user equipment and advanced satellite payloads have produced a low-cost satellite radio service, and new two-way interactive schemes are on the horizon. Other advances help maintain this medium as important and cost effective for decades to come. Not least of those is HDTV, which the terrestrial networks strain to address but which becomes a normal facet of modern digital DTH systems.

This picture provided an introduction to the following discussion of the lineup of valuable services that satellites will provide in decades ahead. We see that the near term is dominated by the technologies, systems, and applications that were under development just prior to 2000. At the same time that satellites are addressing new markets and services, the terrestrial networks are evolving and improving. Keep in mind that applications often appear out of nowhere and some of the more important ones to come have not even been thought of yet.

12.7.2   Emphasis on the Broadcast Feature

The point frequently made in previous chapters is that one of the satellite's principal advantages is its wide-area coverage capability. In broadcasting, the downlink

signal is available everywhere within the footprint. That capability will continue to be attractive for video, audio, and data delivery purposes. The decade of the 1990s has, in fact, seen the introduction of true DBS service, building on the foundation of medium- and high-power bent-pipe GEO satellites and digital compressing and multiplexing technologies. C-band DTH service offerings from cable TV programmers have given way to more attractive Ku-band approaches, which permit smaller antennas on the ground. The ability to deliver over 200 television channels from one orbit position is in place in many countries. Yet, DTH will become more competitive vis-à-vis terrestrial systems as DBS operators experiment with new media formats.

One exceptional aspect of DBS is its ability to reach a narrow segment of users dispersed throughout a nation. Programs in a particular foreign language such as Armenian or Mandarin Chinese can be sent directly to the home for convenient viewing by families. Likewise, custom training and specialized cultural events can be targeted to the individual rather than the community. That greatly simplifies the logistics of delivering programming that otherwise would have to come by way of the local cable TV system or UHF station, both of which may be unavailable due to limited channel capacity.

Another strong area for DBS is in the delivery of HDTV programming. The digital standards for HDTV broadcast have been adopted in the United States and Europe, yet the technology still appears to be bound by the reluctance of broadcasters in other regions to make the required investments. The DBS satellite can easily transmit the HDTV signal, which requires two to four times the compressed data as a standard definition signal.

The DBS platform, with its ability to deliver a high-speed digital multiplex, also provides an important opportunity to provide a variety of information services. Broadcast data perhaps is an underexploited capability of bent-pipe and processor satellites. The same dish that provides over 200 channels of compressed full-motion video is also an effective way to distribute multimedia accessible with a Web browser. An important extension of this medium comes from adding the return channel. The latter can be supplied at either Ku- or Ka-band, using a low-power transmitter from the user terminal.

12.7.3   Paralleling the Terrestrial Networks

The evolution of satellite applications shows that while satellite transmission was at first attractive for long-distance links, the situation has shifted significantly since the advent of fiber optics and digital wireless services like PCS. However, satellite links will continue to be useful as diverse and alternative routes. That is demonstrated by large private VSAT networks operated by companies like Wal*Mart and Chevron, and agencies like the civil aviation ministry of the People's Republic of China. Private networks employ satellites because of the flexibility of the services offered and the ability of the user to own the ground equipment and exercise total control of the network.

While it often is said that satellites and terrestrial services compete with each other in the global markets, it is better to recognize the complementary nature of the two media. At times, one will dominate a given application segment, such as

terrestrial cellular for mobile telephone and satellites for digital video broadcasting, both in the 1990s. The nature of that dominance and competition between terrestrial communications and satellite communications will change and shift from time to time. The first innovations in high-capacity long-distance transmission were in the terrestrial area. Those developments included microwave transmission and analog FDM. Satellites took off next, relying on the same technologies but greatly reducing the cost of transmission by eliminating the unnecessary microwave repeater sites. Satellites became well established and pioneered new technologies in digital processing and transmission, notably TDMA, digital echo cancellation, and video compression. That began to bleed over to the terrestrial side, providing the basis for advances in terrestrial communications such as digital switching and high-speed digital transmission. More recently, the obstacles to HDTV and digital TV broadcasting are being overcome through the innovation of digital DTH service, pioneered by DIRECTV. We are poised for greater expansion of satellite services by virtue of Ka-band projects like Spaceway and ViaSat. Quite logically, forthcoming innovations in both satellite and terrestrial technologies will aid in the redefinition of telecommunications services that can be provided by the next generation of both approaches.

