Third Edition

Satellite Communications

Timothy Pratt • Jeremy Allnutt

WILEY

Satellite Communications

Satellite Communications

Timothy Pratt
Emeritus Professor of Electrical and Computer Engineering, Virginia Tech, Virginia, USA

Jeremy Allnutt
Emeritus Professor of Electrical and Computer Engineering, George Mason University Virginia, USA

Third Edition

WILEY

This edition first published 2020

Registered Office(s)
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data:

Names: Pratt, Timothy, author. | Allnutt, J. (Jeremy), author.
Title: Satellite communications / Timothy Pratt, Jeremy Allnutt.
Description: 3rd edition. | Hoboken, NJ : Wiley, 2020. | Includes bibliographical references and index. |
Identifiers: LCCN 2019015618 (print) | LCCN 2019018672 (ebook) | ISBN 9781119482147 (Adobe PDF) | ISBN 9781119482055 (ePub) | ISBN 9781119482178 (hardback)
Subjects: LCSH: Artificial satellites in telecommunication. | Artificial satellites in telecommunication–Problems, exercises, etc. | Telecommunication–Problems, exercises, etc.
Classification: LCC TK5104 (ebook) | LCC TK5104 .P725 2020 (print) | DDC 621.382/5–dc23
LC record available at https://lccn.loc.gov/2019015618

Cover Design: Wiley
Cover Image: © 2018 Intelsat, S.A. and its affiliates. All rights reserved.

Set in 10/12pt WarnockPro by Aptara Inc., New Delhi, India

10 9 8 7 6 5 4 3 2 1

This book is dedicated to our wives, Maggie and Norma, in gratitude for their love and support for over 50 years.

Contents

Preface

The first edition of *Satellite Communications* was published in 1986, with the second edition following in 2003. There have been many changes in the 33 years since the first edition appeared, with a complete transition from analog to digital communication systems. The launch of satellites, once the province of government agencies, is now a thriving commercial business. By the time this third edition reaches the market, a number of private citizens will have entered the lower reaches of space as tourists. Analog transmission techniques have been replaced by digital modulation and digital signal processing. Spinner satellites have virtually disappeared, replaced by a much wider range of satellites from cubesats with mass less than 1 kg to large GEO satellites with mass exceeding 6000 kg. While distribution of television programming remains the largest sector of commercial satellite communications, earning approximately half of the worldwide revenue from satellite communication systems, low earth orbit constellations of satellites for internet access are set to challenge that dominance.

Satellite communication systems have made a very significant contribution to world economics and society. An international telephone call that cost US$1 per minute in 1960 could be dialed for less than US$0.02 per minute in 2000. Taking account of inflation, the cost of communications has been reduced by a factor of more than 1000, a claim that very few other services can make. Access to the internet will become available to 3 billion people in countries that lack a terrestrial communication system as new constellations of LEO satellites are launched. Global satellite navigation systems help motorists to find their way to their destination and make travel by ships and aircraft safer. Two way television links via satellite enable news from anywhere in the world to be available 24 hours a day. Fiber optic systems have contributed significantly to these achievements, but satellite systems provide service wherever there is a need to broadcast to many locations. These contributions to quality of life have been made possible by the efforts of thousands of telecommunications engineers who design, produce, and maintain the systems that allow us to communicate with almost anyone, anywhere. Rarely do these engineers receive credit from the general public for these achievements.

In writing the third edition of *Satellite Communications* we have followed the intent of the first two editions; to provide a text that can be used in undergraduate and beginning graduate courses to introduce students to the subject, and also by engineers in industry and government to gain a sound understanding of how a satellite communication system works. The subject of satellite communications is extensive and we make no claims to have provided comprehensive coverage of the subject. An internet search for satellite communications yielded more than 250 000 entries in 2018, and there are textbooks available that expand on the topics of each of our individual chapters. In the

third edition, chapters 1–3 cover topics that are specific to satellites, including orbits, launchers, and spacecraft. Chapters 4–7 cover the principles of digital communication systems, radio frequency communications, digital modulation and multiple access techniques, and propagation in the earth's atmosphere, topics that are common to all radio communication systems. The chapter in the second edition on VSATs has been significantly expanded with the addition of low throughput satellite systems, otherwise known as SmallSats or cubeSats. These satellites range from experimental payloads assembled by undergraduate students to SmallSats that accompany advanced missions, such as the two that accompanied the Insight lander, which landed on Mars in late 2018. Also significantly expanded is the chapter dealing with rockets and launchers. Chapters 8–12 cover applications that include non-geostationary satellite systems, low throughput systems, direct broadcast satellite television, internet access by satellite, and global navigation satellite systems. The chapter on internet access by satellite is new to the third edition, and each of the chapters has been extensively revised to include the many changes in the field since 2003. Two new appendices have been added that cover digital transmission of analog signals, and antennas. These are topics that, in our experience, many students do not understand well yet are vital to most communication systems.

One of the most far reaching changes in communication systems technology has been the introduction of digital signal processing. High density integrated circuits are available that implement almost all of the functions required in transmitters and receivers in one or two devices. This is also true of spacecraft components such as three-axis control, which can now consist of a miniaturized digital controller rather than a group of two or three heavy momentum wheels. Liquid bi-propellant thrusters have been supplemented, and in some cases completely replaced by electric thrusters using xenon-ion propulsion systems.

Our text makes extensive use of block diagrams to explain how successive operations are performed on signals to obtain a specific result, for example, modulation of a digital signal onto an RF carrier or selection of a specific signal from a wide band multiplex of many signals. The blocks correspond to identifiable parts of a traditional analog system working in the frequency domain that could previously have been found in a transmitter or receiver, but are now part of a digital processor working in the time domain. Block diagrams are essential in understanding how communication systems are built up from successive operations on signals, but we recognize that in many cases the blocks are now implemented as digital operations.

The internet, and powerful search engines, have made it possible to find information about almost any subject in a few minutes. The reference section of the chapters of the third edition contain fewer references to papers and text books than previous editions and more references to internet sites. Although the specific sites may disappear with time, a search for the relevant topic will usually provide many alternative references. The internet has forced another change in the third edition. Our experience in teaching university courses has shown that the solutions to any problems issued to students for homework and exams appear very quickly at internet sites, and this is often the first place that students go to find answers, regardless of any rules that prohibit such action. As a result, we will not provide a solutions manual for the third edition. We have included exercises at the end of each chapter that instructors can use as the basis for homework problems, but our advice is to change the parameters of the questions each time one of the exercises is used. This forces students to work through the problem even if a similar internet solution is found, rather than just copying the solution. Changing

the first sentence of the question also makes it harder for students to find an internet solution.

The authors would like to thank their colleagues and students who, over the years, have made many valuable suggestions to improve this text. Their advice has been heeded and the third edition is the better for it. In particular, we want to acknowledge the contributions of Dr. Charles Bostian, Alumni Distinguished Professor Emeritus of Electrical and Computer Engineering, co-author of the first and second editions, who first suggested that we should write a book on satellite communications. Dr. Bostian's writing can be found in parts of several chapters of the third edition that cover the basic theory of satellite communications. Dr. Bostian founded the Satellite Communications Group at Virginia Tech and led research that has contributed significantly to the success of many satellite communications systems.

About the Authors

Timothy Pratt is an Emeritus Professor of the Bradley Department of Electrical and Computer Engineering at Virginia Tech, having retired in 2013. He received his B.Sc. and Ph.D. degrees in electrical engineering from the University of Birmingham, UK, and taught courses on satellite communications in the UK and the United States for 40 years. Dr. Pratt is a lifetime senior member of the IEEE. He lives on a farm outside Blacksburg with his wife and several dogs and cats, and many white tail deer.

Jeremy Allnutt is an Emeritus Professor of the Electrical and Computer Engineering Department of George Mason University, having retired in 2014. His primary interest is radiowave propagation effects on satellite links, which he pursued at research establishments in England and Canada, before working at INTELSAT in the US from 1979 to 1994. Prior to joining George Mason University in 2000, he was a professor at the University of York, UK, and at Virginia Tech. Dr. Allnutt obtained his B.Sc. and Ph.D. in Electrical Engineering from Salford University, UK, and is a Fellow of IET and a Fellow of the IEEE. He lives in Blacksburg with his wife, two dogs, two cats, and several birds, rabbits, and deer that consider his backyard to be their home as well.

1

Introduction

Two developments in the nineteenth and twentieth century changed the way people lived: the automobile and telecommunications. Prior to the widespread availability of personal automobiles, individuals had to travel on foot, by bicycle, or on horseback. Trains provided faster travel between cities, but most people's lives were centered on their home town and immediate surroundings. A journey of 100 miles was a major expedition for most people, and the easy mobility that we all take for granted in the twenty-first century was unknown. Before the telegraph and telephone came into widespread use, all communication was face to face, or in writing. If you wanted to talk to someone, you had to travel to meet with that person, and travel was slow and arduous. If you wanted to send information, it had to be written down and the papers hand-carried to their destination.

Telecommunication systems have now made it possible to communicate with virtually anyone at any time. Early telegraph and telephone systems used copper wire to carry signals over the earth's surface and across oceans, and high frequency (HF) radio made possible intercontinental telephone links.

The development and installation of optical fibers and optical transmission techniques has greatly increased the capacity of terrestrial and oceanic links. Artificial earth satellites have been used in communications systems for more than 50 years and have become an essential part of the world's telecommunications infrastructure. Satellites allow people to receive hundreds of television channels in their homes, either by receiving direct broadcast satellite television signals, or via cable TV from a satellite distribution center. Virtually all cable TV systems collect their signals from satellites that distribute television programming nationwide. Access to the internet via satellite from areas that are not served by cable is also available, providing many people in rural areas with much faster service than can be achieved over telephone lines.

1.1 Background

The origins of satellite communications can be traced to an article written by Arthur C. Clarke in the British radio magazine *Wireless World* in 1945 (Clarke 1945). At the time, Clarke was serving in the British Royal Air Force, working on precision approach radar systems that could guide World War II aircraft to a safe landing when the airport was fogged in. He was interested in long distance radio communication and was among the first to propose a practical way to communicate using satellites. He later became

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

famous as the author of *2001: A Space Odyssey*, and other science fiction books (Clarke 1968). In 1945, HF radio was the only available method for radio communication over transcontinental distances, and it was not at all reliable. Sun spots and ionospheric disturbances could disrupt HF radio links for days at a time. Telegraph cables had been laid across the oceans as early as the mid-1800s, but cables capable of carrying voice signals across the Atlantic did not begin service until 1953. Clarke suggested that a radio relay satellite in an equatorial orbit with a period of one sidereal day would remain stationary with respect to the earth's surface and make possible long distance radio links. (A sidereal day is the time it takes for the earth to make one complete revolution on its axis. It is 3 minutes 55.91 seconds shorter than a clock day of 24 hours, accounting for the progress of the earth around the sun in 365 days, which adds one additional revolution.)

Clarke's *Wireless World* paper is available on the internet and makes fascinating reading (Clarke 1945). Solar arrays had not been developed in 1945, so Clarke proposed a solar collector driving a steam engine to generate electrical power; a manned space station was needed to run the complicated systems. In most other respects, Clarke accurately predicted the development of geostationary earth orbit (GEO) satellites for direct broadcast television and data communications using transmitter powers much lower than the kilowatt levels of terrestrial broadcasting, and small parabolic mirrors (dishes) for receiving terminals.

At the time Clarke wrote his paper there were no satellites in orbit nor rockets powerful enough to launch them. But his ideas for what we now know as a geostationary satellite system were not science fiction, as the launch of the Russian satellite *Sputnik* in 1957 and subsequent GEO satellites was to prove. In 1965 the first geostationary communications satellite, *Early Bird*, began to provide telephone service across the Atlantic Ocean, fulfilling Clarke's vision of 20 years earlier. Intelsat launched a series of satellites between 1967 and 1969 that provided coverage of the Atlantic, Pacific, and Indian ocean regions, making worldwide coverage by GEO satellite possible, just in time for the Apollo 11 mission that first sent humans to the moon.

Satellite communication systems were originally developed to provide long distance telephone service. In the late 1960s, launch vehicles had been developed that could place a 500 kg satellite in geostationary earth orbit, with a capacity of 5000 telephone circuits, marking the start of an era of expansion for telecommunication satellites. Geostationary satellites were soon carrying transoceanic and transcontinental telephone calls. For the first time, live television links could be established across the Atlantic and Pacific oceans to carry news and sporting events. From its early beginnings in the 1960s, revenue earned from satellite communication systems has increased at an average of about 5% every year, and was valued at US$260B in 2016. Growth was rapid in the early 2000s, falling to 2% by 2016 (SIA 2017).

By year 2016, there were a total of 1459 active satellites in orbit with over 500 GEO communication satellites serving every part of the globe. Although television accounts for much of the traffic carried by these satellites, international and regional telephony, data transmission, and internet access are also important. In the populated parts of the world, the geostationary orbit is filled with satellites every two or three degrees, operating in almost every available frequency band. The global positioning system (GPS) uses 24 satellites in medium earth orbit (MEO) to provide worldwide navigation data for automobiles, ships, and aircraft. The worldwide revenue from Global Navigation Satellite Systems (GNSS) installations, mainly in automobiles, was US$74B in 2016 (SIA 2017).

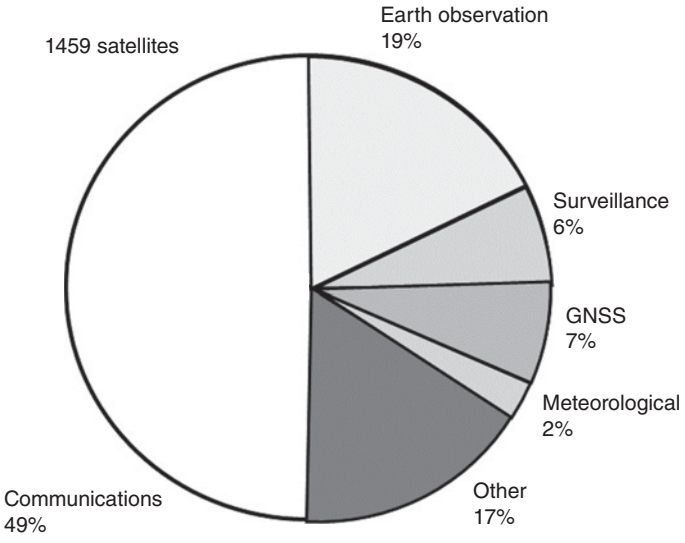


Figure 1.1 Distribution of satellites in orbit in 2016 by application. More than 500 satellites were in geostationary orbit. Communications includes DBS-TV, civil, and military links. Earth observation by small satellites increased quickly between 2014 and 2017 with the introduction of cubesats. Source: Adapted from data in (SIA 2017).

Figure 1.1 shows how the 1459 active satellites in orbit in 2016 were divided by application. Direct broadcast satellite television (DBS-TV) and video distribution services were the dominant uses of satellites, while navigation services made a major contribution. The large number of earth observation satellites were mainly cubesats.

Figure 1.2 shows the distribution of revenues generated by the worldwide satellite industry, divided by application. As in Figure 1.1, DBS-TV and video distribution generate more than half the revenue.

GEO satellites have grown steadily in mass, size, lifetime, and cost over the years. Some of the largest satellites launched to date are the KH and Lacrosse surveillance satellites of the US National Reconnaissance Office weighing an estimated 13 600 kg (30 000 lb) (KH-11_Kennen 2017). By 2000, commercial telecommunications satellites weighing 6000 kg with lifetimes of 15 years were being launched into geostationary orbit at a typical cost around US$125M for the satellite and launch. These costs did not change greatly over the following 15 years, although larger satellites with much higher capacity, and higher cost, have also been launched since 2011. The revenue earning capacity of a GEO satellite costing US$125M in orbit must exceed US$20M per year for the venture to be profitable, and must compete with optical fibers in carrying voice, data, and video signals. A single optical fiber can carry 10 Gbps at a single wavelength of light, and 100 Gbps by employing multiple wavelengths, a capacity similar to that of the largest GEO satellites, and optical fibers are never laid singly but always in bundles. The latest trans-Pacific optical fiber cable can transport 60 terabits per second using multiple optical fibers and optical wavelengths, equivalent to the capacity of 50 large GEO satellites in 2018 (The Verge 2017). GEO satellites cannot compete with optical fibers for point to point communications, but have the advantage of broadcasting to millions of receiving terminals simultaneously. Any place within the satellite coverage can be served by simply installing an earth terminal. To do the same with a fiber optic link requires fiber

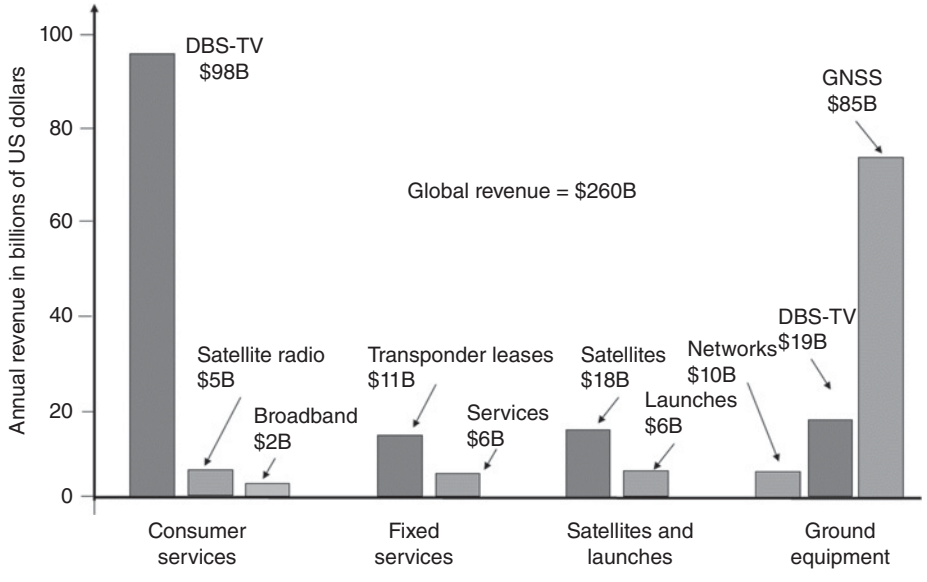


Figure 1.2 Distribution of global revenue earned from all satellite activity in 2016. Direct broadcast satellite television (DBS-TV) and Global Navigation Satellite Systems (GNSS) dominate with US$183B in revenue out of a total of US$261B. Source: Adapted from data in (SIA 2017).

to be laid. Fiber optic transmission systems dominate where there is a requirement for high capacity point-to-point links; GEO satellites succeed best when broadcasting.

The high capacity of both optical fibers and satellites, and the steady move of telecommunications traffic from analog signals to digital has lowered the cost of long distance telephone calls and increased enormously the number of circuits available. In 1960, prior to the advent of satellite communications, the United States had 550 overseas telephone circuits. Calls to Europe cost more than US$1.00 per minute at 1960 prices, and had to be placed through an operator, with delays of many hours being common. By 2016, virtually all international calls could be dialed by the end user, and rates to Europe had dropped to below US$0.02 per minute. To put the reduction in the cost of an international telephone call in perspective, we must remember that incomes have risen significantly over this time period. In the 1950s, the average wage in the United States was US$2.10 per hour, so the average worker would have had to work for 30 minutes to pay for a one minute call to Europe. In 2017, the average wage in the United States was US$26.10 per hour, and required less than 10 second's earnings to pay for the same international call. The United States now has hundreds of thousands of overseas telephone circuits, and video links daily carry live news reports from all over the globe. Texts and emails can be sent over the internet anywhere in the world for free. Telecommunications and computers lowered costs by a factor approaching 2000 between 1960 and 2010, something no other sector of the economy has ever achieved. The electrical and computer engineers who have made this possible rarely get the credit from the general public that they deserve.

GEO satellites have been supplemented by low and medium earth orbit satellites for some applications. Low earth orbit (LEO) satellites can provide satellite telephone and

data services over continents or the entire world, and are also used for earth imaging and surveillance. The delay incurred in a telephone link via a LEO satellite is much lower than with a GEO satellite, but because LEO satellites travel across the sky complicated handoff procedures are needed to ensure continuous communication. The dominance of GEO satellites for internet access by satellite will be challenged after 2020 as the proposed 12 000 LEO satellites operating in Ku-, Ka-, and V-band begin to provide worldwide access to the internet.

The global positioning system uses 24 medium earth orbit satellites to broadcast signals to the entire earth's surface. GPS, and Galileo, a similar European position location system have revolutionized navigation by vehicles, ships and aircraft, and GPS receivers have become a consumer product. Every cellular telephone has a GPS receiver built into it and cars are now available with built-in GPS receivers so that drivers should not get lost. Emergency calls from cellular phones carry information about the phone's location based on received GPS data.

1.2 A Brief History of Satellite Communications

Satellite communications began in October 1957 with the launch by the USSR of a small satellite called *Sputnik I*. This was the first artificial earth satellite, and it sparked the space race between the United States and the USSR. Sputnik I carried only a beacon transmitter and did not have communications capability, but demonstrated that satellites could be placed in orbit by powerful rockets. The first satellite successfully launched by the United States was Explorer I, launched from Cape Canaveral on 31 January 1958 on a Juno I rocket. The first voice heard from space was that of President Eisenhower, who recorded a brief Christmas message that was transmitted back to earth from the Project Score satellite in December 1958. The Score satellite was essentially the core of the Atlas intercontinental ballistic missile (ICBM) booster with a small payload in the nose. A tape recorder on Score had a storage capacity that allowed a four-minute message received from an earth station to be retransmitted. The batteries on Score failed after 35 days in orbit.

After some early attempts to use large balloons (Echo I and II) as passive reflectors for communication signals, and some small experimental satellite launches, the first true communications satellites, Telstar I and II, were launched in July 1962 and May 1963. The Telstar satellites were built by Bell Telephone Laboratories and used transponders adapted from terrestrial microwave link equipment. The uplink was at 6389 MHz and the downlink at 4169 MHz, with 50 MHz bandwidth. The satellites carried solar cells and batteries that allowed continuous use of the single transponder, and demonstrations of live television links and multiplexed telephone circuits were made across the Atlantic Ocean, emphatically demonstrating the feasibility of satellite communications.

The Telstar satellites were launched into what is now called a medium earth orbit, with periods of 158 and 225 minutes. This allowed transatlantic links to operate for about 20 minutes while the satellite was mutually visible. The orbits chosen for the Telstar satellites took them through several bands of high energy radiation, which caused early failure of the electronics on board. However, the value of communication satellites had been demonstrated and work was begun to develop launch vehicles that could deliver a payload to geostationary orbit, and to develop satellites that could provide useful communication capacity (Telstar 2018).

On 24 July 1961, US President John F. Kennedy defined the general guidelines of US policy in regard to satellite communications and made the first unambiguous references to a single worldwide system. On 20 December 1961, the US Congress recommended that the International Telecommunications Union (ITU) should examine the aspects of space communications for which international cooperation would be necessary. The most critical step was on 20 December 1961, when the US Congress passed the Communications Satellite Act. This set the stage for commercial investment in an international satellite organization and, on 19 July 1964, representatives of the first 12 countries to invest in what became Intelsat (the International Telecommunications Satellite Organization) signed an initial agreement. The company that represented the United States at this initial signing ceremony was Comsat, an entity Congress created to act for the United States within Intelsat. At this point the Bell System had a complete monopoly of all long distance telephone communications within the United States. When Congress passed the Communications Satellite Act, the Bell System was specifically barred from directly participating in satellite communications, although it was permitted to invest in Comsat.

Comsat essentially managed Intelsat in the formative years and should be credited with the remarkable success of the international venture. The first five Intelsat series of satellites (INTELSAT I through V) were selected, and their procurement managed, by teams put in place under Comsat leadership. Over this same phase, though, large portions of the Comsat engineering and operations groups transferred over to Intelsat so that, when the Permanent Management Arrangements came into force in 1979, many former Comsat groups were now part of Intelsat. Intelsat was eventually sold to private investors in 2001, and the proceeds divided up among the member counties. Intelsat was sold for US$3.1B in January 2005 to four private equity firms. The company acquired PanAmSat on 3 July 2006, and is now the world's largest provider of fixed satellite services, operating a fleet of 52 GEO satellites. In June 2007 BC Partners announced they had acquired 76% of Intelsat for about 3.75 billion euros (Intelsat 2017). (Fixed satellite services provide communications between earth stations that do not move, in contrast to mobile satellite services.)

In mid-1963, 99% of all satellites had been launched into LEO, and the higher MEO were much easier to reach than GEO with the small launchers available at that time. The intense debate was eventually settled on launcher reliability issues rather than on payload capabilities. The first six years of the so-called space age was a period of both payload and launcher development. The new frontier was very risky, with about one launch in four being fully successful. The system architecture of the first proposed commercial communications satellite system employed 12 satellites in an equatorial MEO constellation. Thus, with the launch failure rate at the time, 48 launches were envisioned to guarantee 12 operational satellites in orbit. Without 12 satellites in orbit, continuous 24-hour coverage could not be offered. Twenty-four hours a day, seven days a week – referred to as 24/7 operation – is a requirement for any successful communications service. A GEO systems architecture requires only one satellite to provide 24/7 operation over essentially one third of the inhabited world. On this basis, four launches would be required to achieve coverage of one third of the earth; 12 for the entire inhabited world. Despite its unproven technological approach, the geostationary orbit was selected by the entities that became Intelsat.

Launching satellites has become more reliable, with the best performance achieved by the United Launch Alliance (ULA) formed by Boeing and Lockheed-Martin. By 2016

ULA had conducted 164 consecutive satellite launches without a single failure. Newer entrants to the launch business offering much lower cost launches than ULA have been less successful, with occasional spectacular launch failures when the rocket exploded on the launch pad.

The first Intelsat satellite, INTELSAT I (formerly Early Bird) was launched on 16 April 1965. The satellite weighed a mere 36 kg (80 lb.) and incorporated two 6/4 GHz transponders, each with 25 MHz bandwidth. Commercial operations commenced between Europe and the United States on 28 June 28 1965. Thus, about two decades after Clarke's landmark article in *Wireless World*, GEO satellite communications began. Intelsat was highly successful and grew rapidly as many countries saw the value of improved telecommunications, not just internationally but for national systems that provided high quality satellite communications within the borders of large countries.

Canada was the first country to build a national telecommunication system using GEO satellites. ANIK 1A was launched in May 1974, just two months before the first US domestic satellite, WESTAR 1. The honor of the first regional satellite system, however, goes to the USSR Molniya system of highly elliptic orbit (HEO) satellites, the first of which was launched in April 1965 (the same month as INTELSAT I). Countries that are geographically spread like the former USSR, which covered 11 time zones, have used regional satellite systems very effectively. Another country that benefited greatly from a GEO regional system was Indonesia, which consists of more than 3000 islands spread out over more than a thousand miles. A terrestrially based telecommunication system was not economically feasible for these countries, while a single GEO satellite allowed instant communications region wide. Such ease of communications via GEO satellites proved to be very profitable. Within less than 10 years, Intelsat was self-supporting and, since it was not allowed to make a profit, it began returning substantial revenues to its Signatories. Within 25 years, Intelsat had more than 100 Signatories and, in early 2000, there were 143 member countries and Signatories that formed part of the international Intelsat community (Intelsat 2017).

The astonishing commercial success of Intelsat led many nations to invest in their own satellite systems, and by 2015 a total of 57 countries were operating one or more active satellites. Many of the original Intelsat Signatories had been privatized by the early 1990s and were, in effect, competing not only with each other in space communications, but with Intelsat. It was clear that some mechanism had to be found whereby Intelsat could be turned into a for-profit, private entity, which could then compete with other commercial organizations while still safeguarding the interests of the smaller nations that had come to depend upon the remarkably low cost communications cost that Intelsat offered. The first step in the move to privatizing Intelsat was the establishment of a commercial company called New Skies and the transfer of a number of Intelsat satellites to New Skies.

In the 1970s and 1980s there was rapid development of GEO satellite systems for international, regional, and domestic telephone traffic and video distribution. In the United States, the expansion of fiber optic links with very high capacity and low delay caused virtually all telephone traffic to move to terrestrial circuits by 1985. However, the demand for satellite systems grew steadily through this period, and the available spectrum in the 6/4 GHz band (C-band) was quickly occupied, leading to expansion into 14/11 GHz band (Ku-band). In the United States, most of the expansion after 1985 was in the areas of video distribution and very small aperture terminals (VSAT) networks. By 1995 it was clear that the GEO orbit capacity at Ku-band would soon be filled, and 30/20 GHz

(Ka-band) satellite systems would be needed to handle the expansion of digital traffic, especially wide band delivery of high speed internet data. Société Européenne de Satellites (SES), based in Luxemburg, began two way multimedia and internet access service in western and central Europe at Ka-band using the Astra 1H satellite in 2001 (SES Astra 2001). Direct to home satellite TV (DHS-TV), also called direct broadcast satellite TV continued to grow its customer base in the United States until 2016 when demand leveled off as subscription TV services became available on the internet.

In 2011 ViaSat launched ViaSat I, a Ka-band satellite with a digital data capacity of 140 Gbps, exceeding the combined capacity of all the Ku-band digital data satellites in orbit at that time (ViaSat I 2012). ViaSat 1 has 72 spot beams, 63 over the United States and 9 over Canada, and 56 Ka-band transponders. ViaSat 1 is intended to provide direct to home internet access, using a system marketed by Echostar as *Exede,* (later called *ViaSat)* over the populated areas of the United States and Canada. Part of the Rocky Mountain region has no spot beams because of low population density, and the Canadian beams are along the country's southern border with the United States. The satellite can also be used for DBS-TV. A similar satellite called Jupiter, later named Echostar 17, was launched by HughesNet for their internet access service. HughesNet became a subsidiary of EchoStar in 2011. The high capacity satellites provide internet access with downlink speeds up to 25 Mbps and uplinks at 3 Mbps, comparable to terrestrial cable speeds. More details of internet access by satellite can be found in Chapter 11.

The ability of satellite systems to provide communication with mobile users had long been recognized, and the International Maritime Satellite Organization (Inmarsat) has provided service to ships and aircraft for several decades, although at a high price. LEO satellites were seen as one way to create a satellite telephone system with worldwide coverage; numerous proposals were floated in the 1990s, with three LEO systems eventually reaching completion by 2000 (Iridium, Globalstar, and Orbcomm). The implementation of a LEO and MEO satellite system for mobile communication has proved much more costly than anticipated, and the capacity of the systems is relatively small compared to GEO satellite systems, leading to a higher cost per transmitted bit. Satellite telephone systems were unable to compete with cellular telephone because of the high cost and relatively low capacity of the space segment. The Iridium system, for example, cost over US$5B to implement, but provided a total capacity for the United States of fewer than 10 000 telephone circuits. Iridium Inc. declared bankruptcy in early 2000, having failed to establish a sufficiently large customer base to make the venture commercially viable. The entire Iridium system was sold to Iridium Satellite LLC for a reported US$25M, approximately 0.5% of the system's construction cost.

Satellite navigation systems, known generically as GNSS have revolutionized navigation and surveying. The global positioning system, created by the US Department of Defense (DoD) took almost 20 years to design and fully implement, at a cost of US$12B. By 2000, GPS receivers could be built in Original Equipment Manufacturer (OEM) form for less than US$25, and the worldwide GPS industry was earning billions of dollars from equipment sales and services. In the United States, aircraft navigation is transitioning to a GPS based system known as Automatic Dependent Surveillance Broadcast (ADS-B), which requires all aircraft operating under air traffic control to carry ADS-B equipment by 2020. ADS-B will replace radar as the main information source for air traffic control, although some radars will be retained for air defense and detection of aircraft without ADS-B capability. ADS-B transponders on Iridium satellites will eventually provide worldwide location of all commercial aircraft. Accurate navigation of ships, especially in

coastal waters and bad weather, is also heavily reliant on GPS. Europe has a comparable satellite navigation system called Galileo and China is building the Beidou system. GPS and ADS-B are the topics of Chapter 12.

1.3 Satellite Communications in 2018

Satellites come in many shapes and sizes. The smallest are cubesats, a low cost satellite that has a standard form called 1 U that is a 0.1 m cube with a maximum weight of 1 kg. Cubes can be joined together in one plane to make 2 U, 3 U … satellites. Cubes come with solar cells, batteries and options for microprocessors, plus standard software. The builder can add scientific experiments, communication systems and antennas, and other extras. Cubesats have proved popular with schools and universities because the basic satellite can be purchased for US$50 000 and a launch as a secondary payload can cost US$100 000 or less. Their development has at led to new ways of building satellites at much lower cost than large GEO satellites, and has spurred the creation of constellations of thousands of LEO satellites that can provide the entire world with internet access. Rockets that have payload space and weight to spare sometimes launch dozens of cubesats into low earth orbit (LEO). Chapter 8 discusses cubesats in more detail and Chapter 11 discusses large LEO satellite constellations.

GEO satellites were the backbone of the commercial satellite communications industry for 50 years. Large GEO satellites can serve one third of the earth's surface and can carry up to 140 Gbps of data, or transmit up to 200 high power DBS-TV signals. The weight and power of GEO satellites has increased. In 2018 a large GEO satellite could weigh 6000 kg (6 tons), generate 16 kW of power and carry 72 transponders, with a trend toward even higher powers but lower weight. Electrical propulsion systems for raising the satellite to GEO orbit and positioning over the satellite's lifetime avoid the need to carry fuel for gas jets; in a large GEO satellite fuel can account for half the initial weight of the satellite when it reaches orbit. Multiple beam antennas allow radio frequencies to be reused many times over, and also to transmit local TV channels from DBS-TV satellites. In between tiny cubesats and large GEO satellites is a range of satellites of medium size and weight that are used for earth observation and meteorology, scientific research, and communications using constellations of LEO or MEO satellites. The one common feature of all these satellites is that they require radio communication systems, the subject of this book.

Television program distribution and DBS-TV have become the major source of revenue for commercial satellite system operators, earning US$98B of the industry's US$122B revenues from communication services in 2016. By 2016 there were over 33 million DBS-TV customers in the United States and 230 million worldwide (SIA 2017). Direct to home satellite television and the distribution of video material to cable TV operators and broadcast stations has become the largest part by far of the satellite communication industry, at least until 2020. The next largest segment is position location systems, where almost all revenue comes from receiving equipment.

To achieve high capacity with a GEO satellite requires the use of high power terrestrial transmitters and high gain earth station antennas. Earth station antenna gain translates directly into communication capacity, and therefore into revenue. Increased capacity lowers the delivery cost per bit for a customer. Systems with fixed directional antennas can deliver bits at a significantly low cost than systems using low gain antennas, such as

those used on mobile terminals. Until the large LEO constellations for internet access come into use, GEO communications satellites will continue to be the largest revenue earners in space, along with the consumer GPS industry that supplies GPS chips and software for automobiles and cell phones.

Low earth orbit satellites are used for surveillance of the earth's surface. Civil uses, termed earth observation, include agricultural surveys to monitor growing crops, production of maps, weather observation, and surveys of archeological sites. Visible and infrared wavelengths yield different information, especially with vegetation. The resolution of commercially available earth observation data in 2017 was about 0.3 m. Military surveillance satellites have become an important part of the defensive capabilities of many countries, and are among the largest and heaviest satellites launched to date. These satellites are in very low earth orbits to obtain the highest possible resolution, and utilize visible and infrared wavelengths as well as radar observations. Infrared emissions have the advantage of being available during the night, whereas visible observations can only be made in daylight. The resolution achieved by military satellites is classified, but is undoubtedly much higher than that of civil earth observation satellites. In 2016 several proposals were approved by the US Federal Communications Commission (US FCC) for constellations of thousands of LEO satellites in low and very low earth orbit. These satellites operate in the Ka- and V-bands (18–50 GHz) providing internet access for homes anywhere in the world, especially in counties that lack a well developed terrestrial communication system. Once completed, these constellations will have many more satellites than all the satellites previously launched into orbit.