While fiber optic and wireless networks are well established in the developed world, satellite communications will continue to play a complementary and therefore important role. Conceptually, a fiber optic network is a given country provides the backbone of transmission between major cities and user locations. The economics of fiber are very favorable as long as the fibers and digital transmission groups within them can be adequately loaded with paying traffic. An unloaded fiber optic cable is not attractive economically, compounded by the fact that once a link is installed between two points it cannot be deployed elsewhere. Wireless systems also depend on fixed infrastructure to provide access, switch calls and route data, and to integrate services with the global Internet and telephone network.

The obvious role of satellite transmission is to provide the flexibility that point-to-point cables and microwave links cannot. Thin route applications will always be attractive by satellite. A hybrid network of heavily loaded fiber and diverse satellite links can produce the lowest cost per call or download where service is to be provided on a widespread basis. The type of satellite that is called for would operate in the FSS portion of the Ku- or Ka-band, taking advantage of the ability to locate uplinks anywhere in the satellite footprint. A high-capacity satellite can achieve cost per call for thin route service that is competitive with conventional service over the public network, even with low-cost fiber optic transmission between major nodes. Another approach to applying satellite technology is through advanced MSS systems. The low cost of the MSS user terminal in these systems makes the service attractive on an investment cost basis, if not in terms of per-minute charges.

12.7.4 Mobility Enhancements

Communications with and between vehicles, ships, and aircraft have been an important facet of radio. More recently, advanced MSS satellites like Inmarsat 4 and

Thuraya offer broadband connections to compact devices. Future mobile satellite networks will draw from past experience, with the benefits of global connectivity being an important attribute. The ATC concept discussed in Section 12.1 has the potential of gaining a substantial market share since it blends broadband wireless service with a broad satellite footprint. This can perhaps be foreseen through the concept of critical mass in networks. Even with all the space and ground factors properly considered, a satellite communications network still can fail as a business. That, of course, is a complex subject and cannot be answered in general. There is, however, a principle that helps explain the timing of when (and if) a new telecommunications service will achieve success in the marketplace. Figure 12.14 presents a conceptual curve of how a network service is developed and taken up by subscribers. We see that in the initial years, only a small percentage of the general population (early adopters) subscribe to the service. Subsequently, more users come on line to experiment and learn two things: what information is available and who else is connected.

Experience has shown that as the subscriber base increases, the network becomes attractive to yet more users. If that works as expected, the network is its own self-fulfilling prophesy. User adoption increases rapidly at the point of inflection in the curve, which is where critical mass is achieved. Network subscriptions accelerate as more potential users sign on and communicate with others. In some networks, the attraction to the network is irresistible, as people find that they must connect either to carry out their business or to enjoy what has become a widely adopted source of information, communication, or entertainment (whatever the case may be).

We can certainly see how critical mass worked to the benefit of fax machines, cellular telephones, the Internet and satellite radio. On the other hand, those areas are subject to intense competition, and it is hard to find a lucrative niche. In some countries, cellular service providers are protected by monopolies and import restrictions. In a few cases, companies have found a means of protecting their position in a market that has reached critical mass. Examples include Microsoft in PC operating systems and personal productivity software (which are not strictly telecommunications services but act like them in some ways) and Qualcomm in CDMA technology, due to their early lead and patent protection.
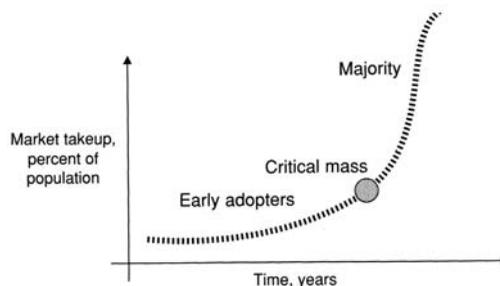
Figure 12.14 Conceptual service adoption curve, indicating the point where critical mass is achieved.

12.7.5 Creating the Future

These observations and projections of satellite communications suggest that there are many possibilities for new services and added competition for digital terrestrial networks. It will be up to the developers of satellite networks to make the decisions at the right time. That is not a simple matter because while it is easy to suggest options, it is much harder to construct a system with long-term economic viability. GEO satellites with bent-pipe repeaters are a known quantity and are capable of achieving very high rates of return. Accomplishing the same result for processing payloads and non-GEO constellations depends on many more factors, as suggested by projects like Iridium and Spaceway.

12.7.6 Advancing Technology

Designers of new generations of spacecraft as well as buyers of those products must keep in mind the substantial risks that exist when new technology is adopted quickly. Risk/benefit trades should be performed to ensure a proper level of technology maturity for a particular application. Important but immature technology cannot be guaranteed, so such a situation demands that one or more alternatives be identified early on to provide what could be called an escape route (a term not used in the industry, but one that nevertheless describes the view).