All radio systems require frequency spectrum, and the delivery of high speed data requires a wide bandwidth. Satellite communication systems started in C-band, with an allocation of 500 MHz, shared with terrestrial microwave links. As the GEO orbit filled up with satellites operating at C-band, satellites were built for the next available frequency band, Ku-band. Both C-band and Ku-band frequency allocations have been expanded over the years to increase the capacity of the GEO orbit, both by moving other services out of the satellite band, or adopting frequency sharing techniques.

There is a continuing demand for ever more spectrum to allow satellites to expand DBS-TV offerings and to provide new services, resulting in a move to Ka-band and even higher frequencies. Access to the internet from small transmitting Ka-band earth stations located at the home offers an alternative to terrestrial cable and telephone networks, especially in rural areas. SES began two way Ka-band internet access in Europe in 1998 with the Astra-K satellite, and ViaSat and Hughes Network Systems offer internet access through their *Exede* and *Hughesnet* systems in the United States, both now owned by Echostar (2017). Worldwide access to the internet via LEO and MEO satellite systems using Ka- and V-bands is also being developed.

Successive World Radio Conferences have allocated new frequency bands for commercial satellite services that now include L, S, C, Ku, K, Ka, V, and W bands. Table 1.1 gives the frequency designations for these letter bands. Letter bands were first used in World War II to obscure the frequencies of newly developed radars. By the end of the war there were seven different letter systems in use, and at least four systems covering radio communications, radar, and electronic warfare are still in widespread use. The frequency designations for letter bands for radio communication were eventually standardized by the IEEE (IEEE Std 521-2002 2012). Mobile satellite systems use very high frequency (VHF), ultra high frequency (UHF), L- and S-bands with carrier frequencies from 137 to 2500 MHz; GEO and LEO satellites use frequency bands extending from

Table 1.1 IEEE standard definitions for radio frequency bands [IEEE Std 521-2002]

Letter band	Frequency range
HF	3–30 MHz
VHF	30–300 MHz
UHF	300 MHz–1 GHz
L	1–2 GHz
S	2–4 GHz
C	4–8 GHz
X	8–12 GHz
Ku	12–18 GHz
K	18–27 GHz
Ka	27–40 GHz
V	40–75 GHz
W	75–110 GHz
mm wave	110–300 GHz

3.2 to 50 GHz. (VHF and UHF bands are defined by the ITU, along with super high frequency (SHF) and extremely high frequency (EHF), using the adjectives very high, ultra high, super high, and extra high, in decades of frequency. These designations are rarely used above 1 GHz, except by the ITU.)

Despite the growth of fiber optic links with very high capacity, the demand for satellite systems continues to increase. Satellites have also become integrated into complex communications architectures that use each element of the network to its best advantage. Examples are very small aperture terminals/wireless local loop (VSAT/WLL) in countries where the communications infrastructure is not yet mature and Local Multipoint Distribution Systems (GEO/LMDS) for the urban fringes of developed nations where the build-out of fiber has yet to be an economic proposition.

1.4 Overview of Satellite Communications

Satellite communication systems exist because the earth is a sphere. Radio waves travel in straight lines at the microwave frequencies used for wideband communications, so a repeater is needed to convey signals over long distances. Satellites, because they can link places on the earth that are thousands of miles apart are a good place to locate repeaters. A radio frequency repeater is simply a receiver linked to a transmitter, using different radio frequencies for transmit and receive, which can receive a signal from one earth station, amplify it, and retransmit it to another earth station. The repeater derives its name from nineteenth-century telegraph links, which had a maximum length of about 50 miles. Telegraph repeater stations were required every 50 miles in a long distance link so that the Morse code signals could be resent before they became too weak to read. Repeaters on satellites are called transponders.

In 2018, the majority of communication satellites were in geostationary earth orbit, at an altitude of 35 786 km, over the equator. A typical path length from an earth station

to a GEO satellite is 38 500 km. Because radio signals get weaker in proportion to the square of the distance traveled, signals reaching a satellite are always very weak. Similarly, signals received on earth from a satellite 38 500 km away are also very weak, because there are limits on the size of the antennas on GEO satellites and the electrical power they can generate using solar cells. The cost to place a geostationary satellite into orbit has fallen over the years as the number of launching options has increased. In 2018, the cost to launch a 8300 kg satellite into GEO varied from US$7600 to US$25 000 per kilogram, and a minimum of US$2700 per kilogram into LEO (Launch cost 2018). This obviously places severe restrictions on the size and weight of GEO satellites, since the high cost of building and launching a satellite must be recovered over a 10 to 15 year lifetime by selling communications capacity. LEO and MEO satellites cost less to launch, but an entire constellation of 12 to 66 satellites is needed to provide continuous coverage. In 2018 there were 160 proposals for LEO satellite systems for internet access using constellations as large as 12 000 satellites. Not all of these proposals will become working systems (Sweeting 2017).

Figure 1.3 illustrates some of the ways that satellites are used to provide communication services. In Figure 1.3a, a one way link is established between two earth stations via a single transponder on a GEO satellite. This configuration is used for the analysis of a satellite link, but not often in practice because two way communication is usually required. The transmission from earth station A to the satellite is called the uplink, and the transmission from the satellite to earth station B is called the downlink. In Figure 1.3b, the one way transmission is received by many receiving earth stations, sometimes as many as 30 million as in a DBS-TV system. DBS-TV satellites carry many transponders so that a large variety of video and audio channels can be sent to subscribers. In Figure 1.3c, a two way link is established through a single transponder. Earth station A transmits to the satellite at a frequency $f1$, which is transposed to a different frequency $f2$ by the transponder, so earth station B receives at a frequency $f2$. Earth station B transmits to the satellite at a frequency $f3$, which occupies a different part of the transponder bandwidth from earth station A's transmission at frequency $f1$. Earth station A receives signals from the satellite at a frequency $f4$. Using radio frequency to separate signals is known as frequency division multiplexing. An alternative technique is time division multiplexing, in which all transmitting stations share the same uplink frequency but transmit at different times such that their signals arrive at the satellite in sequence. All the receiving earth stations receive all the transmitted uplink signals and use time division techniques to extract the wanted signals.

Figure 1.3d illustrates a position location system such as GPS. GPS employs a constellation of 24 MEO satellites such that four satellites are always visible to a GPS receiver. The receiver compares the time of arrival of a spread spectrum sequence from each satellite and calculates the location of the receiver, and also the exact time referenced to atomic clocks on GPS satellites. All GPS receivers know time within one microsecond, which allows systems such a cellular telephones to be synchronized with great accuracy. Because a GPS receiver must simultaneously accept signals from different parts of the sky, an omnidirectional antenna is needed. Compared to the dish antennas used in DBS-TV, an omnidirectional antenna has a very low gain, so GPS signals are extremely weak.

Satellite communication systems are dominated by the need to receive very weak signals. In the early days, very large receiving antennas with diameters up to 30 m were needed to collect sufficient signal power to drive video signals or multiplexed telephone

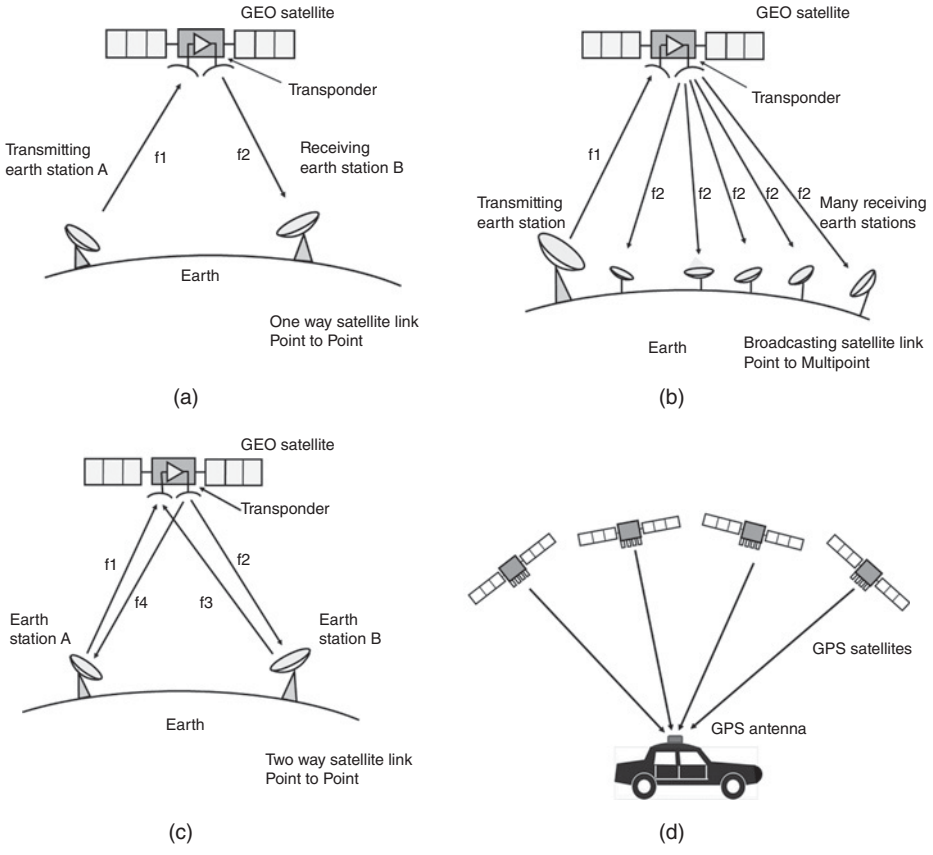


Figure 1.3 Illustration of different application of satellites. (a) One way satellite link from earth station A to earth station B. Uplink frequency is *f* 1, downlink frequency is *f* 2. (b) Point to multipoint link (broadcasting) from a single uplink transmitting station to many receiving stations. Uplink frequency is *f* 1 and all downlinks are at the same frequency *f* 2. (c) Two way connection between earth station A and earth station B. Station A transmits at frequency *f* 1 and receives at frequency *f* 4. Station B transmits at frequency *f* 3 and receives at frequency *f* 2. (d) Illustration of four GPS satellites broadcasting to an automobile. The GPS receiver uses an omnidirectional antenna.

channels. As satellites have become larger, heavier, and more powerful, smaller earth station antennas have become feasible, and DBS-TV receiving systems can use dish antennas as small as 0.5 m in diameter. Satellite systems operate in the microwave and millimeter wave frequency bands, using frequencies between 1 and 50 GHz. Above 10 GHz, rain causes significant attenuation of the signal and the probability that rain will occur in the path between the satellite and an earth station must be factored into the system design. Above 20 GHz, attenuation in heavy rain (usually associated with thunderstorms) can cause sufficient attenuation that the link will fail.

For the first 20 years of satellite communications, analog signals were widely used, with most links employing frequency modulation (FM). Wideband FM can operate at low carrier to noise ratios (CNRs), in the 5 to 10 dB range, but provides a signal to noise improvement so that video and telephone signals can be delivered with signal to noise ratios (SNRs) of 50 dB. The penalty for the SNR improvement is that the RF

signal occupies a much larger bandwidth than the baseband signal. In satellite links that penalty results because signals are always weak and the improvement in SNR is essential. Analog satellite communications is now obsolete for commercial use, although US amateur radio enthusiast still use FM voice links with their OSCAR series of experimental satellites.

Almost all communication signals are now digital – telephony, data, DBS-TV, radio and television broadcasting, and navigation with GPS all use digital signaling techniques. However, sound radio still uses amplitude modulation (AM) and FM analog transmissions for the majority of terrestrial radio broadcasting because of the enormous numbers of existing radio sets. All of the LEO and MEO mobile communication systems are digital, taking advantage of voice compression techniques that allow a digital voice signal to be compressed into a bit stream at 4.8 kbps. Similarly, the Motion Pictures Expert Group developed the MPEG-2 and MPEG-4 video compression techniques allowing video signals to be transmitted in full fidelity at rates less than 4 Mbps.

The most profitable application of satellite communications to date has been broadcasting. One GEO satellite can broadcast its signals to an entire continent, North America and Europe being typical examples. The population of the United States in 2017 was estimated to be 332 million people, in approximately 110 million households. DirecTV and Dish network together had 33 million subscribers to their DBS-TV transmissions, or nearly one third of all households in the United States. That is why Figure 1.2 shows that distribution of television programming is by far the largest revenue earner worldwide. However, this may change in the next decade if the proposed constellations of thousands of LEO satellites providing worldwide internet access are successfully completed.

The constellation of 24 GPS satellites is designed to provide continuous navigation services to every part of the earth. This is another example of satellite broadcasting, this time from a medium earth orbit. The manufacture and sale of GPS receivers represent 19% of the worldwide revenue from satellite communications systems. By comparison, satellites that provide links between individual users, as illustrated in Figure 1.3b, have a much smaller number of users and do not have the earning power of broadcasting satellites unless a worldwide constellation of thousands of LEO satellites is constructed. User terminals for LEO satellites need phased array antennas to track the satellites across the sky, at a much lower price than any such antennas available before 2017. A target price for the phased array antenna of US$200 is needed to make LEO internet access terminals available to a worldwide customer base. The challenge for satellite internet access systems is to serve a sufficiently large user base at a data rate and price that is comparable to other internet providers.

1.5 Summary

Satellite communication systems have become an essential part of the world's telecommunications infrastructure, serving billions of people with video, data, internet access, telephone, and navigation services. Despite the growth of fiber optic links, which have much greater capacity than satellite systems and a lower cost per bit, satellite systems continue to thrive and investment in new systems continues. Satellite services have shifted away from telephony to video and data delivery, with television broadcasting directly to the home emerging as one of the most powerful applications. GEO satellites

carried the majority of services in 2018, because the use of high gain fixed antennas at earth stations maximizes the capacity of the satellite. Over the years, there has been a trend away from trunk communications using very large earth station antennas toward delivery from more powerful satellites to individual users with much smaller antennas. VSAT networks using small antennas and low power transmitters are popular for linking together many locations in a single organization, such as retail stores and automobile dealerships. LEO and MEO satellites are used for mobile communications and navigation systems and, as the need for Geographic Information Systems grows with a variety of applications, LEO earth imaging satellites have the potential to provide strong revenue streams. Internet access by satellite is likely to be the largest sector of the industry by 2025.

1.6 Organization of This Book

Chapter 1 introduces the history of satellite communications and some of the ways that satellites are used.

Chapter 2 sets out the basics of satellite orbits and the factors that influence a satellite once in orbit. Calculation of look angles – where to look for a satellite in the sky – is restricted to GEO satellites. Software is needed to calculate look angles for LEO and MEO satellites as they move across the sky.

Chapter 3 describes the subsystems required to keep a communications satellite in orbit and functioning correctly. The communication system is covered in greater detail, including the organization of transponders, transmit and receive antennas, and use of frequency bands.

Chapter 4 covers the theory of radio communications and the calculation of carrier to noise ratio in a satellite link, and also low noise receiver design. The design of satellite links to achieve specific CNR values to maintain a particular error rate under conditions of atmospheric attenuation on the radio links is described in detail.

Chapter 5 concentrates on digital modulation methods and their performance in a satellite link. Forward error correction and error control methods used in DBS-TV and data links are discussed.

Chapter 6 covers multiple access techniques used in satellite communication systems to share the available resources of the satellite between multiple users. Frequency division multiple access, time division multiple access and code division multiple access (spread spectrum) are the main methods, with random access often used in the acquisition of a channel. Onboard processing and satellite switching techniques are discussed.

Chapter 7 explains the effect of the atmosphere on satellite-earth links, with rain being the most important. Techniques for the prediction of attenuation by clouds and rain are presented and methods for assessing the availability of satellite links are described. The RF frequency of a satellite link has a strong influence, as rain attenuation increases approximately as the square of frequency above 10 GHz.

Chapter 8 discusses some of the many low throughput applications for cubesats, mobile voice links, and VSAT networks. The commercial world of satellite communications is dominated by large geostationary satellites that can deliver tens or hundreds of gigabits per second, but there are many applications for small satellites with much lower throughput.

Chapter 9 describes the many orbits that are employed by satellite systems, generically grouped as non-geostationary (NGSO). These include LEO and MEO, which are becoming increasingly important for internet access via satellite.

Chapter 10 explains how direct broadcast satellite television and radio satellites provide hundreds of video and audio channels to subscribers. Techniques to mitigate the effect of rain on the RF path between the satellite and earth are described, and the probability of outages is discussed.

Chapter 11 is new to the third edition of *Satellite Communications*, covering the topic of internet access via satellite. This has become an important service for people in rural areas who are not served by cable companies or cellular telephone. Satellite internet access is also important in poorer countries where infrastructure is less well developed and satellite access can provide service over a wide area. Both GEO and LEO access systems are discussed and compared in terms of cost and capacity.

Chapter 12 covers satellite navigation systems, with emphasis on GPS. As discussed earlier in this chapter, GPS has become a major part of the satellite communications industry accounting for 19% of all revenue in 2016. The design of GPS receivers and the acquisition of GPS signals is covered in detail, and the system's vulnerability to jamming is discussed.

References

Clarke, A.C. (1945). Extra-terrestrial relays. *Wireless World* 1945: 305–308.

Clarke, A.C. (1968). *2001: A Space Odyssey*. UK: Arrow Books Ltd., Random House Publishing Group.

Echostar (2017). www.echostar.com (accessed 23 May 2018).

IEEE Std 521-2002 (2012). *Standard letter designations for radar-frequency bands*. https://ieeexplore.ieee.org/document/1160089/versions (accessed 23 May 2018).

Intelsat (2017). Intelsat 201. https://en.wikipedia.org/wiki/Intelsat (accessed 21 May 2018).

KH-11 Kennen (2017). https://en.wikipedia.org/wiki/KH-11_Kennen (accessed 23 May 2018).

Launch cost (2018.) https://en.wikipedia.org/wiki/Space_launch_market_competition (accessed 5 June 2018).

SES Astra (2001). https://en.wikipedia.org/wiki/Astra_(satellite) (accessed 5 June 2018).

SIA (2017). www.sia.org/wp-content/uploads/2017/07/SIA-SSIR-2017.pdf (accessed 21 May 2018).

Sweeting (2017). www.aerosociety.com/media/5612/2017-banquet-speech-by-sir-martin-sweeting-group-executive-chairman-sstl.pdf (accessed 23 May 2018).

Telstar (2018). https://en.wikipedia.org/wiki/Telstar (accessed 10 December 2018).

The Verge (2017). www.theverge.com/2017/9/25/16359966/microsoft-facebook-transatlantic-cable-160-terabits-a-second (accessed 21 May 2018).

ViaSat 1 (2012). https://en.wikipedia.org/wiki/ViaSat-1 (accessed 21 May 2018).

2

Orbital Mechanics and Launchers

Satellites are an integral part of our everyday life; so much so, we take many of their services for granted, such as global positioning system (GPS) navigation and pictures of weather systems taken from orbit. We will look at what constitutes a stable orbit first, and then find out how these orbits are achieved, the different types of orbits, and the various categories of launch vehicles that place them in orbit.

2.1 Introduction

Humankind has long sought to emulate birds in their seemingly effortless ability to fly wherever they choose. The flapping motion of avian wings was copied many times by early pioneers in an effort to fly, but without success. Attempts at gliding proved to be more fruitful, with early pioneers like Otto Lilienthal (Britannica 2018) achieving quite long flights. However, it was not until a propulsion system was added to a glider that true sustained flight was possible. The first to successfully demonstrate this were the Wright brothers in 1903, but each succeeding advance in aviation relied on a propulsive system that thrust against the surrounding air to move the aircraft. There is little atmosphere above 20 miles and so some other propulsive system must be found to achieve spaceflight, and this proved to be some form of rocket propulsion. There is a long history of rockets (NASA.gov 2018a,b), first demonstrated in 100 BCE by Hero (Wikipedia 2018a) using steam as a reaction force, and succeeding inventions have led to not just more powerful engines, but to new kinds of propulsion systems. In the sections that follow, we will find out how earth orbit is achieved, the laws that describe the motion of an object orbiting another body, how satellites maneuver in space, and the determination of the look angle to a satellite from the earth using ephemeris data that describe the orbital trajectory of the satellite.

2.2 Achieving a Stable Orbit

To achieve a stable orbit around the earth, a spacecraft must first be beyond the bulk of the earth's atmosphere; that is, in what is popularly called space. There are many definitions of space. US astronauts are awarded their "space wings" if they fly at an altitude that exceeds 50 miles (~80 km); some international treaties hold that the space frontier above a given country begins at a height of 100 miles (~160 km). Below

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

100 miles, permission must be sought to overfly any portion of the country in question. On re-entry, atmospheric drag starts to be felt at a height of about 400 000 ft. (~76 miles, ~122 km). Most satellites, for any mission of more than a few months, are placed into orbits of at least 250 miles (~400 km) above the earth. Even at this height, atmospheric drag is significant. As an example, the initial payload elements of the International Space Station (ISS) were injected into orbit at an altitude of 397 km when the shuttle mission left those modules on 9 June 1999. By the end of 1999, the orbital height had decayed to about 360 km, necessitating a maneuver to raise the orbit. Without onboard thrusters and sufficient orbital maneuvering fuel, the ISS would not last more than a few years at most in such a low orbit. To appreciate the basic laws that govern celestial mechanics, we will begin first with the fundamental Newtonian equations that describe the motion of a body. We will then give some coordinate axes within which the orbit of the satellite can be set and determine the various forces on the earth satellite.

Newton's laws of motion can be encapsulated into four equations:

$$s = ut + (1/2)at^2 \tag{2.1a}$$

$$v^2 = u^2 + 2at \tag{2.1b}$$

$$v = u + at \tag{2.1c}$$

$$P = ma \tag{2.1d}$$

where s is the distance traveled from time $t = 0$; u is the initial velocity of the object at time $t = 0$ and v the final velocity of the object at time t; a is the acceleration of the object; P is the force acting on the object; and m is the mass of the object. Note that the acceleration can be positive or negative, depending on the direction it is acting with respect to the velocity vector. Of these four equations, it is the last one that helps us understand the motion of a satellite in a stable orbit (neglecting any drag or other perturbing forces). Put into words, Eq. (2.1d) states that the force acting on a body is equal to the mass of the body multiplied by the resulting acceleration of the body. Alternatively, the resulting acceleration is the ratio of the force acting on the body to the mass of the body. Thus, for a given force, the lighter the mass of the body, the higher the acceleration will be. When in a stable orbit, there are two main forces acting on a satellite: a centrifugal force due to the kinetic energy of the satellite, which attempts to fling the satellite into a higher orbit, and a centripetal force due to the gravitational attraction of the planet about which the satellite is orbiting, which attempts to pull the satellite down toward the planet. If these two forces are equal, the satellite will remain in a stable orbit. It will continually fall toward the planet's surface as it moves forward in its orbit but, by virtue of its orbital velocity, it will have moved forward just far enough to compensate for the *fall* toward the planet and so it will remain at the same orbital height. This is why an object in a stable orbit is sometimes described as being in *free fall*. Figure 2.1 shows the two opposing forces on a satellite in a stable orbit.

Force = mass × acceleration and the unit of force is a Newton, with the notation *N*. A Newton is the force required to accelerate a mass of 1 kg with an acceleration of 1 m/s^2. The underlying units of a Newton are therefore (kg) × m/s^2. In Imperial Units, one Newton = 0.2248 ft. lb. The standard acceleration due to gravity at the earth's surface is $9.806\,65 \times 10^{-3}$ km/s^2, which is often quoted as 981 cm/s^2. This value decreases with

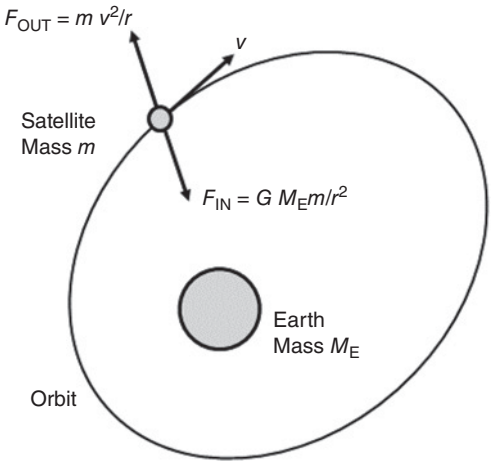


Figure 2.1 Forces acting on a satellite in a stable orbit around the earth. Gravitational force is inversely proportional to the square of the distance between the centers of gravity of the satellite and the planet the satellite is orbiting, in this case the earth. The gravitational force inward ($\boldsymbol{F}_{IN}$, the centripetal force) is directed toward the center of gravity of the earth. The kinetic energy of the satellite ($\boldsymbol{F}_{OUT}$, the centrifugal force) is directed diametrically opposite the gravitational force. Kinetic energy is proportional to the square of the velocity v of the satellite. When these inward and outward forces are balanced, the satellite moves around the earth in a *free fall* trajectory: the satellite's orbit. For a description of the units, please see the text.

height above the earth's surface. The acceleration, a, due to gravity at a distance r from the center of the earth is (Gordon and Morgan 1993)

$$a = \mu/r^2 \text{ km/s}^2 \tag{2.1e}$$

where the constant μ is the product of the universal gravitational constant G and the mass of the earth M_E.

The product GM_E is called Kepler's constant and has the value $3.986\,004\,418 \times 10^5 \text{ km}^3/\text{s}^2$.

The universal gravitational constant is

$$G = 6.672 \times 10^{-11} \text{ Nm}^2/kg^2 \text{ or } 6.672 \times 10^{-20} \text{ km}^3/\text{kg s}^2$$

in the older units. Since force = mass × acceleration, the centripetal force acting on the satellite, F_{IN}, is given by

$$F_{IN} = m \times (\mu/r^2) \tag{2.2a}$$

$$= m \times \left(GM_E/r^2\right) \tag{2.2b}$$

In a similar fashion, the centrifugal acceleration is given by

$$a = (v^2/r) \tag{2.3}$$

which will give the centrifugal force, F_{OUT}, as

$$F_{OUT} = m \times (v^2/r) \tag{2.4}$$

If the forces on the satellite are balanced, $F_{IN} = F_{OUT}$ and, using Eqs. (2.2a) and (2.4),

$$m \times \mu/r^2 = m \times v^2/r$$

hence the velocity v of a satellite in a circular orbit is given by

$$v = (\mu/r)^{1/2} \tag{2.5}$$

If the orbit is circular, the distance traveled by a satellite in one orbit around a planet is $2\pi r$, where r is the radius of the orbit from the satellite to the center of the planet.

Table 2.1 Orbital velocity, height, and period for five satellite systems

Satellite system	Orbital height (km)	Orbital velocity (km/s)	Orbital period (h)	(min)	(s)
Intelsat (GEO)	35 786.03	3.074 7	23	56	4.08
Other 3 billion (O3B) (MEO)	8 062	5.253 9	4	47	0.01
Globalstar (LEO)	1 414	7.152 2	1	54	5.35
Iridium (LEO)	780	7.462 4	1	40	27.0
SpaceX (VLEO)	345.6	7.699 51	1	31	26.90

Since distance divided by velocity equals time to travel that distance, the period of the satellite's orbit, T, will be

$$T = (2\pi r)/v = (2\pi r)/[(\mu/r)^{1/2}]$$

giving

$$T = (2\pi r^{3/2})/(\mu^{1/2}) \tag{2.6}$$

Table 2.1 gives the velocity, v, and orbital period, T, for four satellite systems that occupy typical low earth orbit (LEO), medium earth orbit (MEO), and geostationary earth orbit (GEO) orbits around the earth. In each case, the orbits are circular and the average radius of the earth is taken as 6378.137 km (Gordon and Morgan 1993).

Note the reduction in orbital period as the satellites move from GEO (essentially zero movement as observed from the ground) to very low earth orbit (VLEO). There are two immediate consequences for non-geostationary satellites as far as connections to a fixed earth station on the surface of the earth: (i) there will be gaps in coverage unless a constellation of the same satellites are orbiting, usually in the same plane; (ii) the observation time is significantly reduced as the orbital altitude is reduced. What is gained in lower signal delay with altitude is lost with the need for complex fixed earth station antennas and smaller observation time per satellite.

Example 2.1

Question: A satellite is in a 322 km high circular orbit. Determine:

a. The orbital angular velocity in radians per second;
b. The orbital period in minutes; and
c. The orbital velocity in meters per second.

Note: Assume the average radius of the earth is 6378.137 km and Kepler's constant has the value $3.986\,004\,418 \times 10^5$ km^3/s^2.

Answer

It is actually easier to answer the three parts of this question backward, beginning with the orbital velocity, then calculating the period, and hence the orbital angular velocity. First we will find the total radius of the orbit $r = 322 + 6378.137$ km $= 6700.137$ km

(c) From Eq. (2.5), the orbital velocity $v = (\mu/r)^{1/2} = (3.986\,004\,418 \times 10^5/6700.137)^{1/2}$ $= 7.713\,066$ km/s $= 7713.066$ m/s.

(b) From Eq. (2.6), $T = (2\pi r^{3/2})/(\mu^{1/2}) = (2\pi 6700.137^{3/2})/(3.986\,004\,418 \times 10^5)^{1/2} = (3\,445\,921.604)/(631.348\,114\,6) = 5\,458.037\,372$ seconds $= 90.967\,2895$ minutes $= 90.97$ minutes.

(a) The orbital period from above is 5 458.037 372 seconds. One revolution of the earth covers 360° or 2π radians. Hence 2π radians are covered in 5458.037 372 seconds, giving the orbital angular velocity as 2π/5458.037 372 rad/s = 0.001 1512 rad/s. An alternative calculation procedure would calculate the distance traveled in one orbit ($2\pi r = 2\pi 6700.137 = 42\,098.202\,36$ km). This distance is equivalent to 2π radians and so 1 km is equivalent to 2π/42 098.202 36 rad = 0.000 149 3 rad. From above, the orbital velocity was 7.713 066 km/s = 7.713 066 × 0.000 149 3 rad/s = 0.001 1512 rad/s.

A number of coordinate systems and reference planes can be used to describe the orbit of a satellite around a planet. Figure 2.2 illustrates one of these using a Cartesian coordinate system with the earth at the center and the reference planes coinciding with the equator and the polar axis. This is referred to as a geocentric coordinate system.

With the coordinate system set up as in Figure 2.2, and with the satellite mass m located at a vector distance r from the center of the earth, the gravitational force $\boldsymbol{F}$ on the satellite is given by

$$\overline{F} = -\frac{GM_E m\overline{r}}{r^3} \tag{2.7}$$

where M_E is the mass of the earth and $G = 6.672 \times 10^{-11}$ Nm^2/kg^2. But force = mass × acceleration and Eq. (2.7) can be written as

$$\overline{F} = m\frac{d^2\overline{r}}{dt^2} \tag{2.8}$$

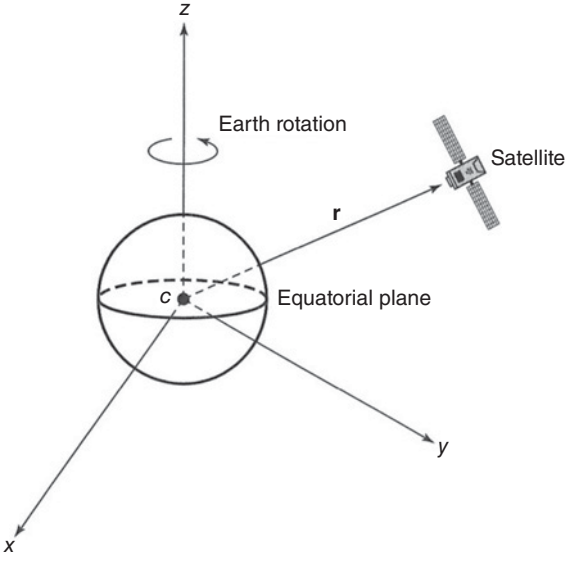


Figure 2.2 The initial coordinate system used to describe the relationship between the earth and a satellite. A Cartesian coordinate system with the geographical axes of the earth as the principal axes is the simplest coordinate system and the origin at the center of the earth. The rotational axis of the earth is about the *z* axis, which passes through the geographic north pole. The *x* and *y* axes are mutually orthogonal to the *z* axis and lie in the earth's equatorial plane. The vector *r* locates the satellite with respect to the center of the earth.

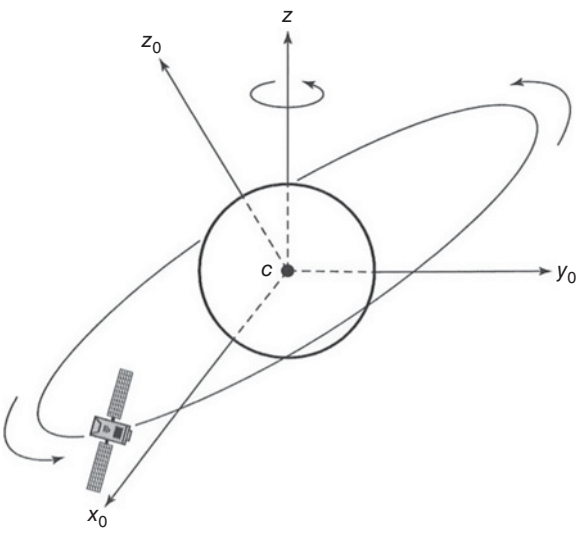


Figure 2.3 The orbital plane coordinate system. In this coordinate system the orbital plane is used as the reference plane. The orthogonal axes x_0 and y_0 lie in the orbital plane. The third axis, z_0 is orthogonal to the x_0 and y_0 axes to form a right hand coordinate set. The z_0 axis is not coincident with the earth's z axis through the earth's north pole unless the orbital plane lies exactly in the earth's equatorial plane.

From Eqs. (2.7) and (2.8) we have

$$-\frac{\bar{r}}{r^3}\mu = \frac{d^2\bar{r}}{dt^2} \tag{2.9}$$

which yields

$$\frac{d^2\bar{r}}{dt^2} + \frac{\bar{r}}{r^3}\mu = 0 \tag{2.10}$$

This is a second order linear differential equation and its solution will involve six undetermined constants called the *orbital elements*. The orbit described by these orbital elements can be shown to lie in a plane and to have a constant angular momentum. The solution to Eq. (2.10) is difficult since the second derivative of r involves the second derivative of the unit vector r. To remove this dependence, a different set of coordinates can be chosen to describe the location of the satellite such that the unit vectors in the three axes are constant. This coordinate system uses the plane of the satellite's orbit as the reference plane. This is shown in Figure 2.3.

Expressing Eq. (2.10) in terms of the new coordinate axes x_0, y_0, and z_0 gives

$$\hat{x}_0\left(\frac{d^2x_0}{dt^2}\right) + \hat{y}_0\left(\frac{d^2y_0}{dt^2}\right) + \frac{\mu\left(x_0\hat{x}_0 + y_0\hat{y}_0\right)}{\left(x_0^2 + y_0^2\right)^{3/2}} = 0 \tag{2.11}$$

Equation (2.11) is easier to solve if it is expressed in a polar coordinate system rather than a Cartesian coordinate system. The polar coordinate system is shown in Figure 2.4.

With the polar coordinate system shown in Figure 2.4 and using the transformations

$$x_0 = r_0\cos\phi_0 \tag{2.12a}$$

$$y_0 = r_0\sin\phi_0 \tag{2.12b}$$

$$\hat{x}_0 = \hat{r}_0\cos\phi_0 - \hat{\phi}_0\sin\phi_0 \tag{2.12c}$$

$$\hat{y}_0 = \hat{\phi}_0\cos\phi_0 + \hat{r}_0\sin\phi_0 \tag{2.12d}$$

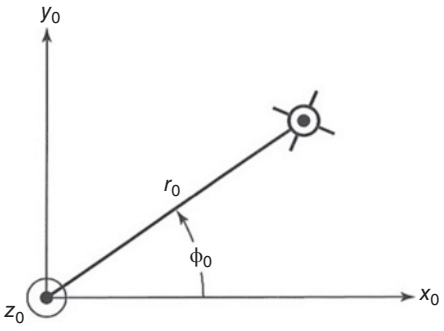


Figure 2.4 Polar coordinate system in the plane of the satellite's orbit. The axis z_0 is straight out of the paper from the center of the earth, and is normal to the plane of the satellite's orbit. The satellite's position is described in terms of the distance r_0 from the center of the earth and the angle this makes with the x_0 axis, ϕ_0.

and equating the vector components of r_0 and ϕ_0 in turn in Eq. (2.11) yields

$$\frac{d^2 r_0}{dt^2} - r_0 \left(\frac{d\phi_0}{dt} \right) = -\frac{\mu}{r_0^2} \tag{2.13}$$

and

$$r_0 \left(\frac{d^2 \phi_0}{dt^2} \right) + 2 \left(\frac{dr_0}{dt} \right) \left(\frac{d\phi_0}{dt} \right) = 0 \tag{2.14}$$

Using standard mathematical procedures, we can develop an equation for the radius of the satellite's orbit, r_0, namely

$$r_0 = \frac{p}{1 + e\cos(\phi_0 - \theta_0)} \tag{2.15}$$

where θ_0 is a constant and e is the eccentricity of an ellipse whose semilatus rectum p is given by

$$p = (h^2/\mu) \tag{2.16}$$

and h is the magnitude of the orbital angular momentum of the satellite. That the equation of the orbit is an ellipse is Kepler's first law of planetary motion.

2.3 Kepler's Three Laws of Planetary Motion

Johannes Kepler (1571–1630) was a German astronomer and scientist who developed his three laws of planetary motion by careful observations of the behavior of the planets in the solar system over many years, with help from some detailed planetary observations by the Hungarian astronomer Tycho Brahe. Kepler's three laws are:

1. The orbit of any smaller body about a larger body is always an ellipse, with the center of mass of the larger body as one of the two foci.
2. The orbit of the smaller body sweeps out equal areas in equal time (see Figure 2.5).
3. The square of the period of revolution of the smaller body about the larger body equals a constant multiplied by the third power of the semimajor axis of the orbital ellipse. That is $T^2 = (4\pi^2 a^3)/\mu$ where T is the orbital period, a is the semimajor axis of the orbital ellipse, and μ is Kepler's constant. If the orbit is circular, then a becomes distance r, defined as before, and we have Eq. (2.6).

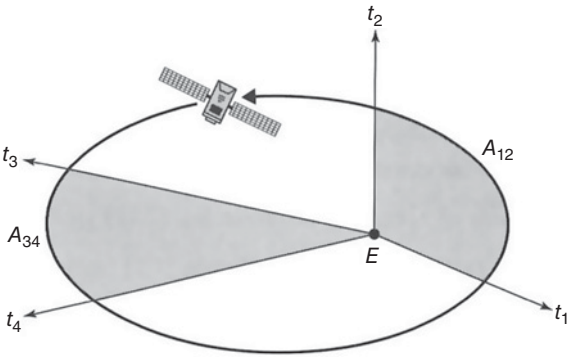


Figure 2.5 Illustration of Kepler's second law of planetary motion. A satellite is in orbit around the planet earth, *E*. The orbit is an ellipse with a relatively high eccentricity, that is, it is far from being circular. The figure shows two shaded portions of the elliptical plane in which the orbit moves, one is close to the earth and encloses the perigee while the other is far from the earth and encloses the apogee. The perigee is the point of closest approach to the earth while the apogee is the point in the orbit that is furthest from the earth. While close to perigee, the satellite moves in the orbit between t_1 and t_2 and sweeps out an area denoted by A_{12}. While close to apogee, the satellite moves in the orbit between times t_3 and t_4 and sweeps out an area denoted by A_{34}. If $t_1 - t_2 = t_3 - t_4$ then $A_{12} = A_{34}$.

Kepler's laws were subsequently confirmed, about 50 years later, by Isaac Newton, who developed a mathematical model for the motion of the planets. Newton was one of the first people to make use of differential calculus, and with his understanding of gravity, was able to describe the motion of planets from a mathematical model based on his laws of motion and the concept of gravitational attraction. The work was published in the *Philosophiae Naturalis Principia Mathematica* in 1687. At that time, Latin was the international language of formally educated people, much in the way English has become the international language of email and business today, so Newton's *Principia* was written in Latin.

Example 2.2

Question: A satellite in an elliptical orbit around the earth has an apogee of 39 152 km and a perigee of 500 km. What is the orbital period of this satellite? Give your answer in hours. Note: Assume the average radius of the earth is 6378.137 km and Kepler's constant has the value $3.986\,004\,418 \times 10^5\ \mathrm{km}^3/\mathrm{s}^2$.

Answer

The mathematical formulation of the third law is $T^2 = (4\pi^2 \mathrm{a}^3)/\mu$, where T is the orbital period, a is the semimajor axis of the orbital ellipse, and μ is Kepler's constant.

The perigee of a satellite is the closest distance in the orbit to the earth; the apogee of a satellite is the furthest distance in the orbit from the earth.

For the last part, draw a diagram to illustrate the geometry.

The semimajor axis of the ellipse = $(39\,152 + (2 \times 6378.137) + 500)/2 = 26\,204.137$ km

The orbital period is

$$T^2 = (4\pi^2 a^3)/\mu = (4\pi^2(26\,204.137)^3)/3.986\,004\,418 \times 10^5 = 1\,782\,097\,845.0\ \mathrm{s}^2$$

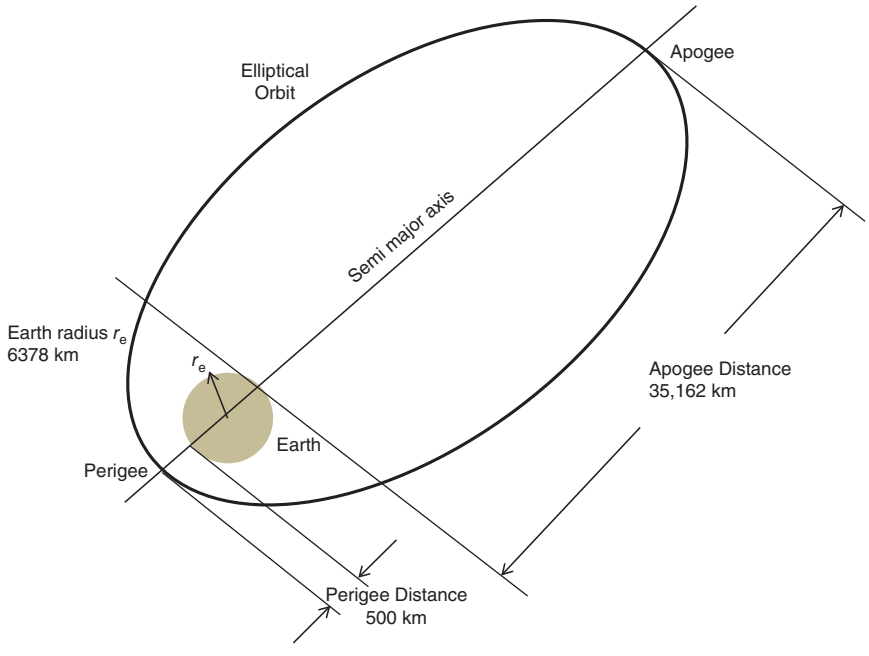


Figure Ex 2.2 The elliptical orbit of the satellite in Example 2.2. This is a Molniya orbit.

Therefore, $T = 42\,214.900\,75$ seconds $= 11.726\,361\,32$ hours $=$ (11 hours 43 minutes 34.9 seconds)

What we have found above is the orbital period of a *Molniya* satellite of the former Soviet Union as shown in Figure Ex 2.2. Describing the orbit of a satellite enables us to develop Kepler's second two laws.

2.4 Describing the Orbit of a Satellite

The quantity θ_0 in Eq. (2.15) serves to orient the ellipse with respect to the orbital plane axes x_0 and y_0. Now that we know that the orbit is an ellipse, we can always choose x_0 and y_0 so that θ_0 is zero. We will assume that this has been done for the rest of this discussion. This now gives the equation of the orbit as

$$r_0 = \frac{p}{1 + e\cos\phi_0} \tag{2.17}$$

The path of the satellite in the orbital plane is shown in Figure 2.6. The lengths a and b of the semimajor and semiminor axes are given by

$$a = p/(1 - e^2) \tag{2.18}$$

$$b = a(1 - e^2)^{1/2} \tag{2.19}$$

The point in the orbit where the satellite is closest to the earth is called the perigee and the point where the satellite is farthest from the earth is called the apogee. The perigee and apogee are **always** exactly opposite each other. To make θ_0 equal to zero, we have

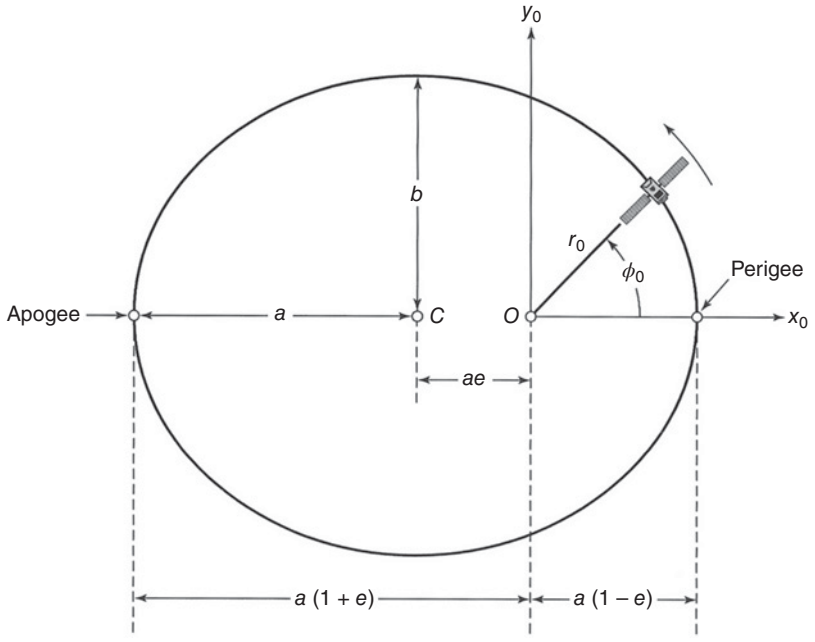


Figure 2.6 The orbit as it appears in the orbital plane. The point *O* is the center of the earth and the point *C* is the center of the ellipse. The two centers do not coincide unless the eccentricity, *e*, of the ellipse is zero (i.e., the ellipse becomes a circle and $a = b$). The dimensions of *a* and *b* are the semimajor and semiminor axes of the orbital ellipse, respectively.

chosen the x_0 axis so that both the apogee and the perigee lie along it and the x_0 axis is therefore the major axis of the ellipse.

The differential area swept out by the vector r_0 from the origin to the satellite in time dt is given by

$$dA = 0.5r_0^2\left(\frac{d\phi_0}{dt}\right)dt = 0.5hdt \tag{2.20}$$

Remembering that h is the magnitude of the orbital angular momentum of the satellite, the radius vector of the satellite can be seen to sweep out equal areas in equal times. This is Kepler's second law of planetary motion. By equating the area of the ellipse (πab) to the area swept out in one orbital revolution, we can derive an expression for the orbital period T as

$$T^2 = (4\pi^2 a^3)/\mu \tag{2.21}$$

This equation is the mathematical expression of Kepler's third law of planetary motion: the square of the period of revolution is proportional to the cube of the semimajor axis. (Note that this is the square of Eq. (2.6) and that in Eq. (2.6) the orbit was assumed to be circular such that semimajor axis a = semiminor axis b = circular orbit radius from the center of the earth r.) Kepler's third law extends the result from Eq. (2.6), which was derived for a circular orbit, to the more general case of an elliptical orbit. Equation (2.21) is extremely important in satellite communications systems. This equation determines the period of the orbit of any satellite, and it is used in every GPS

receiver in the calculation of the positions of GPS satellites. Equation (2.21) is also used to find the orbital radius of a GEO satellite, for which the period T must be made exactly equal to the period of one revolution of the earth for the satellite to remain stationary over a point on the equator.

An important point to remember is that the period of revolution, T, is referenced to inertial space, that is, to the galactic background. The orbital period is the time the orbiting body takes to return to the same reference point in space with respect to the galactic background. Nearly always, the primary body will also be rotating and so the period of revolution of the satellite may be different from that perceived by an observer who is standing still on the surface of the primary body. This is most obvious with a GEO satellite (see Table 2.1). The orbital period of a GEO satellite is exactly equal to the period of rotation of the earth, 23 hours 56 minutes 4.1 seconds, but, to an observer on the ground, the satellite appears to have an infinite orbital period: it always stays in the same place in the sky.

To be perfectly geostationary, the orbit of a satellite needs to have three features: (i) it must be exactly circular (i.e., have an eccentricity of zero); (ii) it must be at the correct altitude (i.e., have the correct orbital period); and (iii) it must be in the plane of the equator (i.e., have a zero inclination with respect to the equator). If the inclination of the satellite is not zero and/or if the eccentricity is not zero, but the orbital period is correct, then the satellite will be in a *geosynchronous* orbit. The position of a geosynchronous satellite will appear to oscillate about a mean look angle in the sky with respect to a stationary observer on the earth's surface. The orbital period of a GEO satellite, 23 hours 56 minutes 4.1 seconds, is one sidereal day. A sidereal day is the time between consecutive crossings of any particular longitude on the earth by any star, other than the sun (Gordon and Morgan 1993). The mean solar day of 24 hours is the time between any consecutive crossings of any particular longitude by the sun, and is the time between successive sunrises (or sunsets) observed at one location on earth, averaged over an entire year. Because the earth moves round the sun once per 365 ¼ days, the solar day is 1440/365.25 = 3.94 minutes longer than a sidereal day.

2.5 Locating the Satellite in the Orbit

We will consider now the problem of locating the satellite in its orbit. The equation of the orbit may be rewritten by combining Eqs. (2.15) and (2.18) to obtain

$$r_0 = \frac{a(1-e^2)}{1+e\cos\phi_0} \tag{2.22}$$

The angle ϕ_0 (see Figure 2.6) is measured from the x_0 axis and is called the true anomaly. (*Anomaly* was a measure used by astronomers to mean a planet's angular distance from its Perihelion, closest approach to the sun, measured as if viewed from the sun. The term was adopted in celestial mechanics for all orbiting bodies.) Since we defined the positive x_0 axis so that it passes through the perigee, ϕ_0 measures the angle from the perigee to the instantaneous position of the satellite. The rectangular coordinates of the satellite are given by

$$x_0 = r_0\cos\phi_0 \tag{2.23}$$

$$y_0 = r_0\sin\phi_0 \tag{2.24}$$

As noted earlier, the orbital period T is the time for the satellite to complete a revolution in inertial space, traveling a total of 2π radians. The average angular velocity η is thus

$$\eta = (2\pi)/T = (\mu^{1/2})/(a^{3/2}) \tag{2.25}$$

If the orbit is an ellipse, the instantaneous angular velocity will vary with the position of the satellite around the orbit. If we enclose the elliptical orbit with a circumscribed circle of radius a (see Figure 2.7), then an object going around the circumscribed circle with a constant angular velocity η would complete one revolution in exactly the same period T as the satellite requires to complete one (elliptical) orbital revolution.

Consider the geometry of the circumscribed circle as shown in Figure 2.7. Locate the point (indicated as A) where a vertical line drawn through the position of the satellite intersects the circumscribed circle. A line from the center of the ellipse (C) to this point (A) makes an angle E with the x_0 axis; E is called the *eccentric anomaly* of the satellite. It is related to the radius r_0 by

$$r_0 = a(1 - e\cos E) \tag{2.26}$$

Thus

$$a - r_0 = ae\cos E \tag{2.27}$$

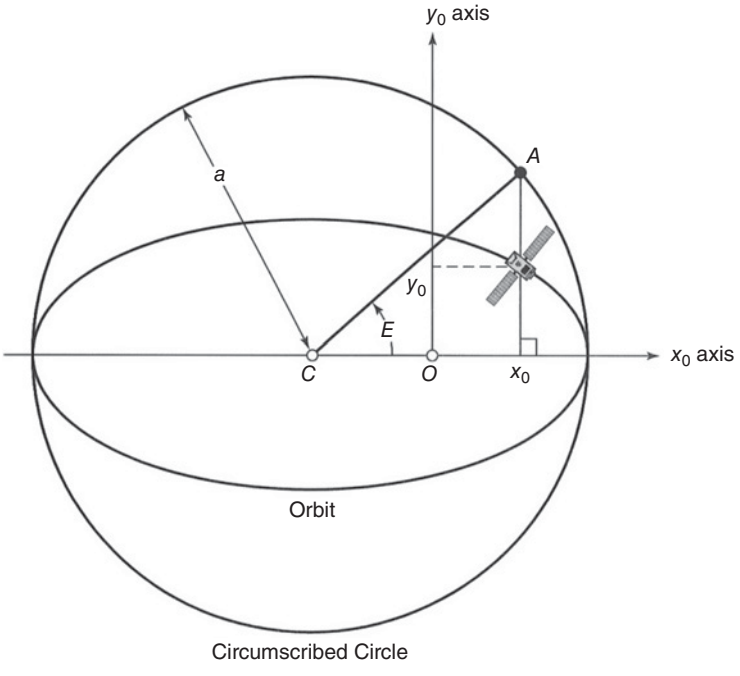


Figure 2.7 The circumscribed circle and the eccentric anomaly E. Point O is the center of the earth and point C is both the center of the orbital ellipse and the center of the circumscribed circle. The satellite location in the orbital plane coordinate system is specified by (x_0, y_0). A vertical line through the satellite intersects the circumscribed circle at point A. The eccentric anomaly E is the angle from the x_0 axis to the line joining C to A.

We can also develop an expression that relates eccentric anomaly E to the average angular velocity η, which yields

$$\eta dt = (1 - e\cos E)dE \tag{2.28}$$

Let t_p be the time of perigee. This is simultaneously the time of closest approach to the earth; the time when the satellite is crossing the x_0 axis; and the time when E is zero. If we integrate both sides of Eq. (2.28), we obtain

$$\eta(t - t_p) = E - e\sin E \tag{2.29}$$

The left side of Eq. (2.29) is called the *mean anomaly*, M. Thus

$$M = \eta(t - t_p) = E - e\sin E \tag{2.30}$$

The mean anomaly M is the arc length (in radians) that the satellite would have traversed since the perigee passage if it were moving on the circumscribed circle at the mean angular velocity η.

If we know the time of perigee, t_p, the eccentricity, e, and the length of the semimajor axis, a, we now have the necessary equations to determine the coordinates (r_0, ϕ_0) and (x_0, y_0) of the satellite in the orbital plane. The process is as follows:

1. Calculate η using Eq. (2.25)
2. Calculate M using Eq. (2.30)
3. Solve Eq. (2.30) for E
4. Find r_0 from E using Eq. (2.27)
5. Solve Eq. (2.22) for ϕ_0
6. Use Eqs. (2.23) and (2.24) to calculate x_0 and y_0

Now we must locate the orbital plane with respect to the earth.

2.6 Locating the Satellite With Respect to the Earth

At the end of the last section, we summarized the process for locating the satellite at the point (x_0, y_0, z_0) in the rectangular coordinate system of the orbital plane. The location was with respect to the center of the earth. In most cases, we need to know where the satellite is from an observation point that is not at the center of the earth. We will therefore develop the transformations that permit the satellite to be located from a point on the rotating surface of the earth. We will begin with a *geocentric equatorial coordinate system* as shown in Figure 2.8. The rotational axis of the earth is the z_i axis, which is through the geographic North Pole. The x_i axis is from the center of the earth toward a fixed location in space called the *first point of Aries* (see Figure 2.8). This coordinate system moves through space; it translates as the earth moves in its orbit around the sun, but it does not rotate as the earth rotates. The x_i direction is always the same, whatever the earth's position around the sun and it is in the direction of the first point of Aries. The (x_i, y_i) plane contains the earth's equator and is called the *equatorial plane*.

Angular distance measured eastward in the equatorial plane from the x_i axis is called *right ascension* and given the symbol *RA*. The two points at which the orbit penetrates the equatorial plane are called nodes; the satellite moves upward through the equatorial plane at the *ascending node* and downward through the equatorial plane

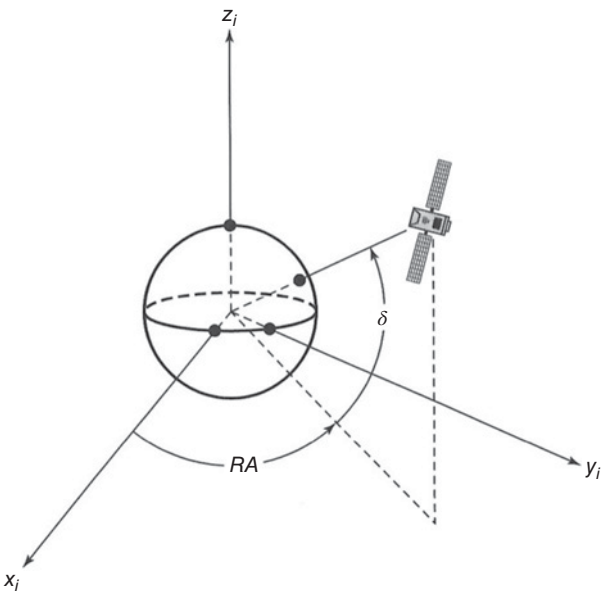


Figure 2.8 The geocentric equatorial system. This geocentric system differs from that shown in Figure 2.1 only in that the x_i axis points to the first point of Aries. The first point of Aries is the direction of a line from the center of the earth through the center of the sun at the vernal equinox (20 or 21 March in the Northern Hemisphere), the instant when the subsolar point crosses the equator from south to north. In the above system, an object may be located by its right ascension *RA* and declination δ.

at the *descending node*, given the conventional picture of the earth, with north at the top, which is in the direction of the positive z axis for the earth centered coordinate set. Remember that in space there is no *up* or *down*; that is a concept we are familiar with because of gravity at the earth's surface. For a weightless body in space, such as an orbiting spacecraft, up and down have no meaning unless they are defined with respect to a reference point. The *right ascension of the ascending node* is called Ω. The angle that the orbital plane makes with the equatorial plane (the planes intersect at the line joining the nodes) is called the *inclination*, *i*. Figure 2.9 illustrates these quantities.

The variables Ω and i together locate the orbital plane with respect to the equatorial plane. To locate the orbital coordinate system with respect to the equatorial coordinate system we need ω, the *argument of perigee west*. This is the angle measured along the orbit from the ascending node to the perigee.

Standard time for space operations and most other scientific and engineering purposes is universal time (UT), also known as *zulu time* (*z*). This is essentially the mean solar time at the Greenwich Observatory near London, England. UT is measured in hours, minutes, and seconds or in fractions of a day. It is 5 hours later than Eastern Standard Time, so that 07 : 00 EST is 12 : 00 : 00 hours UT. The civil or calendar day begins at 00 : 00 : 00 hours UT, frequently written as 0 hours. This is, of course, midnight (24 : 00 : 00) on the previous day. Astronomers employ a second dating system involving *Julian days* and *Julian dates*. Julian days start at noon UT in a counting system whereby noon on 31 December 1899, was the beginning of Julian day 2415020, usually written as 241 5020. These are extensively tabulated in (The American Ephemeris and Nautical Almanac n.d., published annually) and additional information is in (Wertz and Larson 1999). As an example, noon on 31 December 2000, the eve of the twenty-first century, is the start of Julian day 245 1909. Julian dates can be used to indicate time by appending a decimal fraction; 00 : 00 : 00 hours UT on 1 January 2001 – zero hour, minute, and second for the third millennium CE – is given by Julian date 245 1909.5. To find the

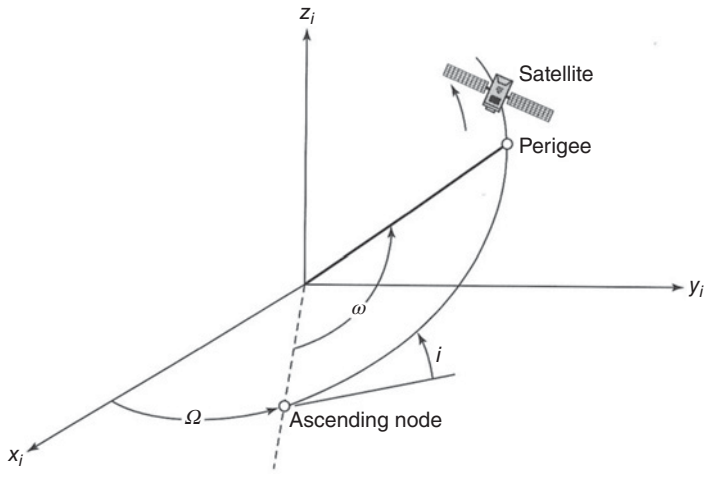


Figure 2.9 Locating the orbit in the geocentric equatorial system. The satellite penetrates the equatorial plane (while moving in the positive *z* direction) at the ascending node. The right ascension of the ascending node is Ω and the inclination *i* is the angle between the equatorial plane and the orbital plane. Angle ω, measured in the orbital plane, locates the perigee with respect to the equatorial plane.

exact position of an orbiting satellite at a given instant in time requires knowledge of the orbital elements.

2.7 Orbital Elements

To specify the absolute (i.e., the inertial) coordinates of a satellite at time t, we need to know six quantities. (This was evident earlier when we determined that a satellite's equation of motion was a second order vector linear differential equation.) These quantities are called the orbital elements. More than six quantities can be used to describe a unique orbital path and there is some arbitrariness in exactly which six quantities are used. We have chosen to adopt a set that is commonly used in satellite communications: eccentricity (e), semimajor axis (a), time of perigee (t_p), right ascension of ascending node (Ω), inclination (i), and argument of perigee (ω). Frequently, the mean anomaly (M) at a given time is substituted for t_p.

Example 2.3 Geostationary Satellite Orbit (GEO) Radius

Question: The earth rotates once per sidereal day (23 hours 56 minutes 4.09 seconds). Use Eq. (2.21) to show that the radius of the GEO is 42 164.17 km as given in Table 2.1.

Answer
Equation (2.21) enables us to find the period of a satellite's orbit given the radius of the orbit. Namely

$$T^2 = (4\pi^2 a^3)/\mu \text{ seconds}$$

Rearranging the equation, the orbital radius a is given by

$$a^3 = T^2\mu/(4\pi^2)$$

For one sidereal day, $T = 86\,164.09$ seconds. Hence

$$a^3 = (86164.1)^2 \times 3.986004418 \times 10^5/(4\pi^2) = 7.496020251 \times 10^{13}\ \text{km}^3$$

Thus $a = 42\,164.17$ km
This is the orbital radius for a geostationary satellite, as given in Table 2.1.

Example 2.4 Low Earth Orbit

Question: A SpaceX mission to the ISS is an example of a LEO satellite mission. Before rendezvousing with the ISS on this mission, SpaceX inserted the *Dragon* capsule into an initial circular orbit 250 km above the earth's surface, where there are still a finite number of molecules from the atmosphere. The mean earth's radius, r_e, is approximately 6378.14 km. Using these numbers, calculate the period of the Dragon capsule of SpaceX in its 250 km orbit. Find also the linear velocity of the Dragon capsule along this orbit.

Answer
The radius from the center of the earth of the 250 km altitude Dragon orbit is $(r_e + h)$, where h is the orbital altitude, and this = 6378.14 + 250.0 = 6628.14 km.
From Eq. 2.21, the period of the orbit is T where

$$T^2 = (4\pi^2 a^3)/\mu = 4\pi^2 \times (6628.14)^3/3.986004418 \times 10^5\ \text{s}^2$$
$$= 2.88401145 \times 10^7\ \text{s}^2$$

Hence the period of the orbit is

$$T = 5370.30\ \text{seconds} = 89\ \text{minutes}\ 30.3\ \text{seconds}$$

This orbit period is about as small as possible. At a lower altitude, friction with the earth's atmosphere will quickly slow the Dragon capsule down and it will return to earth. Thus, all spacecraft in stable earth orbit tend to have orbital periods exceeding 89 minutes 30 seconds.
The circumference of the orbit is $2\pi a = 41\,645.83$ km.
Hence the velocity of the Dragon in orbit is

$$2\pi a/T = 41645.83/5370.30 = 7.755\ \text{km/s}$$

Alternatively, you could use Eq. (2.5): $v = (\mu/r)^{1/2}$.
The term $\mu = 3.986\,004\,418 \times 10^5\ \text{km}^3/\text{s}^2$ and the term $r = (6378.14 + 250.0)$ km, yielding $v = 7.755$ km/s.
Note: If μ and r had been quoted in units of m^3/s^2 and m, respectively, the answer would have been in meters/second. Be sure to keep the units the same during a calculation procedure. A velocity of about 7.8 km/s is a typical velocity for a LEO satellite. As the altitude of a satellite increases, its velocity becomes smaller.

Example 2.5 Elliptical Orbit

Question: A satellite is in an elliptical orbit with a perigee of 1000 km and an apogee of 4000 km. Using a mean earth radius of 6378.14 km, find the period of the orbit in hours, minutes, and seconds, and the eccentricity of the orbit.

Answer
The major axis of the elliptical orbit is a straight line between the apogee and perigee, as seen in Figure 2.6. Hence, for a semimajor axis length a, earth radius r_e, perigee height h_p, and apogee height h_a,

$$2a = 2r_e + h_p + h_a = 2 \times 6378.14 + 1000.0 + 4000.0 = 17\,756.28 \text{ km}$$

Thus the semimajor axis of the orbit has a length $a = 8878.14$ km. Using this value of a in Eq. (2.21) gives an orbital period T seconds where

$$\begin{aligned} T^2 &= (4\pi^2 a^3)/\mu = 4\pi^2 \times (8878.07)^3/3.986004418 \times 10^5 \text{ s}^2 \\ &= 6.930872802 \times 10^7 \text{ s}^2 \end{aligned}$$

which gives

$$\begin{aligned} T &= 8\,325.186\,4 \text{ seconds} = 138 \text{ minutes } 45.19 \text{ seconds} \\ &= 2 \text{ hours } 18 \text{ minutes } 45.19 \text{ seconds} \end{aligned}$$

The eccentricity of the orbit is given by e, which can be found from Eq. 2.27 by considering the instant at which the satellite is at perigee. Referring to Figure 2.7, when the satellite is at perigee, the eccentric anomaly $E = 0$ and $r_0 = r_e + h_p$. From Eq. (2.27), at perigee

$$r_0 = a(1 - e\cos E) \text{ and } \cos E = 1$$

Hence

$$r_e + h_p = a(1 - e)$$

which gives

$$e = 1 - (r_e + h_p)/a = 1 - 7378.14/8878.14 = 0.169.$$

2.8 Look Angle Determination

Navigation around the earth's oceans became more precise when the surface of the globe was divided up into a grid-like structure of orthogonal lines: latitude and longitude. Latitude is the angular distance, measured in degrees, north or south of the equator and longitude is the angular distance, measured in degrees, from a given reference longitudinal line. At the time that this grid reference became popular, there were two major sea-faring nations vying for dominance: England and France. England drew its reference zero longitude through Greenwich, a town close to London, England, and France, not surprisingly, drew its reference longitude through Paris, France. Since the British Admiralty chose to give away their maps and the French decided to charge a fee for theirs, it was not surprising that the use of Greenwich as the zero reference longitude became dominant within a few years. (It was the start of .com market dominance through giveaways three centuries before E-commerce!) Geometry was a lot older science than navigation and so 90° per quadrant on the map was an obvious selection to make. Thus, there are 360° of longitude (measured from 0° at the Greenwich Meridian, the line drawn from the North Pole to the South Pole through Greenwich, England) and ±90° of latitude, plus being measured north of the equator and minus south of the equator. Latitude 90°N (or +90°) is the North Pole and Latitude 90°S (or −90°) is the South Pole. When GEO satellite

systems are registered in Geneva, their (subsatellite) location over the equator is given in degrees east to avoid confusion. Thus, the INTELSAT primary location in the Indian Ocean is registered at 60°E and the primary location in the Atlantic Ocean at 335.5°E (not 24.5°W). Earth stations that communicate with satellites are described in terms of their geographic latitude and longitude when developing the pointing coordinates that earth station must use to track the apparent motion of the satellite.

Frequencies and orbital slots for new satellites used to be registered with the International Frequency Registration Board (IFRB), part of the International Telecommunications Union (ITU) located in Geneva. The re-organization of the ITU led to the responsibilities of the IFRB being taken over by the Radio Regulations Board (RRB) and the Radiocommunication Bureau (RB) of the ITU. The initial application by an organization or company that wants to orbit a new satellite is made to the national body that controls the allocation and use of radio frequencies – the Federal Communications Commission (FCC) in the United States, for example – which must first approve the application and then forward it to the appropriate branch of the ITU (either the RRB or the RB). The first organization to file with the ITU for a particular service is deemed to have protection from newcomers. Any other organization filing to carry the same service at, or close to that orbital location (within 2°) must coordinate their use of the frequency bands with the first organization to file. The first user may cause interference into subsequent filer's satellite systems, since they were the first to be awarded the orbital slot and frequencies, but the later filers' satellites must not cause interference with the first user's system.

The coordinates to which an earth station antenna must be pointed to communicate with a satellite are called the *look angles*. These are most commonly expressed as *azimuth* (Az) and *elevation* (El), although other pairs exist. For example, right ascension and declination are standard for radio astronomy antennas. Azimuth is measured eastward (clockwise) from geographic north to the projection of the satellite path on a (locally) horizontal plane at the earth station. Elevation is the angle measured upward from the local horizontal plane at the earth station to the satellite path. Figure 2.10 illustrates these look angles. In all look angle determinations, the precise location of the satellite is critical. A key location in many instances is the *subsatellite point*.

2.8.1 The Subsatellite Point

The subsatellite point is the location on the surface of the earth that lies directly between the satellite and the center of the earth. It is the *nadir* pointing direction from the satellite and, for a satellite in an equatorial orbit, it will always be located on the equator. Since geostationary satellites are in equatorial orbits and are designed to stay stationary over the earth, it is usual to give their orbital location in terms of their subsatellite point. As noted in the example given earlier, the Intelsat primary satellite in the Atlantic Ocean Region (AOR) is at 335.5°E longitude. Operators of international geostationary satellite systems that have satellites in all three ocean regions (Atlantic, Indian, and Pacific) tend to use longitude east to describe the subsatellite points to avoid confusion between using both east and west longitude descriptors. For US geostationary satellite operators, all of the satellites are located west of the Greenwich meridian and so it has become accepted

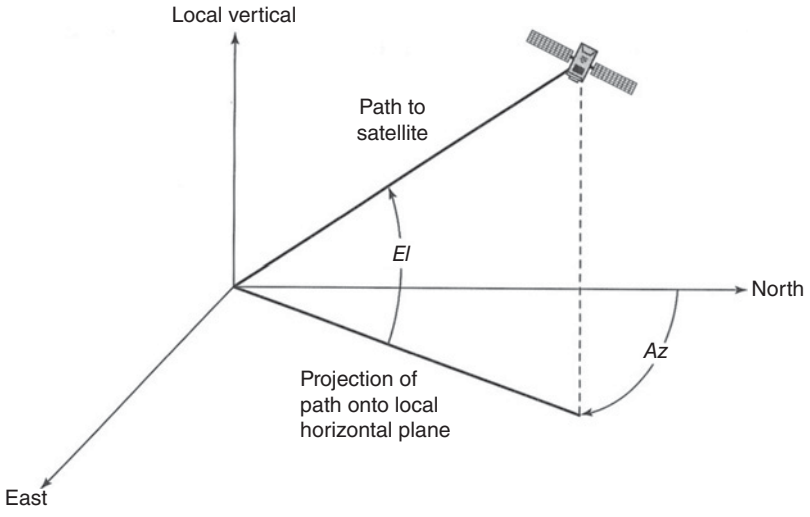


Figure 2.10 The definition of elevation (*El*) and azimuth (*Az*). The elevation angle is measured upward from the local horizontal at the earth station and the azimuth angle is measured from true north in an eastward direction to the projection of the satellite path onto the local horizontal plane.

practice for regional systems over the United States to describe their geostationary satellite locations in terms of degrees W.