We can see many examples over the years of commercial and government satellites in which technology maturity was overestimated. The grandest example is the development of the Space Shuttle, which was put on a tight budget and followed what is termed a success-oriented schedule. The delivery of the first craft was delayed for years because many technology choices did not pan out, and alternatives had to be identified and developed. That added substantially to the cost of the program, which overran by a significant factor. This should not dissuade readers from considering new technology. There are numerous situations in which a newly qualified innovation provides big gains and a competitive advantage.

In nearly 50 years, the satellite communications industry has clearly come a long way. Once seen as a technical feat and a curiosity, the geostationary communications satellite is now commonplace and indispensable in many sectors. There has been a maturation process at work: first the technology had to make economical sense, then the applications for satellite communication had to prove themselves in a competitive marketplace. Clearly, those steps have been accomplished but more innovation is needed.

The technology of the satellite itself is also advancing, although the changes now tend to be less dramatic. Essentially, the applications now drive the design of the satellites to be launched in the future. It is not uncommon for a technology to be developed specifically to provide a certain capability required for an application.

Since the second edition of this work was published, several new systems went into operation. Iridium proved that one could provide a practical voice service throughout the globe using LEO satellites; it also demonstrated that commercial success is not assured and, in fact, is a substantial challenge. Satellite radio from XM and Sirius fared much better in the marketplace. From this point forward, we can expect GEO satellites to continue to dominate the communications realm, both from commercial and government perspectives. Innovation in ground-based

wireless and computer networks will continue to contribute to new application opportunities with satellites. It is more likely to be evolutionary rather than revolutionary; our challenge is to do the systems engineering job right and to play to the economics of satellites in ways that make sense and, not incidentally, money as well.

12.8   Conclusions for the Next Generation

Arthur C. Clarke, the father of the GEO satellite system concept, identified the challenge of looking ahead: "It's hard to predict the future." Looking back is relatively easy in an industry that is within the lifespan of those currently working in it. One purpose of this book is to provide that history and perspective for the generation to come.

The satellites under construction at the time of this writing seem like skyscrapers when compared to the "cottage-sized" satellites in the 1960s. These big birds have the power and capacity to justify the investment, and can be operated with relative ease. More satellites are carrying processors and complex antennas, and the higher frequency bands are now filling.

Compare this with the equivalent cellular network or fiber optic backbone, with their respective hundreds of critical points of presence and high maintenance expenses.

Looking ahead on the service side, it is comfortable to predict that communications satellites retain strength in two areas: broadcasting information across a wide region, and allowing remote communications access to isolated places (including mobile platforms). These uses are secure from terrestrial competition, except in heavily built-up areas such as cities and suburban communities. The economy of scale and critical mass aspects of satellite broadcasting provide for a bright future, as long as attractive content continues to flow. Also, users of data communications applications (which include voice and Internet access) expect to be able to carry on their business no matter where they happen to be. The coming generations of processor satellites could provide operating leverage with rising demand from a widely dispersed customer base.

Our industry has been transformed from a government-mandated and provided structure into one based on private equity. In recent years, in fact, the ownership has moved from a combination of telecommunications companies and stock corporations to private equity funds and firms. This has actually provided more capital to grow the business, and at the same time put expectations of near-term profitability. Perhaps these are in conflict, but they do impose some discipline on managers who need vital criteria when making decisions.

Another interesting trend is toward joint ventures and other combinations among competitors in satellite operation or satellite services. In 2007, XM Satellite Radio and Sirius Satellite Radio decided to merge. Rather than having two systems duplicating the same content, the merged company can adapt its service to new formats such as mobile TV. An innovative joint venture between Asia's largest satellite operator, JSAT Corporation, and Intelsat produced two Ku-band satellites

to serve North America. In 2007, SkyPerfect TV and JSAT Corporation merged to combine their operations and resources.

While satellites are getting larger and more capable, there is still room in years ahead for more focused vehicles that do a particular job very efficiently and for a lower launch investment. Orbital Sciences Corporation of the United States has positioned itself in this market, with both Horizons 2 and Intelsat 11 (shown in Figure 12.15) being good examples. While a large satellite like Anik F2 may have a lower investment cost per transponder, a smaller vehicle like Intelsat 11 could produce a higher return on investment percentage. The right approach, then, is to fully evaluate the technical and economic criteria for the satellite and system in question.