To an observer of a satellite standing at the subsatellite point, the satellite will appear to be directly overhead, in the *zenith* direction from the observing location. The zenith and nadir paths are therefore in opposite directions along the same path (see Figure 2.11).

Designers of satellite antennas reference the pointing direction of the satellite's antenna beams to the nadir direction. The communications coverage region on the earth from a satellite is defined by angles measured from nadir at the satellite to the edges of the coverage. Earth station antenna designers, however, do not reference their pointing direction to zenith. As noted earlier, they use the local horizontal plane at the earth station to define elevation angle and geographical compass points to define azimuth angle, thus giving the two look angles for the earth station antenna toward the satellite (Az, El).

2.8.2 Elevation Angle Calculation

Figure 2.12 shows the geometry of the elevation angle calculation.

In Figure 2.12, r_s is the vector from the center of the earth to the satellite; r_e is the vector from the center of the earth to the earth station; and d is the vector from the earth station to the satellite. These three vectors lie in the same plane and form a triangle. The central angle γ measured between r_e and r_s is the angle between the earth station and the satellite, and ψ is the angle (within the triangle) measured from r_e to d. Defined so that it is non-negative, γ is related to the earth station north latitude L_e (i.e., L_e is the number of degrees in latitude that the earth station is north from the equator) and west longitude l_e (i.e., l_e is the number of degrees in longitude that the earth station is west

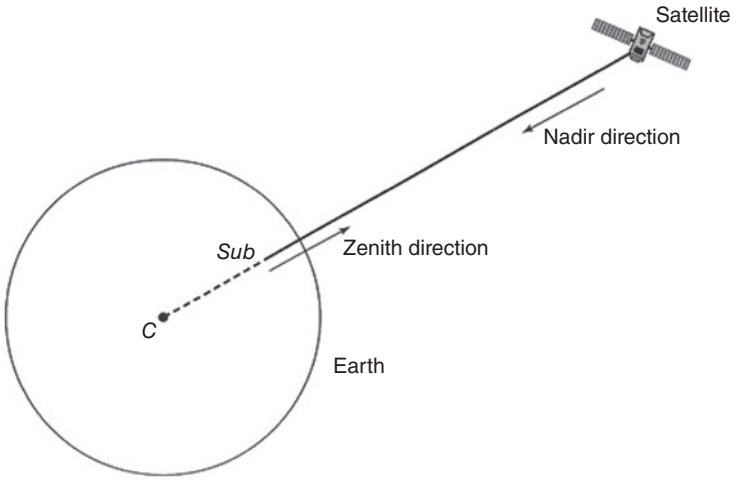


Figure 2.11 Zenith and Nadir pointing directions. The line joining the satellite and the center of the earth, *C*, passes through the surface of the earth at point *Sub*, the subsatellite point. The satellite is directly overhead at this point and so an observer at the subsatellite point would see the satellite at zenith (i.e., at an elevation angle of 90°.) The pointing direction from the satellite to the subsatellite point is the nadir direction from the satellite. If the beam from the satellite antenna is to be pointed at a location on the earth that is not at the subsatellite point, the pointing direction is defined by the angle away from nadir. In general, two off-nadir angles are given: the number of degrees north (or south) from nadir; and the number of degrees east (or west) from nadir. East, west, north, and south directions are those defined by the geography of the earth.

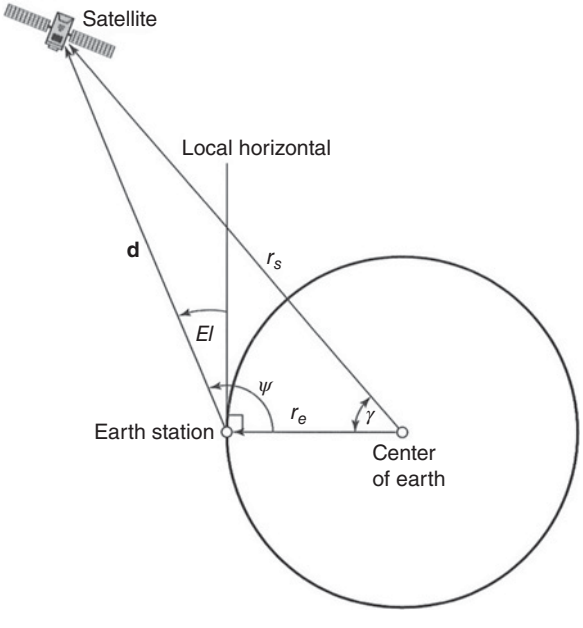


Figure 2.12 The geometry of elevation angle calculation. The plane of the paper is the plane defined by the center of the earth, the satellite, and the earth station. The central angle is γ. The elevation angle *El* is measured upward from the local horizontal at the earth station.

from the Greenwich meridian) and the subsatellite point at north latitude L_s and west longitude l_s by

$$\cos(\gamma) = \cos(L_e)\cos(L_s)\cos(l_s - l_e) + \sin(L_e)\sin(L_s) \tag{2.31}$$

The law of cosines allows us to relate the magnitudes of the vectors joining the center of the earth, the satellite, and the earth station. Thus

$$d = r_s\left[1 + \left(\frac{r_e}{r_s}\right)^2 - 2\left(\frac{r_e}{r_s}\right)\cos(\gamma)\right]^{1/2} \tag{2.32}$$

Since the local horizontal plane at the earth station is perpendicular to r_e, the elevation angle El is related to the central angle ψ by

$$El = \psi - 90^o \tag{2.33}$$

By the law of sines we have

$$\frac{r_s}{\sin(\psi)} = \frac{d}{\sin(\gamma)} \tag{2.34}$$

Combining the last three equations yields

$$\begin{aligned}\cos(El) &= \frac{r_s\sin(\gamma)}{d} \\ &= \frac{\sin(\gamma)}{\left[1 + \left(\frac{r_e}{r_s}\right)^2 - 2\left(\frac{r_e}{r_s}\right)\cos(\gamma)\right]^{1/2}}\end{aligned} \tag{2.35}$$

Equations (2.35) and (2.31) permit the elevation angle El to be calculated from knowledge of the subsatellite point and the earth station coordinates, the orbital radius r_s, and the earth's radius r_e. An accurate value for the average earth radius is 6378.137 km, but a common value used in approximate determinations is 6370 km (Gordon and Morgan 1993).

2.8.3 Azimuth Angle Calculation

Since the earth station, the center of the earth, the satellite, and the subsatellite point all lie in the same plane, the azimuth angle Az from the earth station to the satellite is the same as the azimuth from the earth station to the subsatellite point. This is more difficult to compute than the elevation angle because the exact geometry involved depends on whether the subsatellite point is east or west of the earth station, and in which of the hemispheres the earth station and the subsatellite point are located. The problem simplifies somewhat for geosynchronous satellites, which will be treated in the next section. For the general case, in particular for constellations of LEO satellites, the tedium of calculating the individual look angles on a second-by-second basis has been considerably eased by a range of commercial software packages that exist for predicting a variety of orbital dynamics and intercept solutions for some open source software developed by the Goddard Space Flight Center (GSFC) for assisting in the orbit determination of satellites, particularly for formation flying (see NASA 2018a,b).

A popular suite of software employed by many launch service contractors is that developed by Analytical Graphics: the *Satellite Tool Kit* (STK.com 2018). The core program was used in early 2001, STK 4.0, and the subsequent subseries, was used by Hughes to rescue AsiaSat 3 when that satellite intended for GEO was stranded in a highly elliptical orbit following the failure of an upper stage in the launch vehicle. Hughes used two lunar flybys to provide the necessary additional velocity to circularize the orbit at geostationary altitude. The acronym STK now stands for *Systems Tool Kit*, and it permits the prediction of craft in the air and on land, and not just in space. A number of organizations offer web sites that provide orbital plots with rapid updates for a variety of satellites (e.g., the NASA site (NASA 2018b). A particularly useful one is n2yo (n2yo.com 2018).

2.8.4 Specialization to Geostationary Satellites

For most geostationary satellites, the subsatellite point is on the equator at longitude l_s, and the latitude L_s is 0. The geosynchronous radius r_s is 42 164.17 km (Gordon and Morgan 1993). Since L_s is zero, Eq. (2.31) simplifies to

$$\cos(\gamma) = \cos(L_e)\cos(l_s - l_e) \tag{2.36}$$

Substituting $r_s = 42\,164.17$ km and $r_e = 6378.137$ km in Eqs. (2.31) and (2.35) gives the following expressions for the distance d from the earth station to the satellite and the elevation angle El at the earth station:

$$d = 42164.17[1.02288235 - 0.30253825\cos(\gamma)]^{1/2}\ \text{km} \tag{2.37}$$

$$\cos(El) = \frac{\sin(\gamma)}{[1.02288235 - 0.30253825\cos(\gamma)]^{1/2}} \tag{2.38}$$

For a geostationary satellite with an orbital radius of 42 164.17 km and a mean earth radius of 6378.137 km, the ratio $r_s/r_e = 6.6107345$ giving

$$El = \tan^{-1}[(6.6107345 - \cos\gamma)/\sin\gamma] - \gamma \tag{2.39}$$

To find the azimuth angle, an intermediate angle α must first be found. The intermediate angle α permits the correct 90° quadrant to be found for the azimuth since the azimuthal angle can lie anywhere between 0° (true north) and clockwise through 360° (back to true north again).

The intermediate angle is found from

$$\alpha = \tan^{-1}\left[\frac{\tan|(l_s - l_e)|}{\sin(L_e)}\right] \tag{2.40}$$

Having found the intermediate angle α, the azimuth look angle Az can be found from:

Case 1: Earth station in the Northern Hemisphere with

(a) Satellite to the SE of the earth station:

$$Az = 180° - \alpha \tag{2.41a}$$

(b) Satellite to the SW of the earth station:

$$Az = 180° + \alpha \tag{2.41b}$$

Case 2: Earth station in the Southern Hemisphere with

(c) Satellite to the NE of the earth station:

$$Az = \alpha \tag{2.41c}$$

(d) Satellite to the NW of the earth station:

$$Az = 360° - \alpha \tag{2.41d}$$

2.8.5 Visibility Test

For a satellite to be visible from an earth station, its elevation angle *El* must be above some minimum value, which is at least 0°. A positive or zero elevation angle requires that (see Figure 2.13)

$$r_s \geq \frac{r_e}{\cos(\gamma)} \tag{2.42}$$

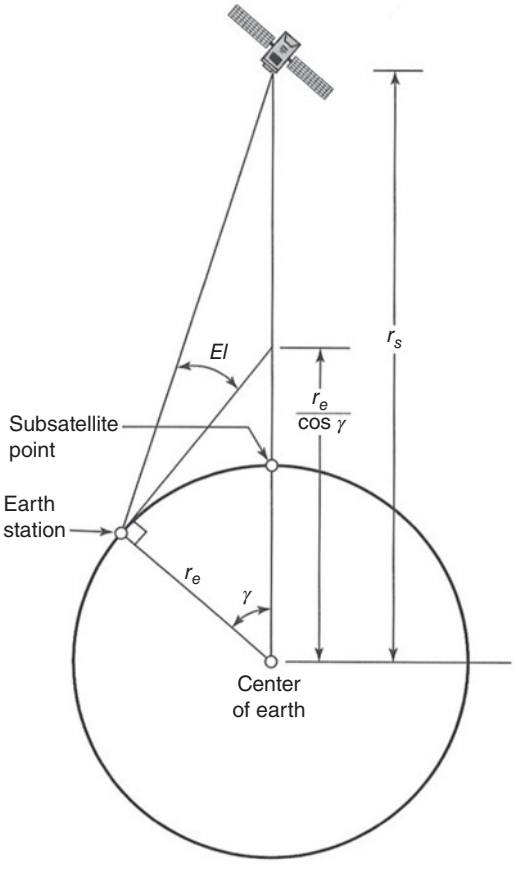


Figure 2.13 The geometry of the visibility calculation. The satellite is said to be visible from the earth station if the elevation angle *El* is positive. This requires that the orbital radius r_s be greater than the ratio $r_e/\cos(\gamma)$ where r_e is the radius of the earth and γ is the central angle.

This means that the maximum central angular separation between the earth station and the sub satellite point is limited by

$$\gamma \leq \cos^{-1}\left(\frac{r_e}{r_s}\right) \tag{2.43}$$

For a nominal geostationary orbit, the last equation reduces to $\gamma \leq 81.3°$ for the satellite to be visible.

Example 2.6 Geostationary Satellite Look Angles

An earth station situated in the Docklands of London, England, needs to calculate the look angle to a geostationary satellite in the Indian Ocean operated by Intelsat. The details of the earth station site and the satellite are as follows:

Earth station latitude and longitude are 52.0°N and 0°, respectively.

Satellite longitude (i.e., the subsatellite point) is 66.0° E.

Step 1:

Find the central angle γ

$$\begin{aligned}\cos(\gamma) &= \cos(L_e)\cos(l_s - l_e)\\ &= \cos(52.0)\cos(6) = 0.2504\end{aligned}$$

yielding $\gamma = 75.4981°$

The central angle γ is less than 81.3° so the satellite is visible from the earth station.

Step 2:

Find the elevation angle *El*

$$\begin{aligned}El &= \tan^{-1}[(6.6107345 - \cos\gamma)/\sin\gamma] - \gamma\\ &= \tan^{-1}[(6.6107345 - 0.2504)/\sin(75.4981)] - 75.4981\\ &= 5.847°\end{aligned}$$

Step 3:

Find the intermediate angle α

$$\begin{aligned}\alpha &= \tan^{-1}\left[\frac{\tan|(l_s - l_e)|}{\sin(L_e)}\right]\\ &= \tan^{-1}[(\tan(66.0 - 0))/\sin(52.0)]\\ &= 70.667°\end{aligned}$$

Step 4:

Find the azimuth angle

The earth station is in the Northern Hemisphere and the satellite is to the south east of the earth station. From Eq. 2.41a, this gives

$$Az = 180^o - \alpha = 180 - 70.667 = 109.333^o \text{ (clockwise from true north)}$$

Note that, in the example above, the elevation angle is relatively low (5.85°). Refractive effects in the atmosphere will cause the mean ray path to the satellite to bend in the elevation plane (making the satellite appear to be higher in the sky than it actually is) and to cause the amplitude of the signal to fluctuate with time. These aspects are discussed

more fully in the propagation effects chapter. While it is unusual to operate to a satellite below established elevation angle minima (typically 5° at C-band, 10° at Ku-band, and in most cases, 20° at Ka-band and above), many times it is not possible to do this. Such cases exist for high latitude regions and for satellites attempting to reach extreme east and west coverages from their given geostationary equatorial location. To establish whether a particular satellite location can provide service into a given region, a simple visibility test can be carried out, as shown earlier in Eqs. (2.42) and (2.43).

A number of geosynchronous orbit satellites have inclinations that are much larger than the nominal 0.05° inclination maximum for current geosynchronous satellites. (In general, a geosynchronous satellite with an inclination of <0.1° may be considered to be geostationary.) In extreme cases, the inclination can be several degrees, particularly if the orbit maneuvering fuel of the satellite is almost exhausted and the satellite's position in the nominal location is only controlled in longitude and not in inclination. This happens with most geostationary communications satellites toward the end of their operational lifetime since the reliability of the payload, or a large part of the payload, generally exceeds that of the lifetime of the maneuvering fuel. Those satellites that can no longer be maintained in a fully geostationary orbit, but are still used for communications services, are referred to as *inclined orbit* satellites. While they now need to have tracking antennas at the earth terminals once the inclination becomes too large to allow the satellite to remain within the 1 dB beamwidth of the earth station antennas, substantial additional revenue can be earned beyond the normal lifetime of the satellite. Those satellites that eventually reach significantly inclined orbits can also be used to communicate to parts of the high latitude regions that were once beyond reach, but only for a limited part of the day. The exceptional reliability of electronic components in space, once they have survived the launch and deployment sequences, has led spacecraft designers to manufacture satellites with two end-of-life criteria. These are: End Of Design Life (EODL), which refers to the lifetime expectancy of the payload components and End Of Maneuvering Life (EOML), which refers to the spacecraft bus capabilities, in particular the anticipated lifetime of the spacecraft with full maneuver capabilities in longitude and inclination.

Current spacecraft are designed with fuel tanks that have a capacity that usually significantly exceeds the requirement for EODL. Once the final mass of the spacecraft (without fuel) is known, a decision can be made as to how much additional fuel to load so that the economics of the launch and the anticipated additional return on investment can be balanced. Having additional fuel on board the spacecraft can be advantageous for many reasons, in addition to adding on-orbit lifetime. In many cases, satellites are moved to new locations during their operational lifetime. Examples for this are opening up service at a new location with an older satellite or replacing a satellite that has had catastrophic failure with a satellite from a location that has fewer customers. Each maneuver, however, consumes fuel. A rule of thumb is that any change in orbital location for a geostationary satellite reduces the maneuvering lifetime by about one month. Moving the satellite's location by 1° in longitude takes as much additional fuel as moving the location by 180°: both changes require an acceleration burn, a drift phase, and a deceleration burn. The 180° location change will clearly take longer, since the drift rates are the same in both cases. Another use for additional fuel is to allow for orbital perturbations at any location.

2.9 Orbital Perturbations

The orbital equations developed in Section 2.1 modeled the earth and the satellite as point masses influenced only by gravitational attraction. Under these ideal conditions, a Keplerian orbit results, which is an ellipse whose properties are constant with time. In practice, the satellite and the earth respond to many other influences including asymmetry of the earth's gravitational field, the gravitational fields of the sun and the moon, and solar radiation pressure. For LEO satellites, atmospheric drag can also be important. All of these interfering forces cause the true orbit to be different from a simple Keplerian ellipse; if unchecked, they would cause the subsatellite point of a nominally geosynchronous satellite to move with time.

Historically, much attention has been given to techniques for incorporating additional perturbing forces into orbit descriptions. The approach normally adopted for communications satellites is first to derive an *osculating orbit* for some instant in time (the Keplerian orbit the spacecraft would follow if all perturbing forces were removed at that time) with orbital elements (a, e, tp, Ω, i, ω). The perturbations are assumed to cause the orbital elements to vary with time and the orbit and satellite location at any instant are taken from the osculating orbit calculated with orbital elements corresponding to that time. To visualize the process, assume that the osculating orbital elements at time t_0 are (a_0, e_0, t_p, Ω_0, i_0, ω_0). Then assume that the orbital elements vary linearly with time at constant rates given by (da/dt, de/dt, etc.). The satellite's position at any time t_1 is then calculated from a Keplerian orbit with elements

$$a_0 + \frac{da}{dt}(t_1 - t_0), e_0 + \frac{d_e}{dt}\left(t_1 - t_0\right), etc.$$

This approach is particularly useful in practice because it permits the use of either theoretically calculated derivatives or empirical values based on satellite observations.

As the perturbed orbit is not an ellipse, some care must be taken in defining the orbital period. Since the satellite does not return to the same point in space once per revolution, the quantity most frequently specified is the so-called *anomalistic period*: the elapsed time between successive perigee passages. In addition to the orbit not being a perfect Keplerian ellipse, there will be other influences that will cause the apparent position of a geostationary satellite to change with time. These can be viewed as those causing mainly longitudinal changes and those that principally affect the orbital inclination.

2.9.1 Longitudinal Changes

2.9.1.1 Effects of the Earth's Oblateness

The earth is neither a perfect sphere nor a perfect ellipse; it can be better described as a triaxial ellipsoid (Gordon and Morgan 1993). The earth is flattened at the poles; the equatorial diameter is about 20 km more than the average polar diameter. The equatorial radius is not constant, although the noncircularity is small: the radius does not vary by more than about 100 m around the equator (Gordon and Morgan 1993). In addition to these non-regular features of the earth, there are regions where the average density of the earth appears to be higher. These are referred to as regions of mass concentration or *Mascons*. The nonsphericity of the earth, the noncircularity of the equatorial radius, and the Mascons lead to a non-uniform gravitational field around the earth. The force on an orbiting satellite will therefore vary with position.

For a LEO satellite, the rapid change in position of the satellite with respect to the earth's surface will lead to an averaging out of the perturbing forces in line with the orbital velocity vector. The same is not true for a geostationary (or geosynchronous) satellite. A geostationary satellite is weightless when in orbit. The smallest force on the satellite will cause it to accelerate and then drift away from its nominal location. The satellite is required to maintain a constant longitudinal position over the equator, but there will generally be an additional force toward the nearest equatorial bulge in either an eastward or a westward direction along the orbit plane. Since this will rarely be in line with the main gravitational force toward the earth's center, there will be a resultant component of force acting in the same direction as the satellite's velocity vector or against it, depending on the precise position of the satellite in the GEO orbit. This will lead to a resultant acceleration or deceleration component that varies with longitudinal location of the satellite.

Due to the position of the Mascons and equatorial bulges, there are four equilibrium points in the geostationary orbit: two of them stable and two unstable. The stable points are analogous to the bottom of a valley, and the unstable points to the top of a hill. If a ball is perched on top of a hill, a small push will cause it to roll down the slope into a valley, where it will roll backward and forward until it gradually comes to a final stop at the lowest point. The satellite at an unstable orbital location is at the top of a gravity hill. Given a small force, it will drift down the gravity slope into the gravity well (valley) and finally stay there, at the stable position. The stable points are at about 75°E and 252°E and the unstable points are at around 162°E and 348°E (Gordon and Morgan 1993). If a satellite is perturbed slightly from one of the stable points, it will tend to drift back to the stable point without any thruster firings required. A satellite that is perturbed slightly from one of the unstable points will immediately begin to accelerate its drift toward the nearer stable point and, once it reaches this point, it will oscillate in longitudinal position about this point until (centuries later) it stabilizes at that point. These stable points are sometimes called the *graveyard* geosynchronous orbit locations (not to be confused with the graveyard orbit for a geosynchronous satellite – the orbit to which the satellite is raised once the satellite ceases to be useful). Note that, due to the nonsphericity of the earth, etc., the stable points are neither exactly 180° apart nor are the stable and unstable points precisely 90° apart.

2.9.2 Inclination Changes

2.9.2.1 Effects of the Sun and the Moon

The plane of the earth's orbit around the sun – the ecliptic – is at an inclination of 7.3° to the equatorial plane of the sun. The earth is titled about 23° away from the normal to the ecliptic, as illustrated in Figure 2.14.

The moon circles the earth with an inclination of around 5° to the equatorial plane of the earth.

Due to the fact that the various planes – the sun's equator, the ecliptic, the earth's geographic equator (a plane normal to the earth's rotational axis), and the moon's orbital plane around the earth – are all different, a satellite in orbit around the earth will be subjected to a variety of out-of-plane forces. That is, there will generally be a net acceleration force that is not in the plane of the satellite's orbit, and this will tend to try to change the inclination of the satellite's orbit from its initial inclination. Under these conditions, the orbit will precess and its inclination will change.

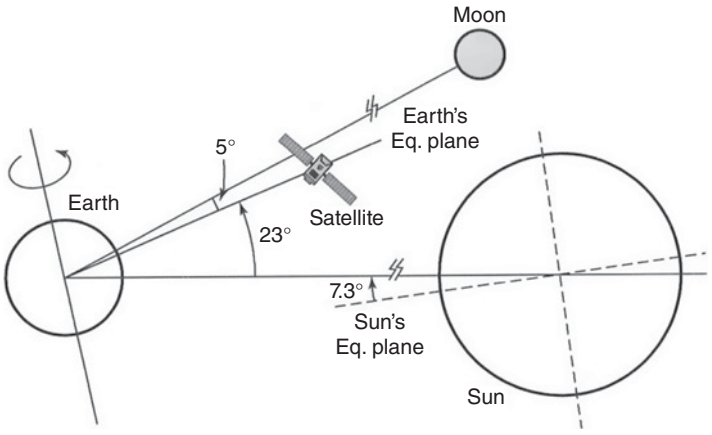


Figure 2.14 Relationship between the orbital planes of the sun, moon, and earth. The plane of the earth's orbit around the sun is the *ecliptic*. The geostationary orbit plane (earth's equatorial plane) is about 23° out of the ecliptic, and leads to maximum out-of-geostationary-orbit plane forces at the solstice periods (approximately 21 June and 21 December). The orbit of the moon is inclined about 5° to the earth's equatorial plane. The moon revolves around the earth in 27.3 days, the earth (and the geostationary satellites) rotates once about every 24 hours, and the earth revolves around the sun every 365.25 days. In addition, the sun – which has a greater girth at the equator than at the poles – has its equator inclined about 7.3° to the ecliptic. All of these various angular differences and orbital periods lead to conditions where all of the out of plane gravitational forces are in one direction with respect to the equatorial (geostationary orbital) plane at a given time as well as to conditions where the various gravitational out-of-plane forces partially cancel each other out. The precessional forces that cause the inclination of the geostationary satellite's orbit to move away from the equatorial plane therefore vary with time.

The mass of the sun is significantly larger than that of the moon but the moon is considerably closer to the earth than the sun (see Table 2.2). For this reason, the acceleration force induced by the moon on a geostationary satellite is about twice as large as that of the sun. The net effect of the acceleration forces induced by the moon and the sun on a geostationary satellite is to change the plane of the orbit at an initial average rate of change of 0.85°/year from the equatorial plane (Gordon and Morgan 1993).

When both the sun and moon are acting on the same side of the satellite's orbit, the rate of change of the plane of the geostationary satellite's orbit will be higher than average. When they are on opposite sides of the orbit, the rate of change of the plane of the satellite's orbit will be less than average. Examples of maximum years are 1988

Table 2.2 Comparative data for the sun, moon, and earth

	Mean radius	Mass	Mean orbit radius	Spin period
Sun	696 000 km	333 432 units	30 000 light years	25.04 earth days
Moon	3 476 km	0.012 units	384 500 km	27.3 earth days
Earth	6 378.14 km	1.0 units	149 597 870 km	1 earth day

Note: The orbit radius refers to the center of the home galaxy (Milky Way) for the sun, center of earth for the moon, and center of the sun for the earth, respectively.

and 2006 (0.94°/year) and examples of minimum years are 1997 and 2015 (0.75°/year) (Gordon and Morgan 1993). These rates of change are neither constant with time nor with inclination. They are at a maximum when the inclination is zero and they are zero when the inclination is 14.67°. From an initial zero inclination, the plane of the geostationary orbit will change to a maximum inclination of 14.67° over 26.6 years. The acceleration forces will then change direction at this maximum inclination and the orbit inclination will move back to zero in another 26.6 years and out to −14.67° over a further 26.6 years, and so on.

In some cases, to increase the orbital maneuver lifetime of a satellite for a given fuel load, mission planners deliberately place a satellite planned for geostationary orbit into an initial orbit with an inclination that is substantially larger than the nominal 0.05° for a geostationary satellite. The launch is specifically timed, however, so as to set up the necessary precessional forces that will automatically reduce the inclination error to close to zero over the required period without the use of any thruster firings on the spacecraft. This will increase the maneuvering lifetime of the satellite at the expense of requiring greater tracking by the larger earth terminals accessing the satellite for the first year or so of the satellite's operational life.

Under normal operations, ground controllers command spacecraft maneuvers to correct for both the in-plane changes (longitudinal drifts) and out-of-plane changes (inclination changes) of a satellite so that it remains in the correct orbit. For a geostationary satellite, this means that the inclination, ellipticity, and longitudinal position are controlled so that the satellite appears to stay within a box in the sky that is bounded by ±0.05° in latitude and longitude over the subsatellite point. Some maneuvers are designed to correct for both inclination and longitude drifts simultaneously in the one burn of the maneuvering rockets on the satellite. In others, the two maneuvers are kept separate: one burn will correct for ellipticity and longitude drift; another will correct for inclination changes. The latter situation of separated maneuvers is becoming more common for two reasons. The first is due to the much larger velocity increment needed to change the plane of an orbit (the so-called north–south maneuver) as compared with the longitude/ellipticity of an orbit (the so-called east–west maneuver). The difference in energy requirement is about 10 : 1. By alternately correcting for inclination changes and in-plane changes, the attitude of the satellite can be held constant and different sets of thrusters exercised for the required maneuver.

The second reason is the increasing use of two completely different types of thrusters to control N–S maneuvers on the one hand and E–W maneuvers on the other. In the mid-1990s, one of the heaviest items that was carried into orbit on a large satellite was the fuel to raise and control the orbit. About 90% of this fuel load, once on orbit, was to control the inclination of the satellite. Newer rocket motors, particularly arc jets and ion thrusters, offer increased efficiency with lighter mass. The first all-electric geostationary satellite was launched on 14 May 2015. The satellite, built by Boeing, used electric thrusters to achieve geostationary orbit once released from the Falcon 9 rocket. Initially, the low thrust, high efficiency electric thrusters were mainly used for N–S maneuvers leaving the liquid propellant thrusters, with their inherently higher thrust (but lower efficiency) for orbit raising and in-plane changes. For SmallSats and CubeSats, where there is little available on-orbit mass, electric propulsion has been universally adopted (see Chapter 8). Increasingly, the higher available on-orbit power available from solar arrays has made the use of electric propulsion more attractive for

all satellite maneuvers. In order to be able to calculate the required orbit maneuver for a given satellite, the controllers must have an accurate knowledge of the satellite's orbit. Orbit determination is a major aspect of satellite control.

Example 2.7 Drift With a Geostationary Satellite

A quasi-GEO satellite is in a circular equatorial orbit close to geosynchronous altitude.

The quasi-GEO satellite, however, does not have a period of one sidereal day: its orbital period is exactly 24 hours – one solar day.

Question: What is

(i) the radius of the orbit;
(ii) the rate of drift around the equator of the subsatellite point in degrees per (solar) day.

An observer on earth sees that the satellite is drifting across the sky; and
(iii) the movement of the satellite – is it moving toward the east or toward the west?

Answer

Part (i)

The orbital radius is found from Eq. (2.21), as in worked Example 2.4. Equation (2.21) gives the square of the orbital period in seconds (remembering that T here is one solar day)

$$T^2 = (4\pi^2 a^3)/\mu$$

Rearranging the equation, the orbital radius a is given by

$$\begin{aligned} a^3 &= T^2\mu/(4\pi^2) \\ &= 7.5371216 \times 10^{13}\text{km}^3 \end{aligned}$$

which gives

$$a = 42\,241.095 \text{ km}$$

Part (ii)

The orbital period of the satellite (one solar day) is longer than a sidereal day by 3 minutes 55.9 seconds = 235.9 seconds. This will cause the subsatellite point to drift at a rate of $360° \times 235.9/864\,00°$ per day or 0.983° per day.

Part (iii)

The earth moves toward the east at a faster rate than the satellite, so the drift will appear to an observer on the earth to be toward the west.

2.10 Orbit Determination

Orbit determination requires that sufficient measurements be made to determine uniquely the six orbital elements needed to calculate the future orbit of the satellite, and hence calculate the required changes that need to be made to the orbit to keep it within the nominal orbital location. Three angular position measurements are needed because there are six unknowns and each measurement will provide two equations. Conceptually, these can be thought of as one equation giving the azimuth and the other the elevation as a function of the six (as yet unknown) orbital elements.

The control earth stations used to measure the angular position of the satellites also carry out range measurements using unique time stamps in the telemetry stream or communications carrier. These earth stations are generally referred to as the TTC&M (Telemetry Tracking Command and Monitoring) stations of the satellite network. Major satellite networks maintain their own TTC&M stations around the world. Smaller satellite systems generally contract for such TTC&M functions from the spacecraft manufacturer or from the larger satellite system operators, as it is generally uneconomic to build advanced TTC&M stations with fewer than three satellites to control. Chapter 3 discusses TTC&M systems.

2.11 Space Launch Vehicles and Rockets

The second decade of the twenty-first century saw an extraordinary surge in both the development of spacecraft, mainly SmallSats (see Chapter 8), and launch vehicles. A total of 345 satellites were launched in 2017, of which 212 were commercially procured CubeSats for earth observation and meteorology. More than a third of the commercial launches were by US companies: US entities and US partners also own 803 of the 1738 operational satellites orbiting the earth at the end of 2017 (Irene Klotz 2018c). As significant as the upsurge in launches in the second decade of the twenty-first century was the evolution of launch vehicles: not only were they to be used to place a satellite into orbit as reliably as possible, but the intention was that major elements of the launch vehicles be recovered and used again. Note that making parts of a launcher re-usable lowers the available payload as a significant additional mass needs to be added to the rocket to bring at least the booster stage(s) back to a designated recovery area. However, the reduced payload mass is more than compensated for by the reduced expense of refurbishing a booster as opposed to having to build a brand new one. A number of approaches have been proposed for the re-use of launch vehicles. The first one that was successful was *Pegasus*.

Pegasus was the first privately developed launch vehicle (Northropgrumman.com 2018). It used a Lockheed 1011 *TriStar* to carry it under one wing up to an altitude around 40 000 ft., where it was released, and the first stage rocket motor ignited. Pegasus was not only the first privately developed launch vehicle; it was the first winged vehicle to exceed 8 times the speed of sound and the first air-launched rocket to place a satellite into orbit. First launched on 5 April 1990, Pegasus is still considered to be operational, although the last launch was on 15 December 2016. A typical mission placed 443 kg into orbit; over 43 missions, Pegasus orbited 93 satellites into LEO at an approximate cost per launch of $40M (Northropgrumman.com 2018). Compare this price to that currently advertised (Time.com 2018) for the SpaceX Falcon 9, which is $62M to place up to 22 800 kg into LEO and 8300 kg into a geosynchronous transfer orbit (GTO). There are two new proposals for air-launching rockets that will place satellites into LEO: one is a six-engined behemoth called *Stratolaunch* (Satellitetoday.com 2018a) that flew for the first time on 13 April 2019, becoming the largest aircraft in the world. The other to be used by Virgin Galactic (Satellitetoday.com 2018a) makes use of a Boeing 747 that has been retired from the Virgin Airways fleet. Clearly, a 747 can utilize many airfields around the world in a similar manner to Pegasus, which was air-launched from the United States, Europe, and the Marshall Islands (Northropgrumman.com 2018). Stratolaunch, on the other hand, will probably be restricted to only a few airfields that have

the required size and support facilities. Nevertheless, it is likely that both air-launched concepts will be utilized in some form, with the Stratolaunch mainly being used to carry multiple launchers while Virgin Galactic carries a single launcher on replenishment missions for a LEO constellation. Inserting a satellite into orbit has become so routine that it is hard to remember a time when each launch made headline news, whether successful or not. However, for a satellite launch to succeed, many facets need to come together simultaneously.

A satellite cannot be placed into a stable orbit unless two parameters are simultaneously correct: the velocity vector and the orbital height. There is little point in obtaining the correct height and not having the appropriate velocity component in the correct direction to achieve the desired orbit. A geostationary satellite, for example, must be in an orbit at a height of 35 786.03 km above the surface of the earth (42 164.17 km radius from the center of the earth) with an inclination of zero degrees, an ellipticity of zero, and a velocity of 3074.7 m/s tangential to the earth in the plane of the orbit, which is the earth's equatorial plane. The further out from the earth the orbit is, the greater the energy required from the launch vehicle to reach that orbit. In any earth satellite launch, the largest fraction of the energy expended by the rocket is used to accelerate the vehicle from rest until it is about twenty miles (32 km) above the earth. To make the most efficient use of the fuel, it is common to shed excess mass from the launcher as it moves up through the atmosphere: this is called *staging*. As noted earlier in this chapter, air-launching a rocket provides two advantages over a vertical launch from a pad on the surface of the earth: a significant portion of the atmosphere is below the rocket, and the airplane has imparted a horizontal velocity vector to augment that of the rocket stage(s).

Most launch vehicles have multiple stages and, as each stage completes its burn that portion of the launcher is expended until the final stage places the satellite into the desired trajectory. Hence the term: Expendable Launch Vehicle (ELV), and more recently (2017) the Evolved Expendable Launch Vehicle (EELV). The ELV work horse of the former Soviet Union was the *Proton*. It has launched more satellites than any other rocket (Wikipedia 2018b) but it is planned to be phased out around 2020 by the Angara rocket (Wikipedia 2018c). Part of the reason behind the change was that the Proton was launched from Baikonur, which is in Kazakhstan, and Russia wanted to have not only the manufactured components of its rockets, but also the launch site in Russia. Figure 2.15 gives a schematic of a Proton launch from the Russian Baikonur complex at Kazakhstan, near Tyuratam. In the example of a Proton launch shown in Figure 2.15, the rocket inserts the payload directly into GEO. As we shall see later in this chapter, most rockets do not launch GEO satellites directly into GEO, but leave the spacecraft in an initial geostationary transfer orbit (GTO). The satellite then completes the circularization and inclination adjustments by using either an apogee kick motor (AKM) or internal guidance rockets to arrive at the correct longitude in GEO. Of equal importance to the orbital height the satellite is intended for is the inclination of the orbit that the spacecraft needs to be launched into.

The earth spins toward the east. At the equator, the rotational velocity of a sea level site in the plane of the equator is $(2\pi \times \text{radius of earth})/(\text{one sidereal day}) = 0.4651$ km/s. This velocity increment is approximately 1000 mph (~1610 km/h). An easterly launch from the equator therefore has a velocity increment of 0.465 km/s imparted by the rotation of the earth. A satellite in a circular, equatorial orbit at an altitude of 900 km requires an orbital velocity of about 7.4 km/s tangential to the surface of the earth. A rocket launched

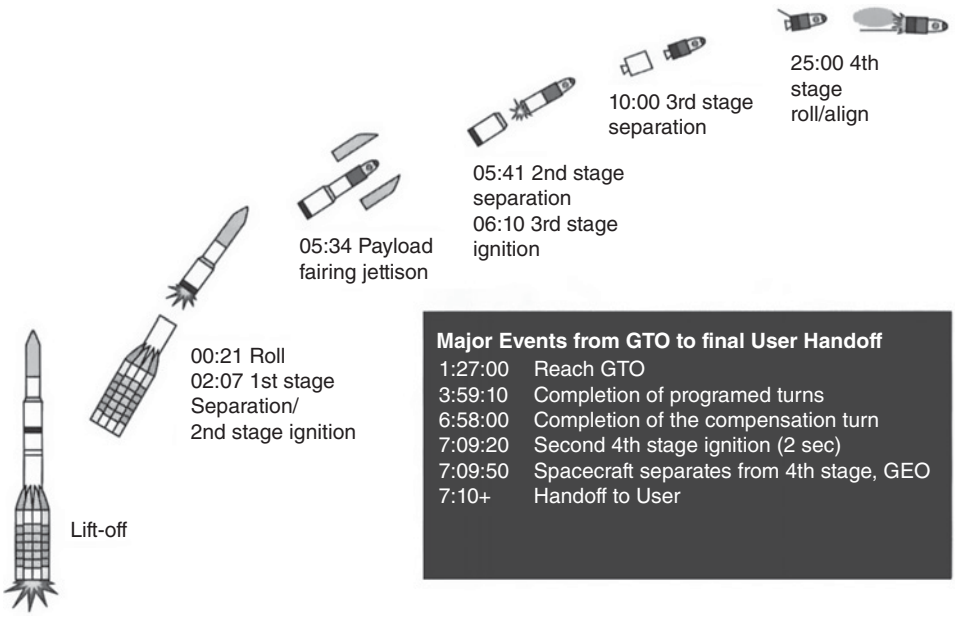


Figure 2.15 Launch sequence of a Proton rocket. After (Walsh and Groves 1997).

from the equator needs to impart an additional velocity of (7.4–0.47) km/s = 6.93 km/s: in other words, the equatorial launch has reduced the energy required by about 6%. This equatorial launch bonus led to the European Space Agency (ESA) choosing *Kourou*, a launch site 5.2° north of the equator in French Guiana. It also led to the concept of a sea launch by Hughes and Boeing (Roundtree 1999). The floating platform was towed out until it was at the equator in order to maximize the added spin of the earth. After some successes, Sea Launch was purchased by Russia and there are plans to have *Zenit* rockets launched from the platform by the end of 2019. If a rocket launch is not to place a satellite into an equatorial orbit, the payload capabilities of any given rocket will reduce as the required orbital inclination increases. An orbit in the same direction as the spin of the earth is called a *prograde* orbit, while one that orbits in the opposite sense is called a r*etrograde* orbit.

A satellite launched into a prograde orbit from a latitude of Φ degrees will enter an orbit with an inclination of Φ degrees to the equator. If the satellite is intended for geostationary orbit, the satellite must be given a significant velocity increment to reorient the orbit into the earth's equatorial plane. For example, a satellite launched from Cape Canaveral at 28.5°N latitude requires a velocity increment of 366 m/s to attain an equatorial orbit from a geosynchronous orbit plane of 28.5°. Ariane rockets are launched from the Guiana Space Center in French Guiana, located at latitude of about 5°N in South America, and Sea Launch can launch from the equator. The lower latitude of these launch sites results in savings in the fuel required by the AKM. Of probably more significance than the additional velocity increment provided by the spin of the earth, a launch close to the equator of a satellite intended for a geostationary orbit is the much lower energy needed to change the plane of the orbit from an inclination of 5.2° (the case for Kourou) to zero inclination. For a given spacecraft, a change in plane uses

approximately 10 times more fuel than a change in velocity in the same plane for a given angular change.

The Space Transportation System, or the *Space Shuttle* as it became known, could launch approximately 65 000 lb. (29 478 kg) into a standard 28.5° orbital inclination at an orbital height of about 200 km from the Kennedy Space Flight Center in Cape Canaveral. If the Vandenberg Air Force Base launch site in California still had the capability of launching the Space Shuttle, the payload capability for a polar launch (inclination 90°) would have been reduced to ~32 000 lb. (14 512 kg). The Space Shuttle was the fourth manned spaceflight program of NASA that achieved orbit (the others being Mercury, Gemini, and Apollo). It was also the first manned vehicle with wings that achieved orbit. After the Challenger accident in January 1996, the shuttle was rarely used to launch civilian payloads, its missions being confined to military payloads (e.g., TDRSS satellites), joint ventures with other agencies (e.g., ESA Spacelab facility), big science missions (e.g., the X-ray telescope Chandra), and ISS flights. In February 2003, the Space Shuttle Columbia broke up on its return to earth due to damage to the wing by ice pieces falling off the external fuel tank on launch. As a result, NASA's safety oversight panel instituted rigid rules for manned launches that were never met by the Space Shuttle and are unlikely to be met by either SpaceX or Boeing. The NASA Commercial Crew requirement is that there should be a 1 in 500 chance of a crew fatality during launch and an overall 1 in 270 chance of a fatality over a 210 day flight. The risk assessment for manned launches showed that the greatest danger was micrometeoroid damage while docked to the ISS or from a parachute deployment failure on landing (Irene Klotz 2018b). The last Space Shuttle flight, STS-135, started on 21 July 2011. The vast majority of the US satellite launches have therefore been conducted by what are referred to as *ELVs*.

2.11.1 Expendable Launch Vehicles (ELVs)

1998 was an important year for ELVs: it was the year when the number of commercial launches in the United States surpassed the number of government launches for the first time (Dekok 1999). The gap between commercial and government launches will continue to grow, particularly with the rapid increase in SmallSat launches. A total of 81 countries can claim to have a satellite that was successfully launched into LEO (Teal Group 2018), although the vast majority of the launches were made by other countries (e.g., the United States, Russia, and ESA). Only 11 countries have built their own satellite and rocket and completed a successful launch, the most recent at the time of writing (July 2018) being New Zealand (Wikipedia 2018c). Most of these satellites have long since re-entered the earth's atmosphere, although – paradoxically – the first successful US satellite, *Explorer 1*, launched in February 1958, is still in orbit. The Teal Group estimated in March 2018 that 12 230 satellites have been launched (Teal Group 2018), and the most recent estimate (15 November 2017) by the United Nations Office of Outer Space Affairs (UNOOSA) is that 4635 satellites are currently in earth orbit (Pixialytics.com 2018). Interestingly, the accelerating trend of adopting electric propulsion and digital payloads for SmallSats has reduced the average mass of these spacecraft to less than 50 kg (Satellitetoday.com 2018b). There is therefore a healthy market for ELVs and a number of companies, consortia, and national entities from around the world are seeking to enter this expanding field, particularly in the United States (Klotz 2019). This has

Table 2.3 Small lift launchers

Rocket	Height	Payload to LEO	Cost per launch	First launch
ISRO PSLV rocket	144 ft. 44 m	3 800 kg 8 400 lb	US$21M to US$31M per launch	Latest version flew 22 October 2008
Rocket Labs Electron rocket	56 ft. 17 m	100–225 kg 220–496 lb	US$5–6M per launch	21 January 2018
Vega	98 ft. 30 m	1 500 kg	US$37M per launch	13 February 2002
Minotaur C[a] (Taurus before)	92 ft. 28 m	1 590 kg 3 500 lb	US$40M to US$55M	13 March 1994
SS-520-S[b]	9.54 m 31.3 ft	3 kg 66 lb	~US$1M	3 February 2018

[a] The Minotaur C is a vertically launched version of the winged Pegasus launch vehicle.
[b] The SS-520-S is a converted sounding rocket that is launched along a rail. The first flight achieved orbit in less than 4.5 minutes.

led to both a search for new launch sites, and for a new family of rockets that are optimally sized for the satellite mass and the desired orbit.

There are currently 22 active rocket launch sites in the United States (FAA.gov 2018a) and this is likely to increase with the number of smaller rockets becoming operational for SmallSat launches. In addition, there will be several airfields that will be used for air-launched rockets. The increase in the number of rocket launches is complicating matters for air traffic control over the United States as it is essential to monitor in real time the location of aircraft that are likely to fly close to a scheduled rocket launch. In 2018 there were 42 000 Federal Aviation Authority (FAA)-controlled aircraft flights per day (FAA.gov 2018b), and this does not include smaller aircraft that fly from uncontrolled air fields, more correctly called *non-towered airports* (AOPA.org 2018). A non-towered airport does not have an operating control tower and requires pilots to strictly observe operation proceedings as set down by the FAA. About 500 airports in the United States have control towers, while there are nearly 20 000 non-towered airfields (AOPA.org 2018). The space launch systems being proposed to orbit spacecraft were initially divided into three broad payload categories: *small lift* (<2000 kg), *medium lift* (>2000 and <22 000 kg), and *heavy lift* (>22 000 kg). As the launch capacity of new rockets has grown, the third category has been split into two categories *heavy lift* (>22 000 and <40 000 kg) and *super heavy lift*, (>40 000 kg). The tables below list the main rockets being used, or proposed to be used, for launching satellites. To bring a common baseline to the data, the launch mass to LEO is used as a comparison parameter. Reference (Wikipedia 2018d) gives an extensive list of all orbital vehicles through mid-2018.

Tables 2.3 to 2.6 provide details on the different rockets used for launching satellites. Table 2.7 gives information on air-launched vehicles, Table 2.8 lists sub-orbital tourist rockets, and Table 2.9 provides a price comparison of the various launch vehicles for LEO satellites. Not included in Table 2.3 are proposals to have high altitude platforms (HAPs) deployed for emergency communications over areas that have suffered damage due to earthquakes or severe flooding. These can be tethered balloons or semi-rigid inflatable craft that execute HALO orbits (Tables 2.4–2.8).

Table 2.4 Medium lift launchers

Rocket	Height	Payload to LEO	Cost per launch	First launch
Ariane 5[a]	179 ft. 54.7 m	21 000 kg 46 297 lb	US$165M	9 March 2008
Ariane 6	207 ft. 63 m	21 500 kg 47 400 lb	US$100M	First launch scheduled for 2020
Soyuz[b]	150 ft. 45.6 m	6 450 kg 14 220 lb	US$81M	28 November 1966
Zenit 2[c]	187 ft. 57 m	13 740 kg 30 290 lb	~US$55M	13 April 1985

[a]There were four variants before Ariane 5, starting with Ariane 1, first launched 24 December 1974.
[b]The Soyuz rocket is the launch vehicle used to send astronauts (US, Russian, and other nations) to the ISS. The price per astronaut varies but was US$75M in mid-2018. The payload capability will increase for a Soyuz launch from Kourou.
[c]The price and payload capability given is for a Zenit 2 launched from Baikonur.

Table 2.5 Heavy lift launchers

Rocket	Height	Payload to LEO	Cost per launch	First launch
Falcon 9[a]	233 ft. 71 m	22 800 kg 50 300 lb	US$62M	7 June 2010
Proton M	191 ft. 58.2 m	23 000 kg 51 000 lb	US$65M	9 March 2008
Delta heavy	236 ft. 72 m	28 970 kg 63 470 lb	US$350M	21 December 2004

[a]There are a number of *blocks* of Falcon 9 rockets; the most recent (2018) is Block 5. This is the version slated to fly the Falcon Crew capsule. The Block 5 rockets are designed to fly 10 times. The first recovery of a Falcon 9 first stage took place on 21 December 2015.

In Table 2.9, the cost per kg was calculated by using the published mass of a satellite launched by the rocket in question and the published cost of the same launch. The numbers were for a LEO with a circular orbit at an altitude of about 500 km. No inclination change was factored into the numbers. A rocket capable of launching a satellite into GTO can launch a satellite into LEO that has a mass approximately 2.75 heavier. Thus the cost numbers given in Table 2.9 for a LEO satellite have been increased by a factor of 2.75, and these numbers are given in the GTO launch column of Table 2.9.

It is interesting to note the values in the GTO column of Table 2.9 and compare them with the trend line shown in Figure 2.16 of around US$12 000 per pound (US$26 450 per kg) to GTO. The data in Figure 2.16 are for 1996 dollars. Using the US consumer price index, US$100 in 1996 is equivalent to US$16 060 in 2018 (Exchange rates 2018), hence the US$26 450 per kilo cost in 1996 would be equal to US$42 480 per kg. Only one launcher in Table 2.9 is above this cost, and many are well below it. Cost is usually only one of the launch vehicle selection factors, as can be seen in Table 2.10 and Figure 2.17. While not the same as buying a jet airliner, there are some similarities. For example, a single brand new Boeing 777-300ER cost US$320M in July 2018, but if an airline wanted

Table 2.6 Super heavy lift launchers

Rocket	Height	Payload to LEO	Cost per launch	First launch
Falcon heavy	230 ft. 70 m	63 800 kg 140 700 lb	US$90M	6 February 2018
New glenn 3-stage	312 ft. 72 m	45 000 kg 99 000 lb	Not available	Scheduled for 2020
Space launch system B2	365 ft. 95 m	130 000 kg 286 601 lb	~US$500M	Scheduled for 2020
Saturn 5	363 ft. 110.6 m	140 000 kg 310 000 lb	US$1.16B	7 November 1967
Long March 9[a]	331 ft. 101 m	140 000 kg 310 000 lb	US$40M to US$55M	Scheduled for 2020
BFR[b]	348 ft. 106 m	250 000 kg 550 000 lb	Not available	First sub-orbital test flights scheduled for 2019

[a] There has been a long series of Long March rockets. The latest, Long March 11 will be all solid fueled and the complete rocket can be stored for long periods, leading to speculation that it is designed for rapid response.
[b] The BFR is either a two stage vehicle (data for which are in the table above) or it can be just a single stage. In a lightly loaded version, it can achieve orbit without the booster stage, leading to a single-stage-to-orbit rocket. The two stage version can also be configured to carry 200 passengers anywhere on the earth in 90 minutes.

Table 2.7 Aircraft launchers

Aircraft	Rocket	Payload to LEO	Cost per launch	First launch
VOX Space[a]	Launcher one	~500 kg ~1 100 lb	Not known but competitive	Launcher one has yet to be tested or flown
Stratolaunch[b]	Not yet available	Small lift to medium lift	Not known but competitive	First taxi run on 21 December 2017

[a] Virgin Orbit X (VOX) consists of a Boeing 747 mother ship that carries a two stage rocket, *Launcher One*, under one wing.
[b] Stratolaunch, founded in 2011 by Paul Allen, is being built by Scaled Composites, a Northrop Grumman subsidiary. It has two bodies and six engines. It is being designed to carry up to three small launchers (similar to *Pegasus*) for small lift satellites, and one larger launcher to orbit medium lift satellites.

Table 2.8 Sub-orbital tourist vehicles

Aircraft	Rocket	Payload to LEO	Cost per launch	First launch
Blue shepard[a]	~22 m (with capsule)	Six passengers to more than 75 miles (48 km)	~US$200 000 per passenger	29 April 2015
Virgin galactic VSS unity[b]	Air-dropped from White Knight mother ship	2 pilots and 6 passengers to >50 miles (80 km)	US$250 000 per passenger	6 April 2018

[a] Blue Shepard is a fully reusable, single stage rocket. It was the first rocket to successfully soft land back at the launch site.
[b] VSS Unity is the second SpaceShip Two to be completed; the first crashed in February 2016.

Table 2.9 Comparison of the price per kg to launch a satellite into LEO

Launch vehicle	Price per kg to LEO	Price per kg to GTO
SS-520-S[a]	US$333 300	Not capable of GTO
Rocket labs electron rocket[b]	US$26 650 to US$50 000	Not capable of GTO
Minotaur C (taurus before)	US$25 150 to US$34 590	Not capable of GTO
Vega	US$24 650	US$67 790
Soyuz	US$12 560	US$34 540
Delta heavy	US$12 080	US$33 220
ISRO PSLV rocket	US$5 525–8 150	US$15 190–22 410
Ariane 5	US$7 850	US$21 590
New glenn 3-stage[c]	US$5 555	US$15 280
Ariane 6	US$4 650	US$12 790
Zenit 2	US$4 000	US$11 000
Space launch system B2	US$3 850	US$10 590
Falcon 9	US$2 720	US$7 480
Proton M	US$2 825	US$7 770
Falcon heavy	US$1 410	US$3 880
BFR[c]	US$1 000	US$2 750
Long March 9[c]	US$535	US$1 470

[a]The launch cost is only US$1 000 000 but the 3 kg payload drives up the per kg cost.
[b]The launch cost is only US$5–6M, but the payload is quite small, hence the high cost per kg.
[c]Assumed US$250 000 000 per launch.

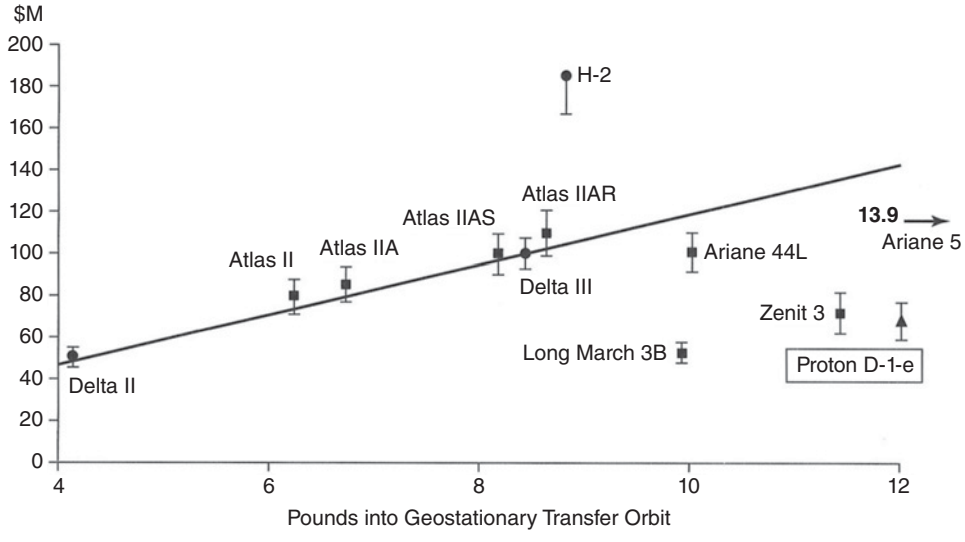


Figure 2.16 Launch vehicle market price versus performance, 1996 prices. After (Walsh and Groves 1997). The launch vehicles have been normalized to a launch into geostationary transfer orbit at an inclination of 28°. The trend line is at US$12 000 per pound. Note that Long march, Zenit, and Proton are well below this trend line, mainly due to aggressive pricing objectives to break into a market long dominated by US and European launchers.

Table 2.10 Some launch vehicle selection factors (Walsh and Groves 1997)

Price/cost
Reliability
Recent launch success/failure history
Dependable launch schedule
Urgency of your launch requirements
Performance
Spacecraft fit to launcher (size, acoustic, and vibration environment)
Flight proven (see recent launch history)
Safety issues
Launch site location
Availability
What is the launcher backlog of orders?
What is the launch site backlog of launchers?
Market issues
What will the market bear at this particular time?

to purchase 20 of these aircraft, there would almost certainly be significant discounts. The same is true for satellite launches.

Some of the launch vehicles deliver the spacecraft directly to geostationary orbit (called a direct-insertion launch) while others inject the spacecraft into a GTO. Spacecraft launched into GTO must carry additional rocket motors and/or propellant to enable the satellites to reach their designated location in geostationary orbit. There are three basic ways to achieve geostationary orbit.

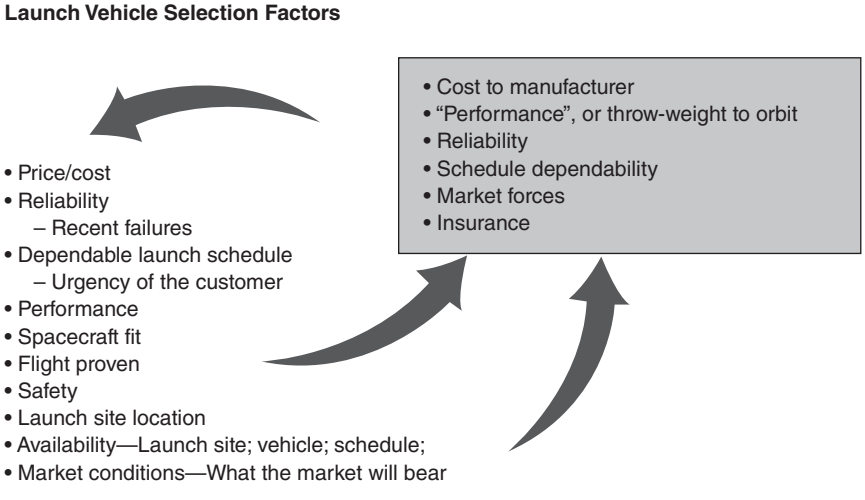


Figure 2.17 Schematic of the decision-making process to select a rocket for a given satellite requirement. After (Walsh and Groves 1997).

2.12 Placing Satellites Into Geostationary Orbit

2.12.1 Geostationary Transfer Orbit and AKM

The initial approach to launching geostationary satellites was to place the spacecraft, with the final rocket stage still attached, into LEO. After a couple of orbits, during which the orbital elements are measured, the final stage is re-ignited and the spacecraft is launched into a GTO. The GTO has a perigee that is the original LEO orbit altitude and an apogee that is the GEO altitude. Figure 2.18 illustrates the process. The position of the apogee point is close to the orbital longitude that would be the in-orbit test location of the satellite prior to it being moved to its operational position. Again, after a few orbits in the GTO while the orbital elements are measured, a rocket motor (usually contained within the satellite itself) is ignited at apogee and the GTO is raised until it is a circular, geostationary orbit. Since the rocket motor fires at apogee, it is commonly referred to as the AKM. The AKM is used both to circularize the orbit at GEO and to remove any inclination error so that the final orbit of the satellite is very close to geostationary.

The first successful GEO satellite was Syncom, launched in 1963. Hughes Corporation built the satellite and the spacecraft was spin stabilized while it was in GTO. In this way, the satellite was correctly aligned for the apogee motor firing. The apogee motor was fairly powerful and the apogee burn was only for a few minutes. During this apogee burn, all of the satellite's deployable elements (e.g., solar panels, antennas) were stowed and locked in place to avoid damage while the AKM accelerated the satellite to GEO. Hughes patented the technique of spin stabilizing the spacecraft in GTO. To avoid infringing this patent, other satellite manufacturers developed a new way to achieve GEO, known as a slow orbit raising technique.

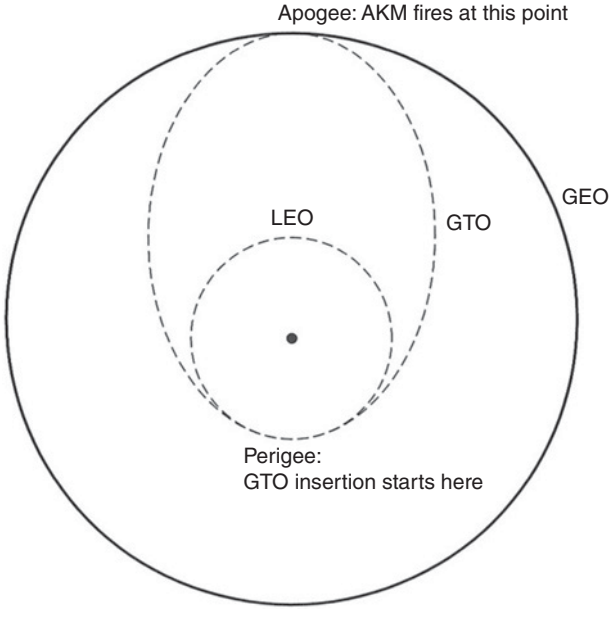


Figure 2.18 Illustration of transfer to geostationary orbit using an apogee kick motor (AKM). (Not to scale.) The spacecraft and final rocket stage are placed in low earth orbit (LEO). After careful orbital determination measurements, the final rocket stage is fired and the satellite placed in an elliptical geostationary transfer orbit (GTO) with apogee at geostationary altitude. The spacecraft is then separated from the rocket casing. After further careful orbital determination measurements, the AKM is fired several times to make the orbit circular, in the earth's equatorial plane, and at the correct altitude. The satellite is now in geostationary orbit (GEO).

Example 2.8

Question: What is the difference, or are the differences, between a *geosynchronous* satellite and a *geostationary* satellite orbit? What is the period of a geostationary satellite? What is the name given to this orbital period? What is the velocity of a geostationary satellite in its orbit? Give your answer in km/s.

A particular launch from Cape Canaveral released a TDRSS satellite into a circular low orbit, with an orbital height of 270 km. At this point, the TDRSS orbit was inclined to the earth's equator by approximately 28°. The TDRSS satellite needed to be placed into a GTO once released from the launch adaptor, with the apogee of the GTO at geostationary altitude and the perigee at the height of the original circular orbit.

(i) What was the eccentricity of the GTO?
(ii) What was the period of the GTO?
(iii) What was the difference in velocity of the satellite in GTO between when it was at apogee and when it was at perigee?

Note: Assume the average radius of the earth is 6378.137 km and Kepler's constant has the value $3.986\,004\,418 \times 10^5$ km^3/s^2.

Answer

A *geostationary* satellite orbit is one that has zero inclination to the equatorial plane, is perfectly circular (eccentricity is zero), and is at the correct orbital height to remain apparently stationary in orbit as viewed from the surface of the earth. A *geosynchronous* satellite orbit has most of the attributes of a geostationary orbit, but is either not exactly circular, not in the equatorial plane, or not at exactly the correct orbital height.

From Table 2.1, the orbital period of a geostationary satellite is 23 hours, 56 minutes, and 4.1 seconds.

The orbital period of a geostationary satellite is called a sidereal day.

From Table 2.1, the velocity of a geostationary satellite is 3.0747 km/s.

(i) The GTO will have an apogee of 35 786.03 km (the geostationary altitude) and a perigee of 270 km (the release altitude of the TDRSS).
The semimajor axis

$$a = (2r_e + h_p + h_a)/2 = (2 \times 6378.137 + 270 + 35\,786.03)/2 = 24\,406.152 \text{ km}$$

From Eq. (2.27) and Example 2.5, $r_0 = r_e + h_p$ and the eccentric anomaly $E = 0$ when the satellite is at perigee. From Eq. (2.27) $r_0 = a(1-e\text{cosE})$, with cosE = 1. Therefore, $r_e + h_p = a(1 - e)$ and, rearranging the equation,

$$e = 1 - (r_e + h_p)/a = 1 - (6378.137 + 270)/24\,406.152 = 0.727\,604.$$

The eccentricity of the GTO is therefore 0.728.

(ii) The orbital period

$$T = ((4\pi^2 a^3)/\mu)^{1/2} = ((4\pi^2 \times 24\,406.152^3)/3.986\,004\,418 \times 10^5)^{1/2}$$
$$= 37\,945.471\,02 \text{ seconds} = 10 \text{ hours } 32 \text{ minutes } 25.47 \text{ seconds.}$$

(iii) Orbital velocities. Eq. (2.5) gives the orbital velocity of a satellite as $v = (\mu/r)^{1/2}$. The perigee value of r = 270 + 6378.137 = 6648.137 km and the apogee value of r = 35 786 + 6478.137 = 42 164.137 km. Using these values in Eq. (2.5) yields a perigee velocity of 7.743117 km/s and an apogee velocity of 3.074660 km/s. The difference in velocity between perigee and apogee is 4.67 km/s.

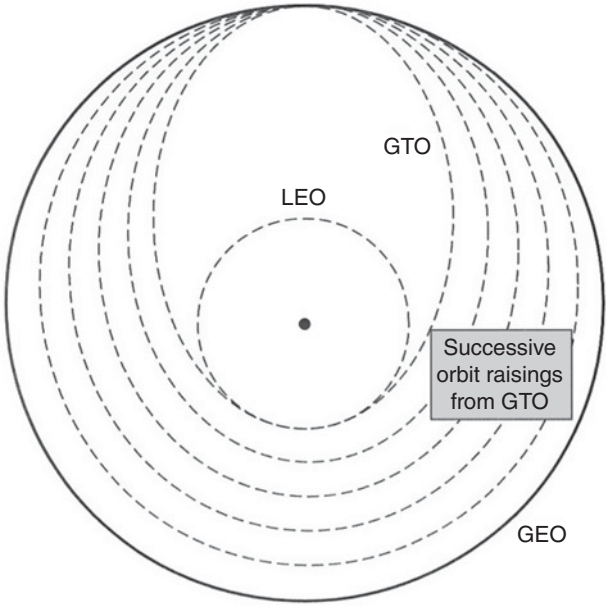


Figure 2.19 Illustration of slow orbit raising technique to geostationary orbit using an ion thrusters. (Not to scale.) The spacecraft and final rocket stage are placed in low earth orbit (LEO) and the satellite is separated from its rocket. The solar panels, antennas, and momentum wheels are deployed so that the satellite can be set to its correct attitude to generate solar power. Ion thrusters are then used to slowly increase the altitude of the satellite until geostationary altitude is achieved. At the same time, other ion thrusters are used to move the satellite's orbit into the equatorial plane. The process may take several months, but significantly reduces the weight of chemical fuel that the satellite has to carry.

2.12.2 Geostationary Transfer Orbit With Slow Orbit Raising

In this procedure, rather than employ an AKM that imparts a vigorous acceleration over a few minutes, the spacecraft thrusters are used to raise the orbit from GTO to GEO over a number of burns. Since the spacecraft cannot be spin stabilized during the GTO (so as not to infringe the Hughes patent), many of the satellite elements are deployed while in GTO, including the solar panels. The satellite normally has two power levels of thrusters: one for more powerful orbit raising maneuvers and one for on-orbit (low thrust) maneuvers. Since the thrusters take many hours of operation to achieve the geostationary orbit, the perigee of the orbit is gradually raised over successive thruster firings. The thruster firings occur symmetrically about the apogee although they could occur at the perigee as well. The burns are typically 60–80 minutes long on successive orbits and up to six orbits can be used. Figure 2.19 illustrates the process.

In the above two cases, AKM and Slow Orbit Raising, the GTO may be a modified orbit with the apogee well above the required altitude for GEO. The excess energy of the orbit due to the higher-than-necessary altitude at apogee can be traded for energy required to raise the perigee. The net energy to circularize the orbit at GEO is therefore less and the satellite can retain more fuel for on-orbit operations. The use of an initial orbit insertion well above that needed for GEO occurs when the launch vehicle has the ability to add additional fuel at launch (due to a lighter satellite or the rocket has increased efficiency due to developments since the original launch agreement was signed).

2.12.3 Direct Insertion to GEO

This is similar to the GTO technique but, in this case, the launch service provider contracts to place the satellite directly into GEO. The final stages of the rocket are used to place the satellite directly into GEO rather than the satellite use its own propulsion system to go from GTO to GEO.

2.13 Orbital Effects in Communications Systems Performance

2.13.1 Doppler Shift

To a stationary observer, the frequency of a moving radio transmitter varies with the transmitter's velocity relative to the observer. If the true transmitter frequency (i.e., the frequency that the transmitter would send when at rest) is f_T, the received frequency f_R is higher than f_T when the transmitter is moving toward the receiver and lower than f_T when the transmitter is moving away from the receiver. Mathematically, the relationship shown in Eq. (2.44a) between the transmitted and received frequencies is

$$\frac{f_R - f_T}{f_T} = \frac{\Delta f}{f_T} = \frac{V_T}{v_p} \tag{2.44a}$$

or

$$\Delta f = V_T f_T / c = V_T / \lambda \tag{2.44b}$$

where V_T is the component of the transmitter velocity directed toward the receiver, $v_p = c$ the phase velocity of light ($2.997\,9 \times 10^8 \approx 3 \times 10^8$ m/s in free space), and λ is the wavelength of the transmitted signal. If the transmitter is moving away from the receiver, then V_T is negative. This change in frequency is called the *Doppler shift*, the *Doppler effect*, or more commonly just *Doppler* after the German physicist who first studied the phenomenon in sound waves. For LEO satellites, Doppler shift can be quite pronounced, requiring the use of frequency-tracking receivers. For geostationary satellites, the effect is negligible.

Example 2.9 Doppler Shift for a LEO Satellite

A LEO satellite is in a circular polar orbit with an altitude, h, of 1000 km. A transmitter on the satellite has a frequency of 2.65 GHz.

Question: Find

(i) The velocity of the satellite in orbit
(ii) The component of velocity toward an observer at an earth station as the satellite appears over the horizon, for an observer who is in the plane of the satellite orbit. Hence
(iii) Find the Doppler shift of the received signal at the earth station. Use a mean earth radius value, r_e, of 6378 km. The satellite also carries a Ka-band transmitter at 20.0 GHz.
(iv) Find the Doppler shift for this signal when it is received by the same observer.