All of these actions indicate how the industry will continue to realign itself in the coming decade. Users of satellite systems will benefit from new service options and stability in terms current satellites and their replacements. Completely new approaches like ATC will develop at the same time as investors seek to exploit the proven technology and economics of constructing and operating communications satellites.
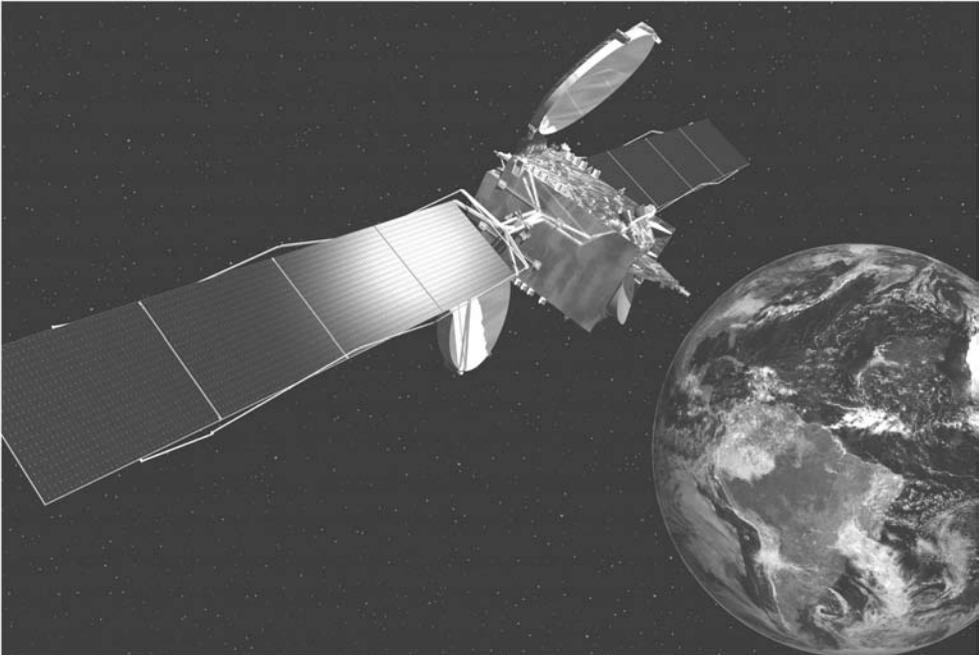
Figure 12.15 Artist's rendering of the Intelsat-11 satellite in geostationary orbit. Built by Orbital Sciences Corporation for Intelsat, Intelsat-11 carries two communications payloads: a C-band payload to provide Fixed Satellite Services (FSS) to the continental United States, Mexico and South America; and a Ku-band payload which serves DirecTV's Latin America downlink coverage in Brazil, with spot coverage over portions of Mexico, Venezuela, and Florida. The satellite was launched in 2007 and is located at 317 degrees east longitude. (*Source:* Orbital Science Corp.)

References

[1]   Jacoby, C., "Overview of Systems Engineering," Jacoby Consulting, Long Beach, CA, 2008.

[2]   Elbert, B. R., *The Satellite Communication Applications Handbook,* 2nd ed., Norwood, MA: Artech House, 2004.

[3]   Watts, T. W., and D. A. Freedman, *Satellite Communications—Instant Infrastructure,* Bear Stearns Equity Research, November 5, 1996.

About the Author

Bruce R. Elbert is a leading authority on satellites and telecommunications, having 35 years of working experience at his own consulting firm, Application Technology Strategy, Inc., and at Hughes, COMSAT, and the U.S. Army Signal Corps. He directed the design of several major satellite projects, including Palapa A, Indonesia's original satellite system; the Galaxy follow-on system, the largest and most successful C-band satellite TV system in the world; and the first GEO mobile satellite network to serve handheld terminals. By considering the technical, business, and operational aspects of satellite systems, Mr. Elbert has contributed to the economic success of several companies in the marketplace. In addition to an international industrial background, he is a prolific writer of books and articles, and has been featured as a speaker at numerous international conferences and seminars. Mr. Elbert is the editor of Artech House's Space Technology and Applications Series, and the Technology Management and Professional Development Series. At UCLA, he teaches satellite communications, information networking, and high-tech marketing to practicing engineers. Mr. Elbert holds BEE and MSEE degrees from CCNY and the University of Maryland, respectively, and an M.B.A. from Pepperdine University. He has two daughters, Sheri and Michelle, and resides with his wife, Cathy, in Thousand Oaks, California.

Index