Answer

Part (i)

The period of the satellite is found from Eq. (2.21):

$$T^2 = (4\pi^2 a^3)/\mu$$
$$T^2 = 4\pi^2 \times (6378 + 1000)^3 / 3.986004418 \times 10^5$$
$$= 3.977754 \times 10^7 \text{ s}^2$$

Giving $T = 6306.94$ s

The circumference of the orbit is $2\pi a = 46\,357.3$ km so the velocity of the satellite in orbit is v_s where

$$v_s = 46\,357.3/6\,306.94 = 7.350 \text{ km/s}$$

Part (ii)

The component of velocity toward an observer in the plane of the orbit as the satellite appears over the horizon is given by $v_r = v_s \cos\theta$, where θ is the angle between the satellite velocity vector and the direction of the observer at the satellite. The angle can be found from simple geometry to be:

$$\cos\theta = r_e/(r_e + h) = 6378/7378 = 0.8645$$

Hence the component of satellite velocity toward the observer is

$$v_r = v_s \cos\theta = 6.354 \text{ km/s} = 6354 \text{ m/s}$$

Part (iii)

The Doppler shift of the received signal is given by Eq. (2.44b). Hence, for this satellite and observer, with a transmitter frequency 2.65 GHz, $\lambda = 0.1132$ m, and the Doppler shift in the received signal is

$$\Delta_f = V_T/\lambda = 6354/0.1132 = 56\,130 \text{ Hz} = 56.130 \text{ kHz}$$

Part (iv)

A Ka-band transmitter with frequency 20.0 GHz has a wavelength of 0.015 m. The corresponding Doppler shift at the receiver is

$$\Delta_f = V_T/\lambda = 6354/0.015 = 423.60 \text{ kHz}.$$

Doppler shift at Ka-band with a LEO satellite can be very large and requires a fast frequency tracking receiver. Ka-band LEO satellites are better suited to wideband signals than narrowband voice communications.

2.13.2 Range Variations

Even with the best station-keeping systems available for geostationary satellites, the position of a satellite with respect to the earth exhibits a cyclic daily variation. The variation in position will lead to a variation in range between the satellite and user terminals. If Time Division Multiple Access (TDMA) is being used, careful attention must be paid to the timing of the frames within the TDMA bursts (see Chapter 6) so that the individual user frames arrive at the satellite in the correct sequence and at the correct time. Range variations on LEO satellites can be significant, as can path loss variations. While guard times between bursts can be increased to help in any range and/or timing inaccuracies, this reduces the capacity of the transponder. The onboard capabilities of some satellites permit both timing control of the burst sequence and power level control of individual user streams.

2.13.3 Solar Eclipse

A satellite is said to be in eclipse when the earth prevents sunlight from reaching it, that is, when the satellite is in the shadow of the earth. For geostationary satellites, eclipses occur during two periods that begin 23 days before the equinoxes (about 21 March

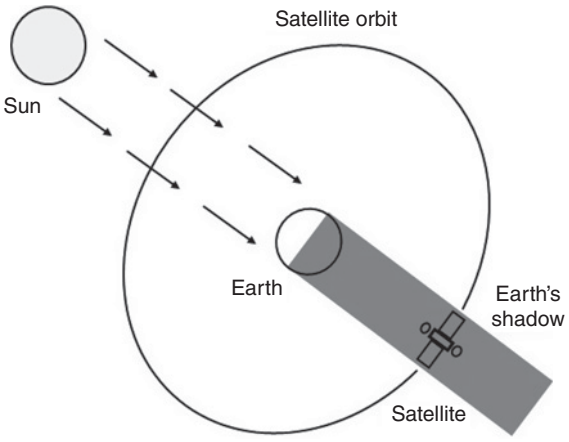


Figure 2.20 Illustration of an eclipse for a GEO satellite. The earth's shadow passes over the satellite twice each year, around the Spring equinox and the Fall equinox. The duration of the eclipse varies from a few minutes to over an hour. The maximum duration of the eclipse is 71 minutes and occurs around March 21 and September 23, as shown in Figure 2.21. This is when the sun crosses the orbital plane of the satellite.

and about 23 September) and end 23 days after the equinox periods. Figures 2.20 and 2.21 illustrate the geometry and duration of the eclipses. Eclipses occur close to the equinoxes, as these are the times when the sun, the earth, and the satellite are all nearly in the same plane.

During full eclipse, a satellite receives no power from its solar array and it must operate entirely from its batteries. Batteries are designed to operate with a maximum depth of discharge; the better the battery, the lower the percentage depth of discharge can be. If the battery is discharged below its maximum depth of discharge, the battery may not recover to full operational capacity once recharged. The depth of discharge therefore sets the power drain limit during eclipse operations. Nickel-hydrogen batteries, long the mainstay of communications satellites, can operate at about a 70% depth of discharge and recover fully once recharged. Lithium-ion batteries are increasingly being used in

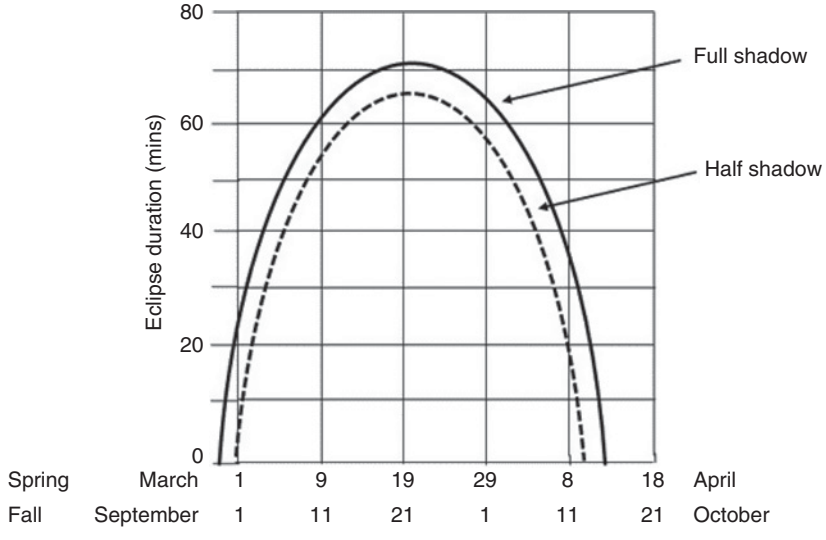


Figure 2.21 Dates and duration of eclipses for a GEO satellite. The longest period that the satellite is in darkness is around 20 March and 22 September, the Spring and Fall equinoxes.

satellites. They can be taken to 75% depth of discharge and do not have a *memory effect* (howstuffworks.com 2018). A memory effect in a battery occurs when it is recharged before being completely empty. A memory effect can decrease the life of a battery.

Ground controllers perform battery-conditioning routines prior to eclipse operations to ensure the best battery performance during the eclipse. The routines consist of deliberately discharging the batteries until they are close to their maximum depth of discharge, and then fully recharging the batteries just before eclipse season begins. The energy density of a nickel-hydrogen battery is only about one-third that of a lithium-ion battery, but it can handle 20 000 charge cycles (Wikipedia 2018g). While GEO satellites only have two eclipse seasons leading to no more than 90 total and partial eclipses, LEO and MEO satellites can have up to 5500 eclipse periods a year. Nickel-hydrogen batteries supplanted nickel-cadmium batteries because of their lighter mass, despite the higher cost and larger volume (esa.int 2018). Lithium-ion batteries have demonstrated similar depth of discharge capabilities as nickel-hydrogen batteries (i.e., to at least 70% depth of discharge without ill effects) and have the ability to handle 1000–2000 cycles, and so are now incorporated in GEO satellites. The Iridium Next series of LEO satellites incorporate 252 lithium-ion cells in each satellite (spaceflight101.com 2018).

The eclipse season is a design challenge for spacecraft builders. Not only is the main power source withdrawn (the sun) but also the rapidity with which the satellite enters and exits the shadow can cause extreme changes in both power and heating effects over relatively short periods. Just like a common light bulb is more likely to fail when the current is switched on as opposed to when it is under steady-state conditions, satellites can suffer many of their component failures under sudden stress situations. Eclipse periods are therefore monitored carefully by ground controllers, as this is when most of the equipment failures are likely to occur.

2.13.4 Sun Transit Outage

During the equinox periods, not only does the satellite pass through the earth's shadow on the unlit side of the earth, the orbit of the satellite will pass directly in front of the sun on the sunlit side of the earth. The sun is a strong microwave source with an equivalent temperature of about 6000–10 000 K, depending on the time within the 11-year sunspot cycle, at the frequencies used by communications satellites (4–50 GHz). The earth station antenna will therefore receive not only the signal from the satellite but also the noise temperature transmitted by the sun as shown in Figure 2.22. The added noise temperature will cause the fade margin of the receiver to be exceeded and an outage will occur. These outages may be precisely predicted. For satellite system operators with more than one satellite at their disposal, traffic can be offloaded to satellites that are just out of, or are yet to enter, a sun outage. The outage in this situation can therefore be limited as far as an individual user is concerned. However, the outages can be detrimental to operators committed to operations during daylight hours.

2.14 Manned Space Vehicles

There are currently four known space vehicles that will carry humans into space. An additional two concepts are included for completeness. Brief details on each are given below.

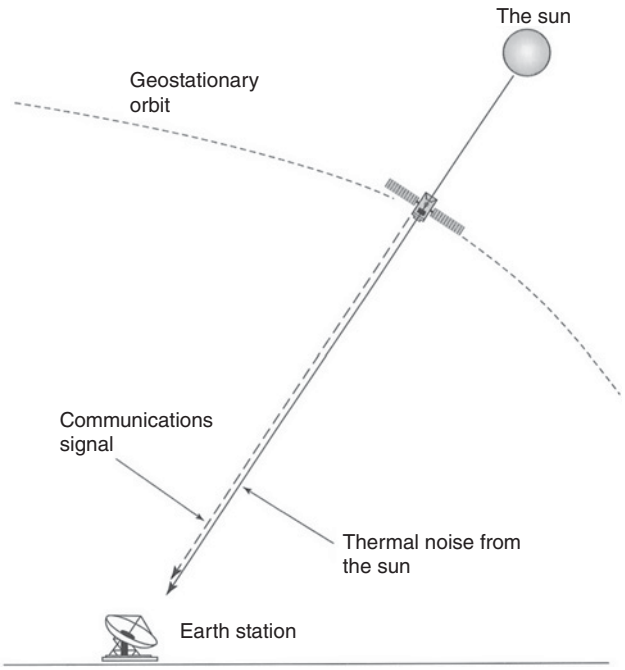


Figure 2.22 Illustration of a sun transit outage. During eclipse periods the sun will appear to pass behind the satellite. The sun is a powerful noise source with a noise temperature between 6000 and 10 000 K, depending on the time within a sunspot cycle. With the antenna beam covering the sun, the receiving system noise power will increase to a point where an outage occurs. With a typical broadcast television (DBS-TV) antenna, the outage lasts only a few minutes.

2.14.1 Dragon Crew

The SpaceX Dragon Crew capsule is a man-rated variant of the Dragon Cargo capsule (Wikipedia 2018e). The Dragon Cargo version has docked many times with the ISS, the first time on 22 May 2012, when it became the first private space vehicle to dock with the ISS. There will be seats for a crew of seven. The first, unmanned, launch of the Dragon Crew variant took place on 16 September 2018. A number of abort tests have been carried out successfully. The orbital launch will be on a man-rated version of the Falcon 9, probably in late 2019. Recovery of the manned capsule will be on water.

2.14.2 Boeing CST-100 Starliner

The CST-100 Starliner when man-rated will be able to carry up to seven crew members (Wikipedia 2018g). With the Dragon Crew capsule above, it was selected by NASA within that organization's Commercial Crew Development (CCDev) program. The first launch will probably be in late-2019.

2.14.3 Orion Deep Space Capsule

This spacecraft was originally conceived as an interplanetary vehicle with seating from two to six depending on the mission. A first mission occurred in December 2014, but there has been a hiatus since then. An abort test will occur in 2019 at the point in flight where maximum dynamic pressure occurs (called *Max Q*) (Arstechnica 2018).

2.14.4 Big Falcon Rocket (BFR)

The Big Falcon Rocket (BFR) is designed as a two stage vehicle capable of carrying cargo or passengers anywhere on earth, or indeed to the moon and beyond (Irene Klotz 2018a).

As a passenger vehicle for destinations on earth, up to 200 passengers could be flown to anywhere on earth in 90 minutes. Construction of the factory that will build the BFR has started in the Port of Los Angeles. Initial tethered tests of a boiler plate version of the first stage of the BFR were conducted in Texas in April 2019.

2.14.5 X-37B

As currently configured, the X-37B is a winged automated orbital vehicle, launched atop an Atlas V that has spent up to 718 days in space. That mission ended on 7 May 2017 (space.com 2018). A proposed new version, the X-37C, would be larger and have space to carry up to 6 astronauts.

2.14.6 Sierra Nevada Dream Chaser

The Dream Chaser vehicle was selected under the Commercial Resupply Service 2 (CRS2) to deliver supplies to the ISS. A minimum of 6 resupply missions are slated to take place between 2019 and 2024 (sncorp.com 2018). There are no firm plans at present (mid-2018) to modify the spacecraft to carry astronauts, but there appears to be no reason why this could not be done in the future.

2.15 Summary

Satellite launches have changed from being a major news event to a routine occurrence noted somewhere on what used to be called the inside pages of a national newspaper. In a like manner, the design and construction of launch vehicles has stopped being something only a national organization like NASA could undertake but a major portion of a privately held company's commercial business operation. The first recovery of a suborbital booster and that of the booster stage from an orbital launch was a commercial undertaking that succeeded within a month of each other in the last quarter of 2015. The ISS is resupplied with cargo by private launch vehicles and capsules designed and built in the United States. The next manned launch from US soil will be carried out by a private organization.

While it might be a common occurrence nowadays to launch satellites into orbit – or, on one occasion, a Tesla vehicle toward the orbit of Mars – the balance between the force pulling a satellite inward to the earth – that is, gravity – and that trying to fling a satellite away from the earth – kinetic energy – is a fine one. Newton's laws of motion explain the forces on a satellite in orbit, and to achieve stable orbit, a satellite must have the correct velocity, be traveling in the right direction, and be at the right height for its velocity. As the orbital height increases, the gravitational acceleration decreases, the orbital velocity decreases, and the period of the satellite increases. Calculation procedures for obtaining the period of a satellite and its velocity are set out and it is seen that Kepler's constant, the product of the universal gravitational constant, G, and the mass of the earth ME, is fundamental to many of the equations that give the forces on the satellite and the velocity of the satellite in its orbit. Kepler's three laws describing the motion of one body orbiting another are given and the terminology employed in satellite ephemeris data is explained.

Locating the satellite in its orbit is a complex process, with a number of possible frames of reference, and different approaches are discussed. Procedures for calculating the look angles from the earth to a geostationary satellite are given. The natural forces that act on

a satellite to cause orbital perturbations are set out and the need for orbital maneuvers explained. The important difference between Orbital Maneuver Life and Orbital Design Life of a spacecraft is explained. Details on launch procedures and launch vehicles are provided, with typical launch campaign information set out. The two basic methods of launching geostationary satellites are described, one using an AKM and the other a slow orbit raising technique. Finally, Doppler shift, range variations, solar eclipse, and sun transit outage are reviewed.

Exercises

2.1 Explain what the terms centrifugal and centripetal mean with regard to a satellite in orbit around the earth.
A satellite is in a circular orbit around the earth. The altitude of the satellite's orbit above the surface of the earth is 1500 km.
- **a.** What are the centripetal and centrifugal accelerations acting on the satellite in its orbit? Give your answer in m/s.
- **b.** What is the velocity of the satellite in this orbit? Give your answer in km/s.
- **c.** What is the orbital period of the satellite in this orbit? Give your answer in hours, minutes, and seconds.

Note: Assume the average radius of the earth is 6378.137 km and Kepler's constant has the value $3.986\,004\,418 \times 10^5$ km^3/s^2.

2.2 A satellite is in a circular orbit at an altitude of 350 km. Determine
- **a.** The orbital angular velocity in radians per second.
- **b.** The orbital period in minutes.
- **c.** The orbital velocity in meters per second.

Note: Assume the average radius of the earth is 6378.137 km and Kepler's constant has the value $3.986\,004\,418 \times 10^5$ km^3/s^2.

2.3 A LEO satellite is in a circular equatorial orbit with an altitude of 1000 km.
What is the orbital period in hours, minutes, and seconds to the nearest 1/100 second?

2.4 An observer is located on the equator at longitude 0°. How long is the LEO satellite described in Question 3 visible to this observer, assuming that the observer can see down to the horizon at zero degrees elevation? Give your answer in minutes and seconds to the nearest second.
Hint. This is a problem in geometry. Calculate the angle at the center of the earth that defines visibility. Then find the relative angular velocity of the satellite, assuming a prograde orbit (satellite travels in same direction as earth's rotation). Visibility time is central angle times angular velocity.

2.5 A LEO satellite has an apogee altitude of 5000 km and a perigee altitude of 800 km.
What is the eccentricity of the orbit?

2.6 What are Kepler's three laws of planetary motion? Give the mathematical formulation of Kepler's third law of planetary motion. What do the terms perigee and apogee mean when used to describe the orbit of a satellite orbiting the earth?

A satellite in an elliptical orbit around the earth has an apogee of 39 152 km and a perigee of 500 km. What is the orbital period of this satellite? Give your answer in hours, minutes, and seconds. Note: Assume the average radius of the earth is 6378.137 km and Kepler's constant has the value $3.986\,004\,418 \times 10^5$ km^3/s^2.

2.7 An observation satellite is to be placed into a circular equatorial orbit so that it moves in the same direction as the earth's rotation. Using a synthetic aperture radar system, the satellite stores data on weather related parameters as it flies overhead. These data will be downloaded to a controlling earth station after each trip around the world.
The orbit is designed so that the satellite is directly above the controlling earth station, which is located on the equator, once every four hours. The controlling earth station's antenna is unable to operate below an elevation angle of 10° to the horizontal in any direction. Taking the earth's rotational period to be exactly 23 hours 56 minutes 4.09 seconds, find the following quantities:
a. The satellite's angular velocity in radians per second.
b. The orbital period in hours, minutes, and seconds.
c. The orbital radius in kilometers.
d. The orbital height in kilometers.
e. The satellite's linear velocity in meters per second.
f. The time interval in minutes and seconds for which the controlling earth station can communicate with the satellite on each pass.

2.8 For a variety of reasons, typical minimum elevation angles used by earth stations operating in the commercial Fixed Services Satellite (FSS) communications bands are as follows:
C-Band 5°; Ku-Band 10°; Ka-Band 20°.
a. Determine the maximum and minimum range in kilometers from an earth station to a geostationary satellite in each of the three frequency bands.
b. What is the *round-trip* signal propagation times for each of the ranges? You may assume the signal propagates with the velocity of light in a vacuum even when in the earth's lower atmosphere.

2.9 Most commercial geostationary communications satellites must maintain their orbital positions to within ±0.05° of arc. If a geostationary satellite meets this condition (i.e., it has an apparent motion ±0.05° of arc N–S and ±0.05° of arc E–W, as measured from the center of the earth), calculate the *maximum* range variation to this satellite from an earth station with a mean elevation angle to the center of the satellite's apparent motion of 5°. You may assume that the equatorial and polar diameters of the earth are the same.

2.10 An interactive experiment is being set up between the University of York, England (latitude 53.5°N, longitude 0.5°W) and the Technical University of Graz, Austria (latitude 47.5°N, longitude 15°E) that will communicate through a geostationary satellite. The earth stations at both universities are constrained to work only above elevation angles of 20° due to buildings near their locations. The groups at the two universities need to find a geostationary satellite that will be visible to both universities simultaneously, with both earth stations operating

at, or above, an elevation angle of 20°. What is the range of subsatellite points between which the selected geostationary satellite must lie?

2.11 A GEO satellite is located at longitude 109° west. The satellite broadcasts television programming to the continental United States.
- **a.** Calculate the look angles for an earth station located near Blacksburg, Virginia, latitude 37.22°N, longitude 80.42°W.
- **b.** Calculate the look angles for an earth station located near Billings, Montana, latitude 46.00°N, longitude 109.0°W.
- **c.** Calculate the look angles for an earth station located near Los Angles, California, longitude 118.0°W, latitude 34.00°N.

2.12 A GEO satellite is located at longitude 343° (17° west), over the Atlantic Ocean. Communication is established through this satellite between two earth stations. One earth station is near Washington, D.C., at latitude 38.9°N, longitude 77.2°W. The other station is near Cape Town South Africa, at latitude 34.0°S, longitude 19.0°E.
- **a.** Calculate the look angles for each earth station.
  Don't forget that the earth station in Africa looks north, and that it has a longitude in degrees east. The Washington, D.C. station has a longitude in degrees west and looks south. The numerical values of the earth station longitudes must be added when finding the separation of the stations in longitude.
- **b.** Calculate the delay, in milliseconds, for a signal to travel from one earth station to the other via the GEO satellite.

2.13 A link is established through a GEO satellite at longitude 30°W between an earth station near Rio de Janeiro, Brazil, latitude 22.91°S, longitude 43.17°W, and an earth station near Santiago, Chile, latitude 33.45°S, longitude 70.67°W.
Calculate the look angles for each earth station.

2.14 A geostationary satellite system is built which incorporates intersatellite links (ISLs) between satellites. This permits the transfer of information between two earth stations on the surface of the earth, which are not simultaneously visible to any single satellite in the system, by using the ISL equipment to link up the satellites. In this question, the effects of ray bending in the atmosphere may be ignored, processing delays on the satellites may initially be assumed to be zero, the earth may be assumed to be perfectly circular with a flat (i.e., not hilly) surface, and the velocity of the signals in free space (whether in the earth's lower atmosphere or in a vacuum) may be assumed to be the velocity of light in a vacuum.
- **a.** What is the furthest apart two geostationary satellites may be so that they can still communicate with each other without the path between the two satellites being interrupted by the surface of the earth? Give your answer in degrees longitude between the subsatellite points.
- **b.** If the longest, one way delay permitted by the ITU between two earth stations communicating via a space system is 400 ms, what is the furthest apart two geostationary satellites may be before the transmission delay of the signal from one earth station to the other, when connected through the ISL system of the

two satellites, equals 400 ms? Assume that the slant path distance between each earth station and the geostationary satellite it is communicating with is 40 000 km.

c. If the satellites in part (b) employ onboard processing, which adds an additional delay of 35 ms in each satellite, what is the maximum distance between the ISL-linked geostationary satellites now?

d. If both of the two earth stations used in parts (b) and (c) must additionally now send the signals over a 2500 km optical fiber line to the end-user on the ground, with an associated transmission delay in the fiber at each end of the link, what is the maximum distance between the ISL-linked geostationary satellites now? Assume a refractive index of 1.5 for the optical fiber and zero processing delay in the earth station equipment and end-user equipment.

References

AOPA.org (2018). https://www.AOPA.org/-/media/files/AOPA/home/pilot-resources/asi/safety-advisors/sa08.pdf?la=en (accessed 26 July 2018).

Arstechnica (2018). https://arstechnica.com/science/2018/05/nasas-orion-spacecraft-getting-closer-to-finally-flying-again (accessed 27 July 2018).

Britannica.com (2018). https://www.britannica.com/technology/glider-aircraft (accessed 7 July 2018).

Dekok, R. (1999). *Spacelift in and Beyond the Next Millennium*, 6. Launchspace.

ESA.int (2018). http://www.esa.int/esapub/bulletin/bullet90/b90dudle.htm (accessed 13 July 2018).

Exchange rates (2018). http://www.in2013dollars.com/1996-dollars-in-2018 (accessed 25 July 2018).

FAA.gov (2018a). https://www.faa.gov/about/office_org/headquarters_offices/ast/industry/media/S (accessed 26 July 2018).

FAA.gov (2018b). https://www.faa.gov/air_traffic/by_the_numbers (accessed 13 July 2018).

Gordon, G.D., Walter, L., and Morgan, W.L. (1993). *Principles of Communications Satellites*. Wiley.

Howstuffworks.com (2018). https://auto.howstuffworks.com/lithium-ion-batteries-improve-hybrids1.htm (accessed 25 July 2018).

Irene, K. (2019). *Little Launchers Lining Up*, 70–71. Aviation Week & Space Technology.

Irene, K. (2018a). *SpaceX Aiming to Start BFR Tests next Year*, 24–26. Aviation Week and Space Technology.

Irene, K. (2018b). *The Launchpad*, 18. Aviation week and Space Technology.

Irene, K. (2018c). *The Launchpad*, 21. Aviation week and Space Technology.

n2yo.com (2018). www.n2yo.com (accessed 7 July 2018).

NASA.gov (2018a). https://history.msfc.nasa.gov/rocketry (accessed 7 July 2018).

Nasa.gov (2018b). https://opensource.gsfc.nasa.gov/projects/ODTBX (accessed 13 July 2018).

Northropgrumman.com (2018). https://www.northropgrumman.com/Capabilities/Pegasus/Pages/default.aspx (accessed 10 July 2018).

Pixialytics.com (2018). https://www.pixalytics.com/sats-orbiting-earth-2017 (accessed 7 July 2018).

Roundtree, K. (1999). *Launching Payloads by Sea*, 38–39. Launchspace.

Satellitetoday.com (2018a). https://www.satellitetoday.com/business/2018/01/17/virgin-orbit-signs-contract-launch-gomspace-nanosatellites (accessed 17 January 2018).
Satellitetoday.com(2018b). https://www.satellitetoday.com/innovation/2018/03/08/firing-it-up-at-both-ends-new-launch-vehicles-extend-mass-range/undefined (accessed 20 July, 2018).
sncorp.com (2018). https://www.sncorp.com/what-we-do/dream-chaser-space-vehicle (accessed 20 July, 2018).
Space.com (2018). https://www.space.com/34633-x-37b-military-space-plane-surprising-facts.html?utm_source=sdc-newsletter/utm_medium=email&utm_campaign=20180425-sdc (accessed 22 July, 2018).
Spaceflight101.com (2018). spaceflight101.com/spacecraft/iridium-next/ (accessed 26 July 2018).
STK.com (2018). www.stk.com (accessed 13 July 2018).
Teal Group (2018). http://www.tealgroup.com/index.php/about/press-releases (accessed 13 July 2018).
The American Ephemeris and Nautical Almanac (n.d.). U.S. Government Printing Office, Washington, DC (published annually).
Time.com (2018). time.com/money/5135565/elon-musk-falcon-heavy-rocket-launch-cost/ (accessed 13 July 2018).
Walsh, D. and Groves, C. (1997). Private communication, EE-4644.
Wertz, J.R. and Larson, W.J. (eds.) (1999). *Space Mission Analysis and Design*, 3e. Kluwer Academic Publishers.
Wikipedia.org (2018a). https://en.wikipedia.org/wiki/Hero_of_Alexandria (accessed 9 July 2018).
Wikipedia.org (2018b). https://en.wikipedia.org/wiki/Proton_(rocket_family) (accessed 10 July 2018).
Wikipedia.org (2018c). https://en.wikipedia.org/wiki/Angara_(rocket_family) (accessed 12 July 2018).
Wikipedia.org (2018d). https://en.wikipedia.org/wiki/Satellite (accessed 13 July 2018).
Wikipedia.org (2018e). https://en.wikipedia.org/wiki/Comparison_of_orbital_launch_systems (accessed 23 July 2018).
Wikipedia.org (2018f). https://en.wikipedia.org/wiki/Dragon_2 (accessed 25 July 2018).
Wikipedia.org (2018g). https://en.wikipedia.org/wiki/CST-100_Starliner (accessed 25 July 2018).

3

Satellites

Maintaining a microwave communication system in orbit in space is not a simple problem, so communications satellites are very complex, extremely expensive to purchase, and also expensive to launch. A typical large geostationary satellite, for example, is estimated to cost from US$100M to US$500M on station (see Chapter 2). A constellation of low earth orbit (LEO) satellites that can maintain continuous coverage and provide capacity that exceeds the capacity of a large geostationary earth orbit (GEO) satellite costs over US$2B. The cost of the satellites and launches is increased by the need to dedicate one or more earth stations to the monitoring and control of the satellite, at a cost of several million dollars per year. The revenue to pay these costs is obtained by selling the communication capacity of the satellite to users, either by way of leasing circuits or transponders, or by charging for circuit use, as in direct to home television (DTH-TV), international telephone connections, and data transmission services.

Communications satellites are usually designed to have an operating lifetime of 10–15 years. The operator of the system hopes to recover the initial and operating costs well within the expected lifetime of the satellite, and the designer must provide a satellite that can survive the hostile environment of outer space for that long. In order to support the communications system, the satellite must provide a stable platform on which to mount the antennas, be capable of station keeping, provide the required electrical power for the communication system, and also provide a controlled temperature environment for the communications electronics. In this chapter we discuss the subsystems needed on a satellite to support its primary mission of communications. We also discuss the communications subsystem itself in some detail, and other problems such as reliability. This chapter discusses satellites in geostationary orbit, which have provided the majority of satellite communication capacity from the 1970s through 2020, and also the non-geostationary satellite orbit (NGSO) satellite constellations being developed for internet access. Communications satellites for LEO are in most cases quite similar to GEO satellites and have similar requirements, but when produced in large quantities (exceeding 1000) there are economies of scale that can lower the unit cost. The discussion of satellites in this chapter is necessarily brief. For more details of the many subsystems used on satellites and their construction and operation the reader should refer to reference (Maral and Bousquet 2002). Much information about individual satellites can be found on the web sites of satellite manufacturers and operators.

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

3.1 Satellite Subsystems

The major subsystems required on the satellite are outlined below. Figure 3.1a shows an illustration of National Aeronautics and Space Administration (NASA)'s Tracking and Data Relay Satellite (TDRS) on orbit. Figure 3.1b shows an exploded view of the TDRS satellite with some of its subsystems labeled. There are nine TDRS satellites in geostationary orbit (2018) providing continuous links between NASA in the United States and low earth orbit spacecraft such as the International Space Station (ISS) and the Hubble astronomical telescope. Two early TDRS satellites have been moved into a *graveyard orbit* 300 km above geostationary altitude, and shut down (TDRS fleet 2017). The TDRS satellites track low earth orbit spacecraft and relay signals back to earth stations placed at strategic locations around the globe. The large antennas are 4.5 m diameter steerable antennas that can track LEO spacecraft; they are folded up for launch (NASA TDRS 2017). The 4.5 m antennas operate at S-band, Ku-band, and Ka-band, with a capability to handle bit rates up to 300 Mbps at Ku-band.

The omni antenna is part of the S-band telemetry and command system. Note that all the important components of the satellite are duplicated for redundancy. If one battery pack fails, for example, there are three left to run the satellite.

3.1.1 Attitude and Orbit Control System (AOCS)

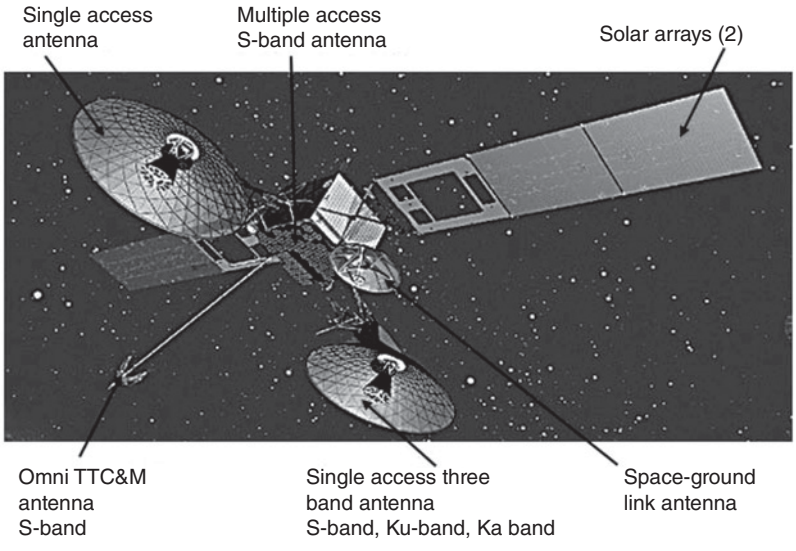
This subsystem consists of rocket motors and electric propulsion systems that are used to move the satellite back to the correct orbit when external forces cause it to drift off station, and gas jets or inertial devices that control the attitude of the satellite.

3.1.2 Telemetry, Tracking, Command, and Monitoring (TTC&M)

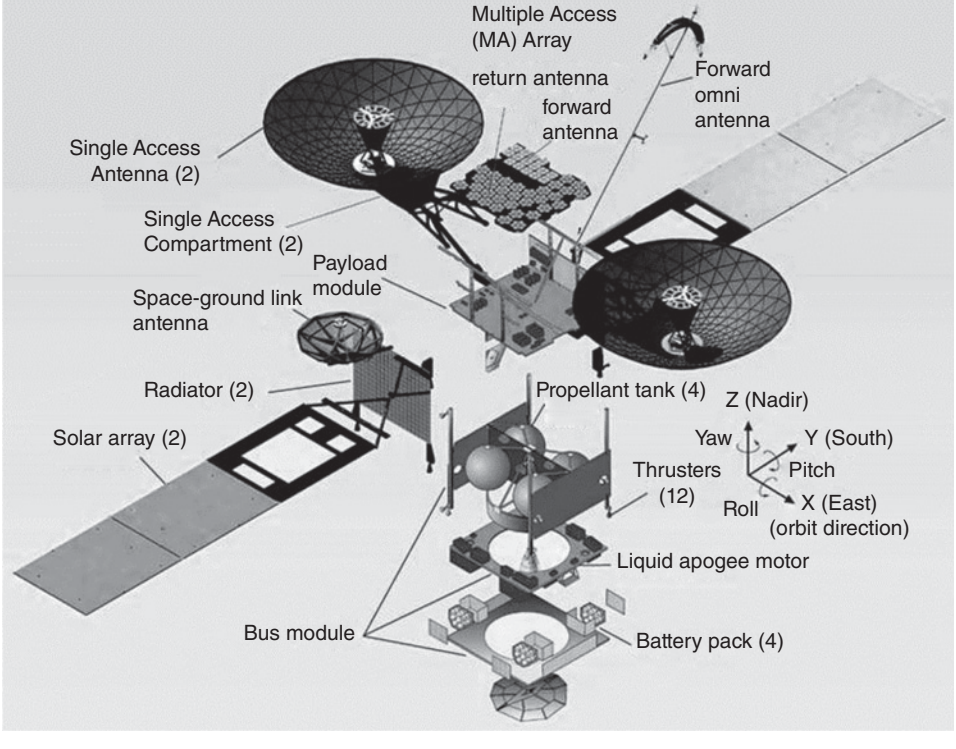
These systems are partly on the satellite and partly at the controlling earth station (Maral and Bousquet 2002 Chapter 10). The telemetry system sends data derived from many sensors on the satellite, which monitor the *satellite's health*, via a telemetry link to the controlling earth station. The tracking system is located at this earth station and provides information on the range and the elevation and azimuth angles of the satellite. Repeated measurement of these three parameters permits computation of orbital elements, from which changes in the orbit of the satellite can be detected. Based on telemetry data received from the satellite and orbital data obtained from the tracking system, the control system is used to correct the position and attitude of the satellite. It is also used to control the antenna pointing and communication system configuration to suit current traffic requirements, and to operate switches on the satellite.

3.1.3 Power System

All communications satellites derive their electrical power from *solar cells*. The power is used by the communication system, mainly in its transmitters, and also by all other electrical systems on the satellite. The latter use is termed *housekeeping,* since these subsystems serve to support the communications system. Power systems generate as little as 1 W of DC power for a 1U cubesat and up to 20 kW for a large GEO platform.



(a)



(b)

Figure 3.1 NASA third generation TDRS satellite. (a) TDRS satellite in orbit. (b) Exploded view of TDRS satellite. The satellite is inverted from Figure 3.1a. Source: NASA.

3.1.4 Communications Subsystems

The communications subsystem is the major component of a communications satellite, and the remainder of the satellite is there solely to support it. Frequently, the communications equipment is only a small part of the weight and volume of the whole satellite. It is usually composed of two or more antennas, which receive and transmit over wide bandwidths at microwave frequencies, and a set of receivers and transmitters that amplify and retransmit the incoming signals. The receiver-transmitter units are known as *transponders*. There are two types of transponder in use on satellites: the linear or *bent pipe* transponder that amplifies the received signal and retransmits it at a different frequency, and the *baseband processing transponder*, used only with digital signals, that converts the received signal to baseband, processes it, and then retransmits a digital signal. The latter approach is known generically as *onboard processing* (OBP).

3.1.5 Satellite Antennas

Although these form part of the communication system, they can be considered separately from the transponders. On large GEO satellites the antenna systems are very complex and produce beams with shapes carefully tailored to match the areas on the earth's surface served by the satellite, or multiple *spot beams* directed at specific points on earth. Most satellite antennas are designed to operate in a single frequency band, for example, C-band, Ku-band, or Ka-band, but some satellites have antennas for both C- and Ku-band, or Ku- and Ka-band, for example. A satellite that uses multiple frequency bands usually has four or more antennas. The communications capacity of a satellite can be increased by using spot beam antennas. The satellite can have one or more antennas creating many individual beams within the *footprint* of the satellite on the earth's surface, using two orthogonal polarizations and multiple frequency bands. LEO satellites for internet access have complex phased array antennas that can produce multiple electronically steered beams. Each beam is pointed to a *gateway* station or a user terminal and steered in angle as the satellite moves in its orbit.

The subsystems listed above are discussed in more detail in this chapter. There are other subsystems that are not discussed here, but which are essential to the operation of the satellite – the thermal control system that regulates the temperature inside a satellite, for example. The reader who is interested in spacecraft design should refer to the literature of that field, particularly the *IEEE Transactions on Aerospace and Electronic Systems* (IEEE Trans AP-S 1963–2018) and the *Journal of the American Institute of Aeronautics and Astronautics Transactions* and annual *Conference Proceedings* (AIAA Journal 2018). A useful reference text on spacecraft engineering is Fortescue et al. (2011). Only a brief review of the subsystems that support the communication mission is included here.

Satellites and space exploration have been popular subjects for postage stamps since the first satellites were launched in the late 1950s. For a selection of stamps go to Satellite Stamps (2012) and Space Exploration Stamps (2015).

3.2 Attitude and Orbit Control System (AOCS)

The attitude and orbit of a satellite must be controlled so that the satellite's antennas point toward earth and so that the user knows where in the sky to look for the satellite. This is particularly important for GEO satellites since the earth station antennas that are used with GEO satellites are normally fixed and movement of the satellite away from its appointed position in the sky will cause a loss of signal. There are several forces acting on an orbiting satellite that cause its attitude and orbit to change, as discussed in Chapter 2. The most important forces for a GEO satellite are the gravitational fields of the sun and the moon, and solar pressure from the sun.

A LEO satellite is less affected by the of the gravity of the sun and moon, but variations in the earth's magnetic field and gravitational constant cause deviations in the orbit. Solar pressure acting on a satellite's solar arrays and antennas, and the earth's magnetic field generating eddy currents in the satellite's metallic structure as it travels through the magnetic field, tend to cause rotation of the satellite body. Careful design of the structure can minimize these effects, but the orbital period of the satellite makes many of the effects cyclic, which can cause *nutation* (a wobble) of the satellite. The attitude control system must damp out nutation and counter any rotational torque or movement.

The presence of gravitational fields of the sun and the moon cause the orbit of a GEO satellite to change with time. At GEO orbit altitude, the sun's gravitational force is about three times as strong as the sun's. The moon's orbit is inclined to the earth's equatorial plane by approximately 5°, which creates a force on the satellite with a component that is normal to the satellite's orbit. The plane of the earth's rotation around the sun (the *ecliptic*) is inclined by 23° to the earth's equatorial plane. As discussed in Chapter 2, there is a net gravitational pull on a GEO satellite that tends to change the inclination of the satellite's orbit, pulling it away from the earth's equatorial plane at an initial rate of approximately 0.86° per year. The orbital control system of the satellite must be able to move the satellite back into the equatorial plane before the orbital inclination becomes excessive. LEO satellites are less affected by gravitational fields of the sun and moon. Since they are much closer to the earth than GEO satellites, earth's gravity is much stronger, and the pull from the sun and moon are proportionately weaker. The gravitational force on an orbiting satellite decreases as the square of the distance from the center of the earth. At an orbital altitude of 1000 km, the reduction in earth's gravity is approximately 35%. At geostationary altitude of 35 786 km, the reduction is 97.7%, or a factor of 43.7. The small earth gravitational force at GEO altitude allows the gravity of the sun and moon to influence the orbit of a GEO satellite much more than with a LEO satellite.

The earth is not quite a perfect sphere. At the equator, there are bulges of about 65 m at longitudes 162°E and 348°E, with the result that a GEO satellite is accelerated toward one of two stable points in the GEO orbit at longitude 75°E and 252°E, as shown in Figure 3.2. To maintain accurate station keeping, the satellite must be accelerated periodically in the opposite direction to the forces acting on it. This is done as a sequence of *station keeping maneuvers,* using small rocket motors called *gas jets* or *thrusters* that can be controlled from earth via the TTC&M system.

Low earth orbit satellites cross the sky in a few minutes and must be tracked by earth station antennas. The orbital elements of each satellite must be determined by the operator of the LEO constellation and provided to earth station users so that the look

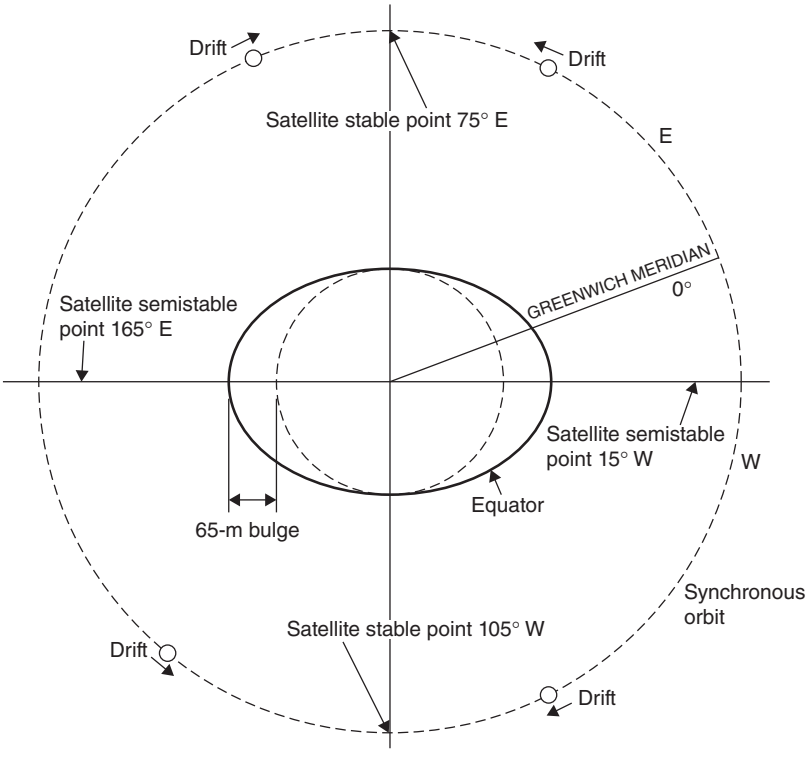


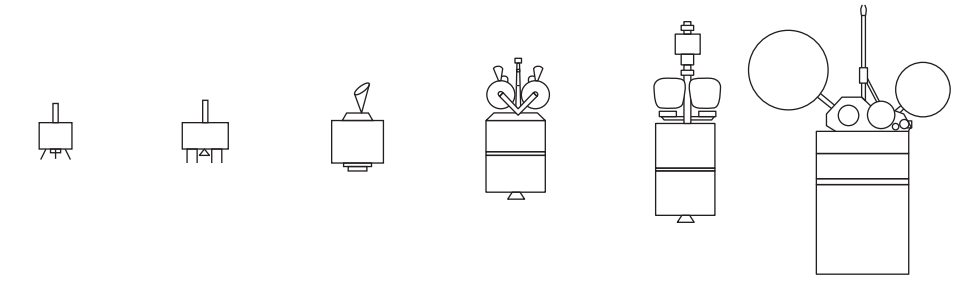
Figure 3.2 Forces on a satellite in geosynchronous orbit. GEO satellites tend to drift around their orbit toward the stable points.

angles of the satellite can be calculated second by second and supplied to the tracking antenna.

3.2.1 Attitude Control System

There are several ways to make a satellite stable in orbit, where it is weightless. Prior to year 1990, the body of most GEO satellites was rotated at a rate between 30 and 100 rpm to create a gyroscopic force that provided stability on the spin axis and kept the satellite pointing in the same direction. Such satellites are known as *spinners*. The Boeing 376 GEO satellite was a spinner design that enjoyed widespread use, but had limited power generation capability because only one third of the solar cells on a spinner satellite body are illuminated by the sun (Boeing 376, n.d.). All large GEO communication satellites are now three-axis stabilized designs and most cubesats have three-axis stabilization.

A *three-axis stabilized satellite* has one or more *momentum wheels*. The momentum wheel is usually a solid metal disk driven by an electric motor. Either there must be one momentum wheel for each of the three axes of the satellite, or a single momentum wheel can be mounted on gimbals and rotated to provide a rotational force about any of the three axes (Reaction Wheel 2018). Increasing the speed of the momentum wheel causes the satellite to precess in the opposite direction, according to the principle of conservation of angular momentum. Figure 3.3a shows some examples of spinner satellites used by Intelsat for international telephone and data transmission.



INTELSAT I | INTELSAT II | INTELSAT III | INTELSAT IV | INTELSAT IV-A | INTELSAT VI

Figure 3.3a Spinner satellites launched or operated by Intelsat between 1965 and 2013. Intelsat V is missing as it was a three-axis stabilized satellite. Subsequent satellites were all three-axis stabilized. Intelsat 1 had an on-orbit mass of 34 kg and generated 46 W of power at end of life. Total effective bandwidth was 50 MHz. Intelsat 6 had an on-orbit mass of 1800 kg and generated 2100 W of power at end of life. Total effective bandwidth was 3360 MHz.

As shown in Figure 3.3b, the body of a spinner satellite consists of a cylindrical drum covered in solar cells with the power systems, fuel storage, and batteries inside. The communications system is mounted at the top of the drum and is driven by an electric motor in the opposite direction to the rotation of the satellite body to keep the antennas pointing toward earth. The satellite is *spun up* by operating small thrusters mounted on the periphery of the drum, at an appropriate point in the launch phase. The despin system is then brought into operation so that the main TTC&M antennas point toward earth. One advantage that spinner satellites had was simpler thermal control. One half of the satellite was in sunlight causing the solar cells to heat up, while the other half faced deep space, allowing heat to be radiated away from the solar cells.

This kept the satellite at a constant temperature. By comparison, one side of a three-axis stabilized satellite is constantly baked by the sun while the opposite side is constantly

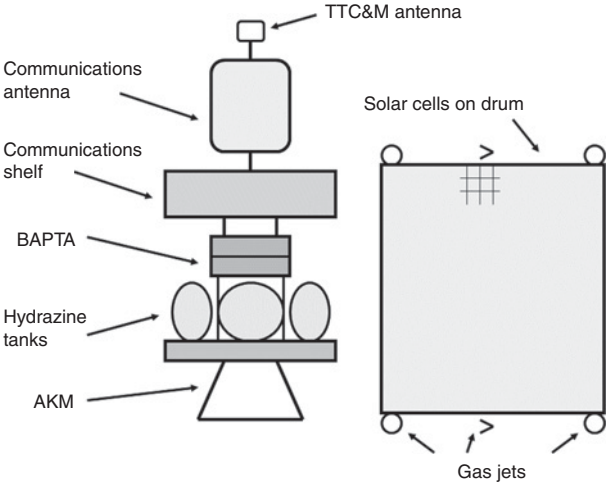


Figure 3.3b A typical spinner satellite from the 1980s. The entire satellite rotated at roughly one revolution per second. The communications equipment and antennas were driven by a motor in the opposite direction to the body of the satellite to keep the antennas pointed at earth. BAPTA, Bearing and power transfer assembly; AKM, Apogee kick motor; TTC&M, Telemetry, tracking, command, and monitoring.

losing heat. Some satellites employ *heat pipes,* tubes filled with a liquid metal that can convey heat from the hot side of the satellite to the cold side.

The Intelsat 6 series of satellites was the largest of the spinner design. Later satellites were all three-axis stabilized. Intelsat 603 was launched on 14 March 1990, but the satellite failed to separate from the second stage of the launcher. This left the satellite and the upper stage attached to each other in low earth orbit, preventing firing of the apogee kick motor (AKM) that would place the satellite in a GEO transfer orbit. Intelsat arranged for the satellite to be rescued by the US Space Shuttle in May 1992. Three of the shuttle astronauts successfully captured the satellite by hand after several attempts using a capture bar. A new AKM was installed by the Space Shuttle crew and the satellite was released from the shuttle. The AKM was successfully ignited and the satellite placed into GEO transfer orbit (Intelsat 603 2016). Intelsat 603 was finally deorbited in January 2013. Intelsat 603 was the only satellite ever to be rescued in orbit. It is usual for the launch of a large GEO satellite to be insured at a premium between 18% and 23% of the launch cost; if the satellite fails to reach its intended orbit, the owner makes a claim against the insurer and the stranded satellite becomes another piece of space junk. The insurance also covers problems found during the checkout phase once the satellite has reached its final orbit, typically the first 100 days.

Figure 3.3c shows some examples of several large geostationary three-axis stabilized satellites. The satellites all have large solar arrays that fold down against the satellite's body for launch, and multiple reflector antennas that also fold in for launch. Typical span across the solar arrays is 30 m once in orbit, and power generated can be up to 20 kW.

A variety of liquid propulsion mixes have been used for the thrusters, the most common being a variant of *hydrazine* (N_2H_4), which is easily liquefied under pressure, but readily decomposes when passed over a catalyst (Hydrazine 1984). Increased power can be obtained from the hydrazine gas jets by electrically heating the catalyst and the gas. Satellites that use liquid fuel thrusters have standardized on bi-propellant fuels, that is, fuels that mix together to form the thruster fuel. The most common bi-propellants used for thruster operations are mono-methyl hydrazine and nitrogen tetroxide. Bi-propellants are hypogolic and ignite spontaneously on contact, so do not need either a catalyst or a heater. The fuel that is stored on a GEO satellite is used for two purposes: to fire the AKM that injects the satellite into its final orbit, and to maintain the satellite in that orbit over its lifetime. If the launch is highly accurate, a minimum amount of fuel is used to attain the final orbit. If the launch is less accurate, more fuel must be used up in maneuvering the satellite into position, and that reduces the amount left for station keeping.

There are two types of rocket motors used on satellites. The traditional bi-propellant thruster described above, and *arc-jets* or *ion thrusters*. Ion thrusters use a high voltage source to accelerate ions to a very high velocity, thus producing thrust. The ion engine thrust is not large, but because the thruster can be driven by power from the solar cells it saves on expendable fuel. Ion engines can also be used to slowly raise a GEO satellite from a transfer orbit to GEO orbit as described in Chapter 2, although the process takes much longer than with a rocket engine.

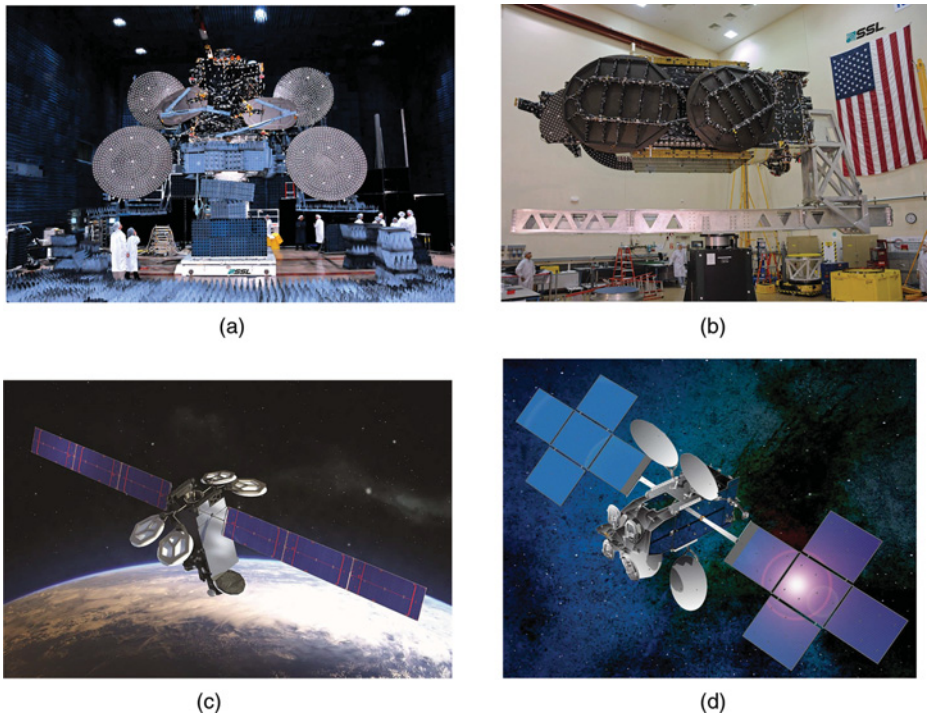


Figure 3.3c Examples of three-axis stabilized communication satellites. (a) A large GEO direct broadcast television satellite built by SSL, under test. (b) Same satellite as (a) folded for launch. The solar cells are folded onto the top and bottom of the body and the antennas are folded against the sides of the body, as viewed in this photograph. (c) Intelsat 35e satellite. (d) ViaSat 1 satellite. Source: Image credits: (a) and (b) Courtesy of SSL, © SSL 2018; (c) © Intelsat, S.A. 2018 and its affiliates. All rights reserved; (d) Courtesy of ViaSat, © ViaSat 2018. For a color version of this figure please see color plate section.

The most common propellant used in ion propulsion is xenon, which is easily ionized and has a high atomic mass. It also is inert and can be stored as a liquid under high pressure. Compared to chemical fuels used in rocket motors, a much smaller weight of xenon is required for station keeping over the lifetime of a GEO satellite. A typical Hall effect ion engine has a hollow cylindrical chamber charged to a high positive voltage and a discharge cathode at one end. Electrons are attracted toward the walls of the chamber, but are directed into a stream toward the open end of the chamber by a powerful magnetic field. The propellant is introduced into the chamber as a gas, where electrons bombard the propellant to produce positively charged ions and release more electrons. The ions are accelerated toward a negatively charged grid and emitted from the ion engine at a high velocity, up to 145 km/s. Compared to a chemical rocket engine, the ion thruster emits a much lower volume of particles but at a much higher velocity. Typical thrust is 0.5 Newtons (Ion Thruster 2004, 2008, 2018).

In a three-axis stabilized satellite, one pair of gas jets or ion thrusters is needed for each axis to provide for rotation in both directions of pitch, roll, and yaw. An additional set of controls, allowing only one thruster on a given axis to be operated, provides for velocity increments in the *X, Y,* and *Z* directions. When motion is required along a given axis,

the appropriate thruster is operated for a specified period of time to achieve the desired velocity. An opposing thruster must be operated for the same length of time to stop the motion when the satellite reaches its new position. Fuel is saved if the velocity of the satellite is kept small, but progress toward the destination is slow. Since fuel on board an orbiting satellite is a finite resource, slow movements are generally preferred even if this results in loss of revenue.

We can define a set of reference Cartesian axes (X_R, Y_R, Z_R) with the satellite at the origin, as shown in Figure 3.4a. The Z_R axis is directed toward the center of the earth and is in the plane of the satellite orbit. It is aligned along the local vertical at the satellite's subsatellite point. The X_R axis is tangent to the orbital plane and lies in the orbital plane. The Y_R axis is perpendicular to the orbital plane. For a satellite serving the Northern Hemisphere, the directions of the X_R and Y_R axes are nominally east and south.

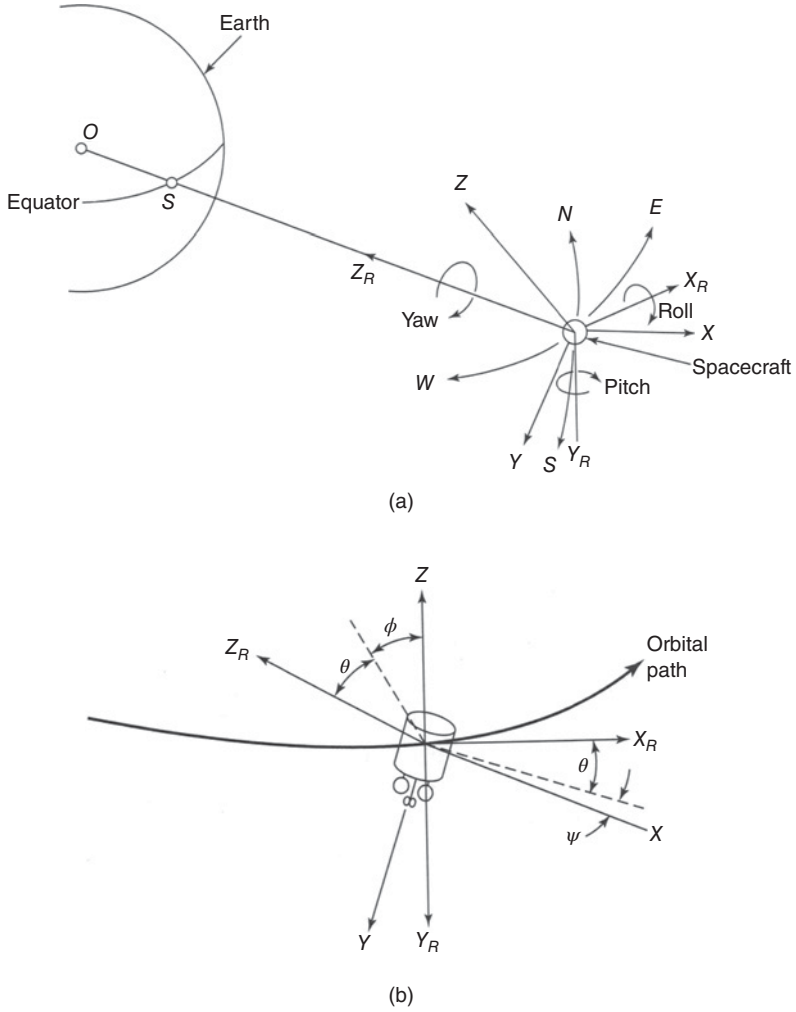


Figure 3.4 (a) Definition of pitch, roll, and yaw for a geostationary satellite. (b) Relationship between axes of a GEO satellite when in orbit. The axes X_R, Y_R, and Z_R are related to the orbit. Axes *X, Y, Z* relate to the satellite. The *Z* axis of the satellite is directed to a specific point on earth.

Rotation about the X_R, Y_R, and Z_R axes is defined as *roll* about the X_R axis, *pitch* about the Y_R axis, and *yaw* about the Z_R axis, in exactly the same way as for an aircraft or ship traveling in the X direction. The satellite must be stabilized with respect to the reference axes to maintain accurate pointing of its antenna beams. The axes X_R, Y_R, and Z_R are defined with respect to the location of the satellite; a second set of Cartesian axes, *X, Y, Z*, as shown in Figure 3.4b, define the orientation of the satellite. Changes in a satellite's attitude cause the angles θ, φ, and ψ in Figure 3.4b to vary as the *X, Y, Z* axes move relative to the fixed reference axes X_R, Y_R, and Z_R. The Z axis is usually directed toward a reference point on earth, called the *Z-axis intercept*. The location of the Z-axis intercept defines the pointing of the satellite antennas; the Z-axis intercept point may be moved to repoint all the antenna beams by changing the attitude of the satellite with the attitude control system.

Attitude control of a three-axis stabilized satellite requires an increase or a decrease in the speed of a momentum wheel. If a constant torque exists about one axis of the satellite, a continual increase or decrease in momentum wheel speed is necessary to maintain the correct attitude. When the upper or lower speed limit of the wheel is reached, it must be *unloaded* by operating a pair of thrusters and simultaneously reducing or increasing the wheel speed in the appropriate sense. Closed-loop control of attitude is employed on the satellite to maintain the correct attitude. When the satellite has narrow beam antennas, the whole satellite may have to be stabilized within $\pm 0.1^\circ$ on each axis. The references for the attitude control system may be the outer edge of the earth's disk, as observed with infrared sensors, the sun, or one or more stars. The control system for a three-axis stabilized satellite employs an onboard computer to process the sensor data and command the thrusters and momentum wheels (Maral and Bousquet 2002). Figure 3.5 is a simplified diagram of the attitude control system for a three-axis stabilized antenna.

The earth sensor illustrated in Figure 3.5 consists of four small telescopes aimed at four points along the edge of the earth as seen from geostationary orbit. The view seen by each telescope is half of the earth and half of dark space beyond the earth. Sensors for 14–16 μm wavelength infrared radiation at the focus of each telescope have identical outputs when the earth is symmetrically aligned with all four sensor views. Long wavelength IR is used to avoid the influence of clouds. If the left hand pair of telescopes has a higher output than the right hand pair, for example, the view of the earth has moved to the left and the attitude controller will adjust the satellite pointing accordingly. The sun sensors work in the same fashion. Star sensors employ a telescope with an infrared detector and must track the selected star as the satellite moves in its orbit. Occasionally, the moon will rise above the rim of the earth and be seen by an infrared earth sensor, a condition known as a *moon hit*. The moon has a noise temperature of 200 K and can confuse the earth sensor system. Fortunately, the moon has a highly predictable orbit and moon hits can be accurately predicted in advance.

3.2.2 Orbit Control System

As discussed in Chapter 2, a geostationary satellite is subjected to several forces that tend to accelerate it away from its required orbit. The most important, for the geostationary satellite, are the gravitation forces of the moon and the sun, which cause inclination of the orbital plane, and the non-spherical shape of the earth around the equator, which causes drift of the subsatellite point. There are many other smaller forces that act on the satellite causing the orbit to change over time. Accurate prediction of the satellite

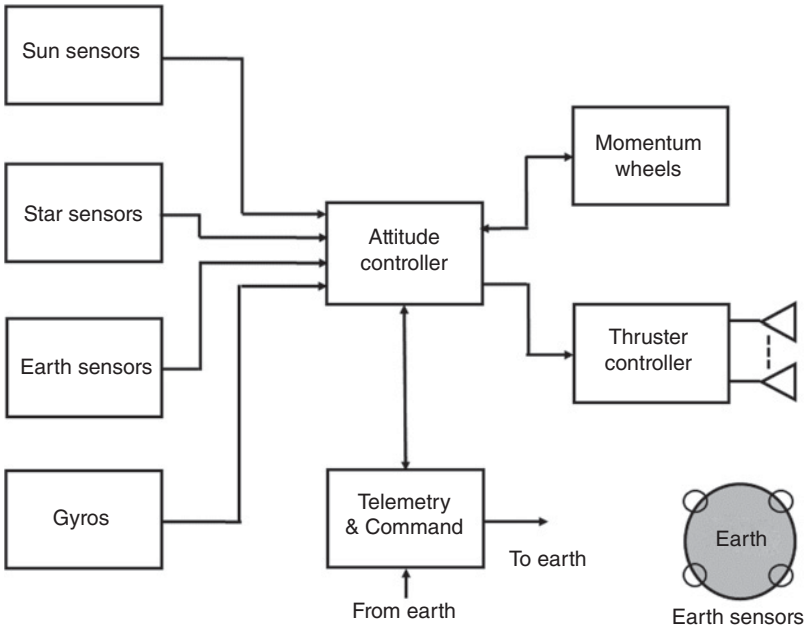


Figure 3.5 Attitude control system for a three-axis stabilized GEO satellite. The sun and star sensors are used to establish orientation with respect to space and the earth sensors are used for z-axis pointing. The gyros detect movement on the three axes of the satellite and send signals to the controller to initiate a correction with the momentum wheels. The thrusters used for attitude control are ion thrusters. The attitude controller operates autonomously and sends all sensor data and command data to the controlling earth station over the telemetry link.

position a week or two weeks ahead requires a computer program with up to 20 force parameters; we shall restrict our discussion here to the two major effects.

Figure 3.6 shows a diagram of an inclined orbital plane close to the geostationary orbit.

For the orbit to be truly geostationary, it must lie in the equatorial plane, be circular, and have the correct altitude. The various forces acting on the satellite will steadily pull it out of the correct orbit; it is the function of the orbit control system to return it to the correct orbit. This cannot be done with momentum wheels since linear accelerations are required. Thrusters that can impart velocity changes along the three references axes of the satellite are required.

If the orbit of a GEO satellite is not circular, a velocity increase or decrease must be made along the orbit, in the X-axis direction in Figure 3.4. On a three-axis stabilized satellite, there will usually be two pairs of X-axis thrusters acting in opposite directions, one pair of which will be operated for a predetermined length of time to provide the required velocity change. The orbit of a geostationary satellite remains approximately circular for long periods of time and does not need frequent velocity corrections to maintain circularity. Altitude corrections are made by operating the Z-axis thrusters.

The inclination of the orbit of a satellite that starts out in a geostationary orbit increases at an average rate of about 0.85° per year, with an initial rate of change of inclination for a satellite in an equatorial orbit between 0.75° and 0.94° per year (see Chapter 2). Most GEO satellites are specified to remain within a box of ±0.05° as seen from earth in azimuth and elevation, and so, in practice, corrections called a

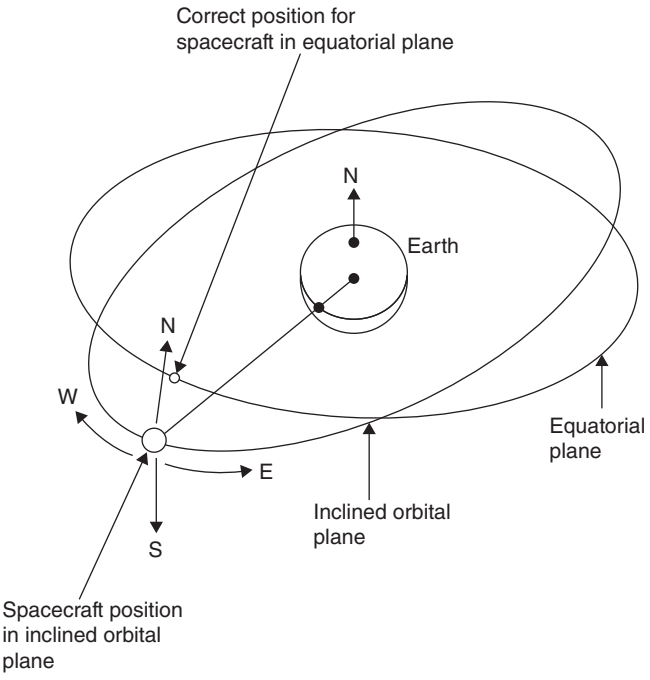


Figure 3.6 Geosynchronous satellite in an inclined orbit. Thrusters must be used to give the satellite a velocity in a northerly direction until it reaches the required position. Opposing thrusters must then be fired to stop the satellite's motion.

North–South station keeping maneuver are made every two to four weeks to keep the error small. It has become normal to split the E–W and N–S maneuvers so that at intervals of two weeks the E–W corrections are made first and then after two more weeks, the N–S corrections are made. When ion thrusters are used for N–S station keeping maneuvers, they tend to operate almost continuously since their thrust levels are low when compared with liquid fueled engines. Correcting the inclination of a satellite orbit requires more fuel to be expended than for any other orbital correction. This places a weight penalty on those satellites that must maintain very accurate station keeping, and reduces the communications payload they can carry. As much as half the total satellite weight at launch may be station keeping fuel when liquid fuel thrusters are employed and the satellite's expected lifetime on orbit is 15 years. Figure 3.7 shows a simplified diagram of the orbital control system of a three-axis stabilized GEO satellite.

GEO satellites may relax their inclination (N–S station keeping) to become *inclined orbit satellites* but may never relax the E–W station keeping tolerance as this would lead to unacceptable interference into other systems. Some systems have used several satellites in the same inclined GEO orbit to provide coverage to extreme north and south latitudes as a GEO satellite in an equatorial orbit cannot be seen from latitudes above 75°. LEO communications satellites do not need the very tight station keeping tolerance of GEO satellites, as they are less affected by the gravitational pull of the moon and sun because of the stronger gravitational force of the earth, so they do not need to carry as large a supply of fuel as a GEO satellite.

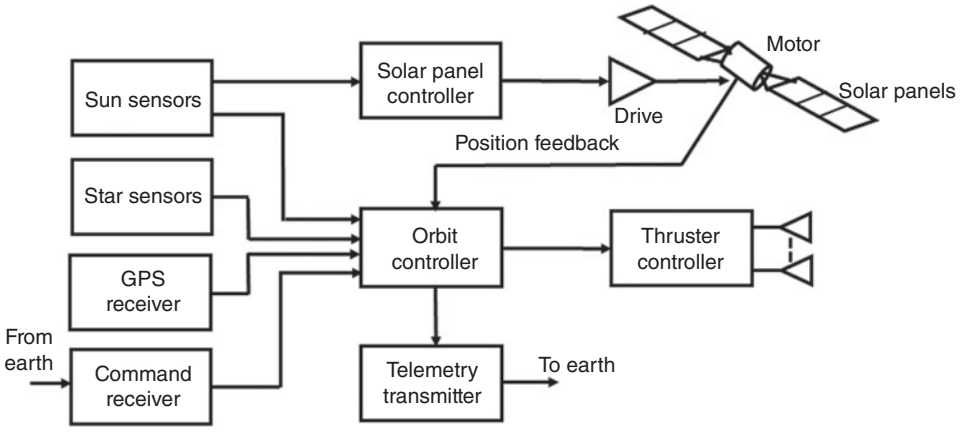


Figure 3.7 Simplified orbital control system. The sun and star sensors and the GPS receiver provide information on the satellite's orbital location. The orbital controller can command E–W maneuvers to keep the satellite on station using ion thrusters. N–S maneuvers can require firing of the gas jets and are performed by command from the controlling earth station. Also shown is the drive system that rotates the solar panels to face the sun, using information from the sun sensors. There are multiple sensors of each type to provide redundancy.

East–West station keeping is effected by use of the *X*-axis thrusters of the satellite. For a satellite located away from the stable points at 75°E and 252°E, a slow drift toward these points will occur. Typically, the *X*-axis jets are pulsed every two or three weeks to counter the drift and add a small velocity increment in the opposite direction. The satellite then drifts through its nominal position, stops at a point a fraction of a degree beyond it, and then drifts back again. East–West station keeping requires only a modest amount of fuel and is necessary on all geostationary communications satellites to maintain the spacing between adjacent satellites. With orbital locations separated by two or three degrees, East–West drifts in excess of a fraction of a degree cannot be tolerated, and most GEO satellites are held within $\pm 0.05°$ of their allotted longitude.

Low earth orbit (LEO) and medium earth orbit (MEO) satellites also need AOC systems to maintain the correct orbit and attitude for continuous communication. Because of the much stronger gravitational force of the earth in LEO orbit, attitude stabilization is often accomplished with a rigid *gravity gradient boom*. This is a long pole that points toward the center of the earth, providing damping of oscillations about the satellite's *z*-axis by virtue of the difference in gravitational field at the top of the pole and at the bottom.

3.3 Telemetry, Tracking, Command, and Monitoring (TTC&M)

The TTC&M system is essential to the successful operation of a communications satellite. It is part of the satellite management task, which also involves an earth station, usually dedicated to that task, and a group of personnel. The main functions of satellite management are to control the orbit and attitude of the satellite, monitor the status of all sensors and subsystems on the satellite, and switch on or off sections of the communication system. The TTC&M earth station may be owned and operated by the satellite

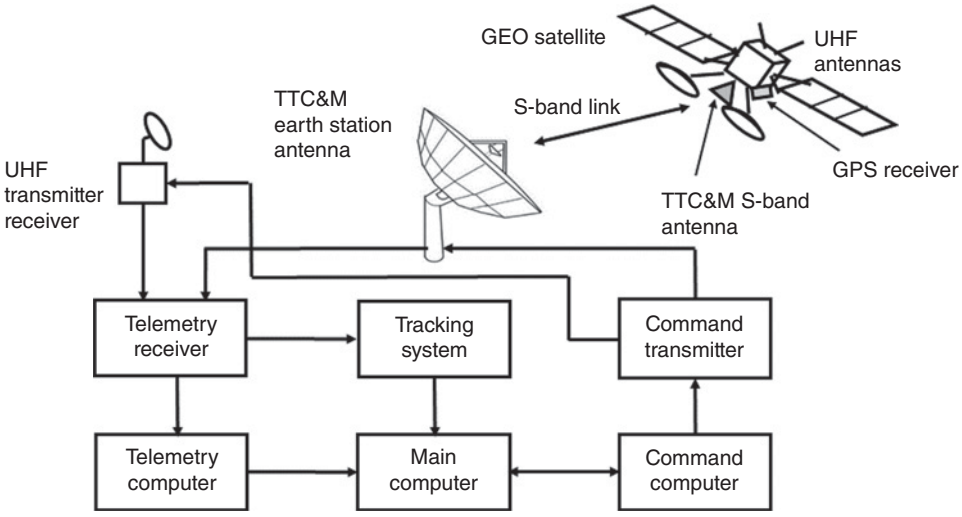


Figure 3.8 Simplified diagram of the earth based control system for a GEO satellite. The main computer receives telemetry data from the satellite's attitude and orbital control systems, as well as all the many sensors on board the satellite. Commands for station keeping maneuvers and changes to switch settings originate in the main computer and are translated to command codes by the command computer. The UHF command and telemetry system is used during the launch phase when the satellite orientation has not been stabilized. Once on station, TTC&M communications are switched to the main frequency, S-band in this example. All telemetry data is stored for later analysis to determine aging trends in critical components such as transponders.

owner, or it may be owned by a third party that provides TTC&M services under contract. On large geostationary satellites, some repointing of individual antennas may be possible, under the command of the TTC&M system. Tracking is performed primarily by the earth station. Figure 3.8 illustrates the functions of a controlling earth station.

3.3.1 Telemetry and Monitoring System

The monitoring system collects data from many sensors within the satellite and sends the data to the controlling earth station. There may be several hundred sensors located on the satellite to monitor pressure in the fuel tanks, voltage, and current in the power conditioning unit, current drawn by each subsystem, and critical voltages and currents in the communications electronics. The temperature of many of the subsystems is important and must be kept within predetermined limits, so many temperature sensors are fitted. The sensor data, the status of each subsystem, and the positions of switches in the communication system are reported back to earth by the telemetry system. The sighting devices used to maintain attitude are also monitored via the telemetry link: this is essential in case one should fail and cause the satellite to point in the wrong direction. The faulty unit must then be disconnected and a spare brought in, via the command system, or some other means of controlling attitude devised.

Telemetry data are usually digitized and transmitted as phase shift keying (PSK) of a low-power telemetry carrier using time division multiplexing (TDM). A low data rate is normally used to allow the receiver at the earth station to have a narrow bandwidth and

thus maintain a high carrier to noise ratio. The entire TDM frame may contain thousands of bits of data and take several seconds to transmit. At the controlling earth station a computer is used to monitor, store, and decode the telemetry data so that the status of any system or sensor on the satellite can be determined immediately by the controller on earth. Alarms can also be sounded if any vital parameter goes outside allowable limits.

3.3.2 Tracking

A number of techniques can be used to determine the current orbit of a satellite. Velocity and acceleration sensors on the satellite can be used to establish the change in orbit from the last known position, by integration of the data. The earth station controlling the satellite can observe the Doppler shift of the telemetry carrier or beacon transmitter carrier to determine the rate at which range is changing. Together with accurate angular measurements from the earth station antenna, range is used to determine the orbital elements. Active determination of range can be achieved by transmitting a pulse, or sequence of pulses, to the satellite and observing the time delay before the pulse is received again. The propagation delay in the satellite transponder must be accurately known, and more than one earth station may make range measurements. If a sufficient number of earth stations with an adequate separation are observing the satellite, its position can be established by triangulation from the earth station by simultaneous range measurements. With precision equipment at the earth stations, the position of the satellite can be determined within 10 m.

Ranging tones are also used for range measurement. A carrier generated on board the satellite is modulated with a series of sine waves at increasing frequency, usually harmonically related. The phase of the sine wave modulation components is compared at an earth station, and the number of wavelengths of each frequency is calculated. Ambiguities in the numbers are resolved by reference to lower frequencies, and prior knowledge of the approximate range of the satellite. If sufficiently high frequencies are used, perhaps even the carrier frequency, range can be measured to millimeter accuracy. The technique is similar to that used in the terrestrial *tellurometer*, in aircraft radar altimeters, and in high accuracy position location using global positioning system (GPS) satellites where the repetition of code sequences provides ranging information. Accurate knowledge of the range to the satellite is essential for multiple access systems using time division. Bursts of data from each earth station must arrive at the satellite in the correct sequence with only a few microseconds between the arriving bursts.

Some satellites carry GPS receivers that report the satellite's position over the telemetry link. GEO satellites orbit at a higher altitude than GPS satellites requiring a somewhat different operating mode to calculate position. The availability of the L5 signal on later GPS satellites makes possible dual frequency measurements with accuracy better than 5 m (GEO satellite positioning with GPS 2005).

3.3.3 Command

A secure and effective command structure is vital to the successful launch and operation of any communications satellite. The command system is used to make changes in attitude and corrections to the orbit and to control the communication system. During launch, it is used to control the firing of the AKM and to extend the solar arrays and antennas of a three-axis stabilized satellite.

The command structure must possess safeguards against unauthorized attempts to make changes to the satellite's operation, and also against inadvertent operation of a control due to error in a received command. Encryption of commands and responses is used to provide security in the command system. Intelsat did not encrypt its command and control systems for the first five series of satellites, believing that no one with a 30 m antenna would attempt to take over a satellite, although this would have been easy to do. With the Intelsat VI series of satellites, the US government insisted that encryption be employed, and all subsequent Intelsat satellites used encryption on their TTC&M links. The *control code* is converted into a *command word,* which is sent in a TDM frame to the satellite. After checking for validity in the satellite, the word is sent back to the control station via the telemetry link where it is checked again in the computer. If it is received correctly, an *execute* instruction is sent to the satellite so that the command is executed. The entire process may take 5 or 10 seconds, but minimizes the risk of erroneous commands causing a satellite malfunction.

The command and telemetry links are usually separate from the communication system on a GEO satellite, although they may operate in the same frequency band, often C-band (6 and 4 GHz). Two levels of command system are used in some satellites; the main system operates in the 6 GHz band, in a gap between the communication channel frequencies; the main telemetry system uses a similar gap in the 4 GHz band. The TTC&M antenna for the S-band system on the TDRS satellites can be seen in Figure 3.1.

During the launch phase and injection into geostationary orbit, the main TTC&M system may be inoperable because the satellite does not have the correct attitude or has not extended its solar arrays. A backup system is used at this time, which controls only the most important sections of the satellite. A great deal of redundancy is built into this system, since its failure will jeopardize the entire mission. Near omnidirectional antennas are used at either ultra high frequency (UHF) or S-band (2–4 GHz), and sufficient margin is allowed in the signal to noise ratio (SNR) at the satellite receiver to guarantee control under the most adverse conditions. The backup system provides control of the AKM, the attitude control system and orbit control thrusters, the solar sail deployment mechanism (if fitted), and the power conditioning unit. With these controls, the satellite can be injected into geostationary orbit, turned to face earth, and switched to full electrical power so that handover to the main TTC&M system is possible. When ion thrusters are used to raise the satellite to geosynchronous orbit, the antennas and solar arrays can be deployed during the orbit raising process, allowing full operation of the satellite and main TTC&M system. This allows all the satellite's systems to be checked out by the time the satellite reaches GEO. In the event of failure of the main TTC&M system, the backup system can be used to keep the satellite on station. It is also used to eject the satellite from geostationary orbit and to switch off all transmitters when the satellite eventually reaches the end of its useful life.

Controlling a satellite in orbit is a complex process that requires considerable care. In one case, an incorrect sequence of command instructions caused loss of control of the Olympus satellite, a European large GEO satellite used for experiments in the 30/20 GHz band. The E–W thrusters fired for a lengthy period, causing the satellite to drift toward the east at 5° per day, and the satellite also began rotating with a period of 90 seconds. All communication with the satellite was lost, and the batteries discharged fully because the solar arrays no longer pointed at the sun. The satellite drifted round the earth over

a period of 2½ months, and was eventually recovered by a team of experts using large antennas in Australia and the United States to send telemetry commands to the satellite. The solar cells provided short bursts of power as they rotated past the sun's direction, allowing commands to be sent for a few seconds every 90 seconds. The rotation of the satellite was eventually stopped and Olympus returned to its correct location, but with shortened life expectancy due to the loss of station keeping fuel.

3.4 Power Systems

All communications satellites obtain their electrical power from solar cells, which convert incident sunlight into electrical energy. Some deep space planetary research satellites have used thermonuclear generators to supply electrical power, but because of the danger to people on the earth if the launch should fail and the nuclear fuel spread over an inhabited area, communications satellites have not used nuclear generators. Thermonuclear generators are needed on deep space probes that go to the most distant planets because the strength of sunlight varies inversely with the square of the distance of a spacecraft from the sun. Solar panels are not effective when a spacecraft is close to Jupiter, for example, at a distance of 1440 million kilometers from the sun, where the intensity of sunlight is 1% of that in an earth orbit.

3.4.1 Solar Power Systems

The sun is a powerful source of energy. In the total vacuum of outer space, at geostationary altitude, the radiation falling on a satellite has an intensity of 1.36 kW/m^2. Solar cells do not convert all this incident energy into electrical power; the efficiency for gallium arsenide (GaAs) cells is typically 33–39% at *beginning of life* (BOL) but falls with time because of aging of the cells and etching of their surfaces by micrometeor impacts. The silicon cells of a typical home solar power installation have efficiencies between 10% and 19%. Since sufficient power must be available at the *end of life* (EOL) of the satellite to supply all the systems on board, about 15% extra area of solar cells is usually provided as an allowance for aging (Solar Panels 2018).

A three-axis stabilized satellite has solar cells arranged on flat panels along the Y-axis of the satellite (see Figure 3.4b) that are rotated by an electric motor to maintain normal incidence of the sunlight. This causes the cells to heat up, typically to 50–80° C, which causes a drop in output voltage. A rotary joint with slip rings must be used with each solar sail to transfer current from the rotating sail to the body of the satellite. Large GEO satellites have solar arrays that can generate up to 20 kW of electric power; with a bus voltage of 50 V and two solar arrays generating 10 kW, each slip ring must carry a current of 200 A. Slip ring failures have occurred on some GEO satellites, cutting the available power in half.

The most powerful solar power system in space is located on the International Space Station, where solar panels were added over a period of years to generate a maximum of 120 kW. Because the Space Station is in low earth orbit it is in sunlight for only half of each orbit; 60% of the output of the solar arrays is used to charge the station's batteries so that normal operation can continue while it is in darkness (Space Station Solar Arrays 2017).

3.4.2 Batteries

Satellites must carry batteries to power the subsystems during launch and eclipses. Eclipses occur twice per year, around the spring and fall equinoxes, when the earth's shadow passes across the satellite, as illustrated in Figures 2.20 and 2.21. The longest duration of eclipse is 70 minutes, occurring around 20 March and 22 or 23 September of each year. By locating the satellite 20°W of the longitude of the service area, the eclipse will occur after 1 a.m. local time for the service area, when shutdown is more acceptable. Batteries of the nickel-hydrogen type, which do not gas when charging, with high reliability and long life were the choice for satellites until lithium-ion batteries became available. The higher capacity per unit weight of lithium-ion batteries has resulted in their widespread use on satellites (Saft 2017).

A power-conditioning unit controls the charging current and dumps excess current from the solar cells into heaters or load resistors on the cold side of the satellite. Sensors on the batteries, power regulator, and solar cells monitor temperature, voltage, and current, and supply data to both the onboard control system and the controlling earth station via the telemetry downlink. Typical battery voltages are 20–50 V with capacities of 20–1000 ampere hours.

Example 3.1

A large GEO satellite requires a total of 12 kW to operate its communication systems and 1.5 kW for housekeeping purposes. The solar cells on the satellite are mounted on two large sails that rotate to face the sun at all times. The efficiency of the solar cells is 36% at BOL and 33% at EOL. Using an average incident solar flux density of 1.36 kW/m^2 calculate the area of each solar sail to meet the power requirements at the end of the satellite's life.

How much power is generated at BOL?

The solar arrays are 2.0 m wide. How long are they?

Answer

The total power required by the satellite is 13.5 kW. At EOL the solar cells' efficiency is 33%. With an incident solar flux of 1.36 kW/m^2, the total area of solar sail required is

$$A = \frac{13.5}{0.33 \times 1.36} = 30.1\ m^2$$

At BOL, the 30.1 m^2 of solar cells will generate a power P kW where

$$P = 30.1 \times 0.36 \times 1.36 = 14.74\ \text{kW}$$

Each solar array must have an area of 15.05 m^2 and will have a length of 7.53 m.

Example 3.2

The large GEO satellite in Example 3.1 is subject to eclipses that last 70 minutes in spring and fall. The satellite is required to maintain full communications capacity during eclipses. Batteries on board the satellite must supply 13.5 kW for 70 minutes. The battery voltage is 50 V and the batteries must not discharge more than 50% during the eclipse.

Calculate the battery capacity required in ampere hours (AHs). A battery with a capacity of one ampere hour can supply one amp for one hour.

If lithium-ion batteries with a capacity of 200 watt hours per kilogram are used, find the weight of the battery.

Answer
First calculate the current required to supply 13.5 kW at 50 V.

$$I = \frac{P}{V} = \frac{13{,}500}{50} = 270\,A$$

The energy supplied by the battery over 70 minutes (1.167) hours is 315 ampere hours, representing 50% of the battery capacity. Hence batteries with a total capacity of 630 AH are needed.

The total battery capacity is 630 AH at 50 V, which is 31.5 kWH. Hence the battery weight is 157.5 kg. A large GEO satellite may have mass up to 6000 kg, so the battery accounts for 2.6% of the satellite's mass in this example.

Example 3.3

Calculate the total power radiated by the sun in watts and in dBW.

Hint: The sun is 93 million miles (about 150 million kilometers) from the earth. At that distance, the sun produces a flux density of 1.36 kW/m^2. This power density is present over all of a sphere with a radius of 150 million km.

Answer
The surface area A of a sphere with radius R m is given by

$$A = 4\,\pi\,R^2\ \mathrm{m}^2$$

The sun is radiating 1.36 kW/m^2 over an area of

$$A = 4\,\pi R^2 = 4 \times \pi \times (150 \times 10^6)^2 = 2.83 \times 10^{17}\ \mathrm{m}^2$$

Hence the power radiated by the sun is P watts where

$$P = 2.83 \times 10^{17} \times 1.36 \times 10^3 = 3.85 \times 10^{20}\ \mathrm{W}$$

Converting to decibels

$$P = 10\log_{10}(3.85 \times 10^{20}) = 205.9\ \mathrm{dB\ W}$$

This represents a maximum value for any power system on earth.

3.5 Communications Subsystems

3.5.1 Description of the Communication System

A communications satellite exists to provide a platform in geostationary orbit for the relaying of voice, video, and data communications. All other subsystems on the satellite exist solely to support the communications system, although this may represent only a small part of the volume, weight, and cost of the satellite in orbit. Since it is the communication system that earns the revenue for the system operator, communications satellites are designed to provide the largest traffic capacity possible. Successive GEO satellites have become larger, heavier, and more costly, but the rate at which traffic capacity has increased has been much greater, resulting in a lower cost per transmitted bit with each succeeding generation of satellite. The satellite transponders have limited output power and the earth stations are at least 35 786 km away from a GEO satellite, so the received power level, even with large aperture earth station antennas, is very small and rarely exceeds 10^{-10} W. For the system to perform satisfactorily, the signal power

typically must exceed the power of the noise generated in the receiver by between 2 and 20 dB, depending on the bandwidth of the transmitted signal and the modulation scheme used. Satellites such as cubesats that have low power transmitters cannot support high bit rates because narrow receiver bandwidths have to be used to maintain the required SNR.

From 1973 to 1998 global space revenue increased at an annual growth rate of 6.3%, from US$15B to US$68.8B, approximately double the world's GDP compound annual growth rate of 2.96%. Over the 16 year period from 1998 to 2014 the world GDP grew at an annual growth rate of 2.71%, while in the same period the space sector economy grew at 10.14%, four times the GDP rate. In 1973 the US government contribution to global space revenue was around 80%, while in 2014 commercial space revenue was 80% of the global space revenue, with a commercial annual growth rate of 13.42% (Space Revenue 2016). The size, weight, and electric power of GEO satellites grew at a corresponding rate until the introduction of multiple beam antennas and Ka-band operation after 2010, when ViaSat I and later GEO satellites achieved a step function in communication capacity with no increase in satellite mass (Viasat 2017).

Early communications satellites were fitted with transponders of 250 or 500 MHz bandwidth, but had low-gain antennas and transmitters of 1 or 2 W output power. The earth stations needed large reflector antennas, typically 25 m diameter, and receivers could not achieve an adequate SNR when the full bandwidth was used with the result that the system was *power limited*. Later generations of communications satellites have transponders with greatly increased output power – up to 2.8 kW for some radio broadcast satellites – and have steadily improved in bandwidth utilization efficiency by frequency reuse. The total channel capacity of a satellite can be increased only if the bandwidth can be increased or reused. The trend in high-capacity satellites has been to reuse the available bands by employing multiple beams at the same frequency (*spatial beam frequency reuse*) and orthogonal polarizations at the same frequency (*polarization frequency reuse*). Large GEO satellites also use both the 14/11 and 30/20 GHz bands to obtain more bandwidth. ViaSat I has 18-fold frequency reuse with 72 antenna beams and dual polarizations and achieves a capacity of over 120 Gbps (Viasat 2017).

The designer of a satellite communication system is not free to select any frequency and bandwidth he or she chooses. International agreements restrict the frequencies that may be used for particular services, and the regulations are administered by the appropriate agency in each country – the Federal Communication Commission (FCC) in the United States, for example. Frequencies allocated to satellite services are listed in Tables 4.1 and 4.2 in Chapter 4. The bands currently used for the majority of services are 6/4, 14/11, and 30/20 GHz. Expansion to higher frequency bands is proposed for some NGSO internet access systems, mainly V-band (40–75 GHz) (V-band 2016).

The 500 MHz bands originally allocated for 6/4 and 14/11 GHz satellite communications quickly became very congested and were extended to cover 1000 MHz bandwidth. By 1980, satellites were using the 14/11 GHz band, which offered another 1000 MHz of bandwidth. The first commercial Ka-band satellites were Wild Blue 1 launched in 2006 to provide internet access to rural areas of the United States, and Astra 3B launched by Société Européenne de Satellites (SES) in 2010 to provide internet access in Europe, the Middle East, and South Africa (Wild Blue 1 2016) (Astra 3B 2018). The spectrum allocated in Ka-band for satellites has a potential bandwidth of 2.4 GHz, although most Ka-band satellites use 1.5 GHz due to international spectrum sharing rules. When combined with the narrow spot beams that can be generated at Ka-band, ViaSat was able to achieve a step function increase in satellite capacity in 2011 from 10 to over 120 Gbps

with the ViaSat I satellite. The standard spacing between GEO satellites was originally set at 3°, but under regulations covering North America and much of the rest of the world, the spacing has been reduced to 2°. The move to 2° spacing opened up extra slots for new satellites.

Satellite systems designed for Ku-band (14/11 GHz) and Ka-band (30/20 GHz), have narrower antenna beams, and better control of coverage patterns than satellites using C-band (6/4 GHz). As the available orbital slots for GEO satellites filled up with satellites using the 6/4 and 14/11 GHz bands, multiple beam satellites began to use the 30/20 GHz band. Originally, this band had 3 GHz bandwidth allocated to satellite services, but part of the band was reallocated to the land multipoint distribution service (LMDS). Approximately 1.5 GHz of bandwidth is available for satellite systems at Ka-band on an exclusive or shared basis, which is similar to the combined allocations of C- and Ku-bands. However, propagation in rain becomes a major factor at frequencies above 10 GHz. Attenuation (in dB) in rain increases as roughly the square of the frequency, so at 20 GHz rain attenuation is four times larger, in dB, than at 10 GHz.

Internet access systems are two way, and establish a virtual connection between the user and the gateway station. When rain attenuation affects the link, the user can request a change in forward error correction rate or a different modulation, to reduce the bit rate on the link and allow a narrower bandwidth to be selected. The narrower bandwidth generates less noise in the receiver and increases the carrier to noise ratio to combat the loss of signal caused by rain attenuation. The European Telecommunications Standards Institute (ETSI) DVB-S2 standard for direct broadcast television has provision for two way connections for internet access using adaptive coding and modulation (ACM) (ETSI EN 302 307 V1.2.1 2009). The concept is to use a high order modulation and low forward error correction rate under clear air conditions when only a small CNR margin is required, to achieve the highest possible bit rate on the link. As rain attenuation starts to reduce the CNR margin, the user terminal sends a message to the gateway station requesting a lower date rate, and reduces its receiver bandwidth. A lower bit rate is preferable to loss of the connection with the gateway and the reduced bit rate is needed only for the duration of the rain event, typically a few minutes. Chapter 4 explains how CNR is set in a satellite link and Chapter 5 explains digital transmission techniques. Chapter 7 discusses propagation through rain on satellite links and Chapter 11 discusses internet access via satellite.

The ETSI DVB-S2 standard is widely used for direct broadcast satellite television (DBS-TV) using constant modulation and coding, but the integrated circuits used in DBS-TV receivers can also be used for internet access. These integrated circuits are built in very large quantities each year – many millions – and are therefore relatively low cost and ideal for internet access receivers. Under clear sky conditions, with a CNR margin of a few decibels, the link can use 8-ASK modulation and 5/6 rate FEC, which requires a CNR around 11.3 dB and delivers 2.5 bits per hertz of link bandwidth. As rain attenuation increases the modulation can change to quadrature phase shift keying (QPSK) and the FEC rate can be reduced to one half, which requires a CNR around 2.8 dB and delivers 1 bit/Hz. A CNR reduction of 8.5 dB in rain can be accommodated with a bit rate reduction by a factor of 2.5. These CNR figures assume an implementation margin of 1.8 dB and are the lowest CNR values at which quasi-error free (QEF) operation can be achieved. For many earth stations in the United States, the link will be lost (an *outage*) for no more than 50 hours in a typical year. See Chapter 11 for details of this approach.

3.5.2 Transponders

A transponder is a device that receives radio signals and retransmits the identical signal at a different frequency, or a different signal at a different frequency. All commercial aircraft carry transponders that are triggered by signals from air traffic control radars. The transmitted signal contains information about the aircraft's identity, altitude, and GPS location, data that the air traffic control system uses to ensure that aircraft are separated by adequate distances and altitudes so as to avoid mid-air collisions. Communication satellite transponders receive signals from uplink earth stations and retransmit the same signal at a different frequency. A transponder that simply amplifies the received signal and changes its frequency for retransmission is known as a linear or *bent pipe* transponder. The analogy is to a pipe that conveys water from one point to another. A transponder that processes the received signal in some way, typically to select which of many downlink beams to use is called an *onboard processing* (OBP) transponder. Bent pipe transponders can carry analog signals; OBP transponders can only handle digital signals.

Transponders on GEO satellites typically have bandwidths of 20–200 MHz and carry many signals. A cubesat or other small satellite might have only one transponder with a bandwidth of 1 MHz or less, designed to convey only one or two signals. The basic principles of all transponders are very similar; signals received by the satellite are generally of low power, typically below −90 dBW and must be amplified to a level of watts for transmission back to earth. End to end gain in a typical GEO satellite transponder exceeds 100 dB, which cannot be achieved at one radio frequency (RF) or intermediate frequency (IF). The transmit frequency of the RF signals must be different from the frequency of the received signals. If the same frequency were used, coupling from the transmitting antenna to the receiving antenna on the satellite would cause feedback and oscillation (sometimes called ringing) within the transponder. This is one reason for uplinks and downlinks to be allocated separate frequency bands. Similarly, the isolation of the output of an RF amplifier from its input rarely exceeds 60 dB, so ringing will occur if gain in excess of 60 dB at one frequency is attempted in a single unit.

Signals (known as *carriers*) transmitted by uplink earth stations are received at the satellite by either a *regional beam* antenna or a *spot beam* antenna. Regional beams can receive from transmitters anywhere within the coverage zone, whereas spot beams have limited coverage. The received signal is taken to two low-noise amplifiers and is recombined at their output to provide *redundancy*. If either amplifier fails, the other one can still carry all the traffic. Since all carriers from one antenna must pass through a low-noise amplifier, a failure at that point is catastrophic. Redundancy is provided wherever failure of one component will cause the loss of a significant part of the satellite's communication capacity.

Figure 3.9a shows a simplified block diagram of a satellite communication subsystem for the 6/4 GHz band. The 500 MHz bandwidth allocated to this satellite is divided up into twelve channels, 36 MHz wide, which are each handled by a separate transponder. There are two sets of transponders, one set for left hand circularly polarized (LHCP) uplink signals and a second set for right hand circularly polarized (RHCP) uplink signals. Each transponder has a band pass filter (BPF) to select the particular channel's band of frequencies, a down converter to change the frequency from 6 GHz at the input to 4 GHz at the output, and a high power output amplifier. There are two receivers operating in parallel; if one fails the other will continue to operate. The parallel

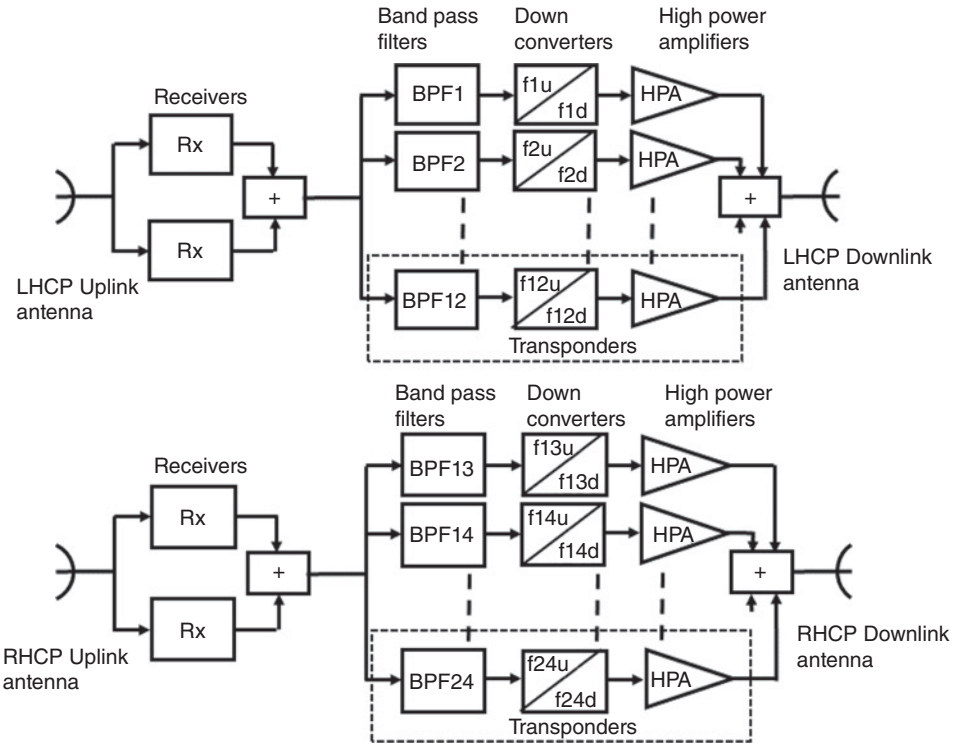


Figure 3.9a Example of a system of 24 transponders for a 6/4 GHz satellite operated with orthogonal circularly polarized signals. Each bandpass filter has a different center frequency according to the frequency plan in Figure 3.9b. The down conversion frequency shift is 2225 MHz. LHCP, left hand circular polarization; RHCP, right hand circular polarization.

arrangement is preferable to an input switch, which creates a catastrophic failure if it fails to operate.

Figure 3.9b shows the frequency plan for this satellite. There are twelve 36 MHz bandwidth transponders for each polarization spaced 40 MHz apart. The uplink channels are labeled *fnu* and the downlink are labeled *fnd*. All downlink channels are centered at a carrier frequency 2225 MHz below the corresponding uplink channel. There is a 20 MHz offset between transponders at the same nominal frequency with orthogonal polarizations. The frequency offset helps the earth station receiver separate the two orthogonally polarized signals and reduces crosstalk between the channels. By employing LHCP and RHCP channels, the effective bandwidth of the communication system, and hence its communication capacity is doubled, compared to a single polarization system. This technique is known as *frequency reuse*.

The satellite's communication system has many transponders, some of which may be spares; typically, 24 to more than 100 active transponders may be carried by a high-capacity GEO satellite. The transponders are supplied with signals from one or more receive antennas and send their outputs to a switch matrix that directs each transponder band of frequencies to the appropriate antenna or antenna beam. In a large satellite there may be four or five beams to which any transponder can be connected. The switch

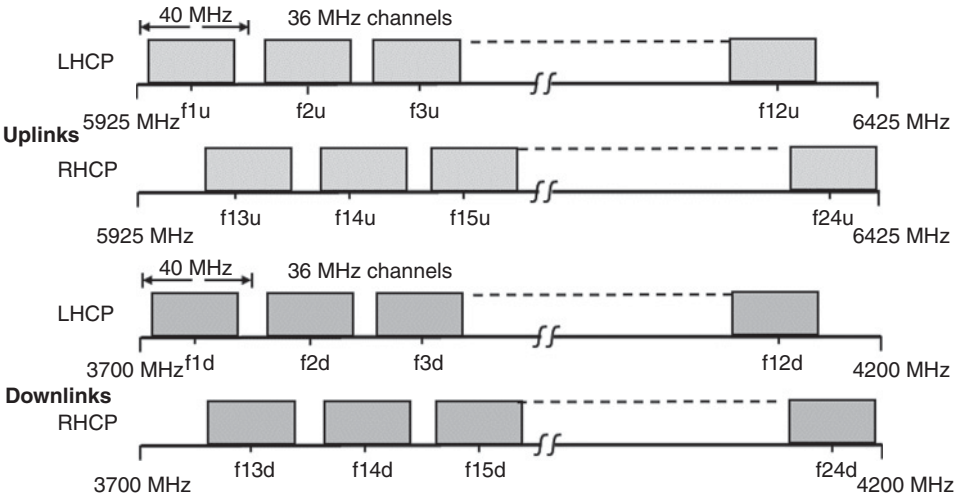


Figure 3.9b Frequency plan for the transponders in Figure 3.9a. Note that there is a 20 MHz frequency shift between transponders for orthogonal polarized signals. *fnu* is an uplink frequency, *fnd* is the corresponding downlink frequency, 2225 MHz below the uplink frequency.

setting can be controlled from earth to allow reallocation of the transponders between the downlink beams as traffic patterns change or transponders fail.

In the early satellites such as INTELSAT I and II, one or two 250-MHz bandwidth transponders were employed. This proved unsatisfactory because of the nonlinearity of the traveling wave tube amplifier (TWTA) used at the output of the transponder, and later GEO satellites for the 6/4 GHz band have used up to 44 transponders each with 36, 54, or 72 MHz bandwidth. The reason for using narrower bandwidth transponders is to avoid excessive *intermodulation* problems when transmitting several carriers simultaneously with a nonlinear transmitter, as discussed in Chapter 6. Intermodulation distortion is likely to occur whenever a high power amplifier (HPA) is driven close to saturation (its maximum output power). Since we generally want to have more than one earth station transmitter sending signals via a satellite, one solution would be to provide one transponder for each earth station's signal, but this could result in a requirement for as many as 100 transponders per satellite. As a compromise, 36 MHz has been widely used for transponder bandwidth, with 54, 72, 100, and 200 MHz adopted for some GEO satellites

Many satellites operating in the 6/4 GHz band carry 24 active transponders, as illustrated in Figure 3.9b. The center frequencies of the transponders are spaced 40 MHz apart, to allow guard bands for the 36 MHz filter skirts. With a total of 500 MHz available, a single polarization satellite can accommodate 12 transponders across the band. When frequency reuse by orthogonal polarizations is adopted, 24 transponders can be accommodated in the same 500 MHz bandwidth. Expansion of the 6/4 GHz band to 1000 MHz allowed additional transponders to be added, or the transponder bandwidth was increased to 72 MHz on some satellites. Internet access satellites need a plethora of beam interconnections – more than 50 in most cases. The only way to achieve this level of beam/path interconnections is via *on board processing*.

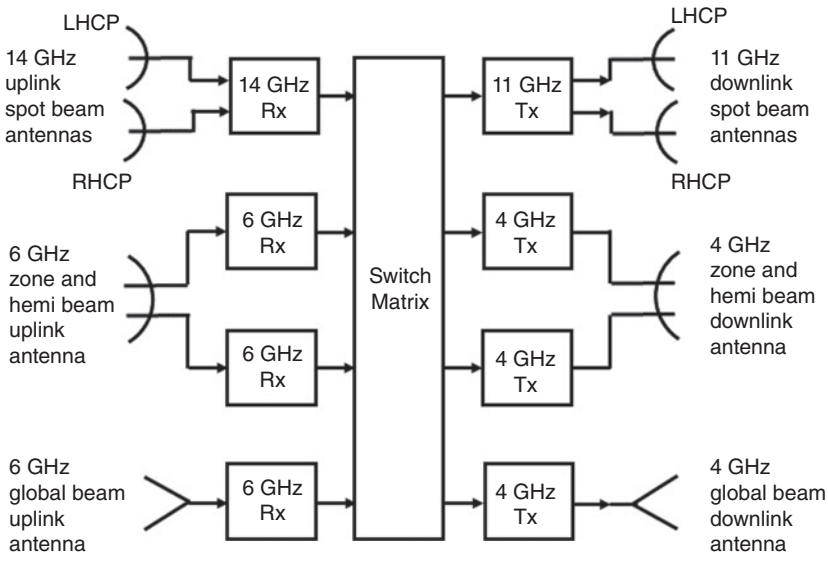
Figure 3.10a shows a simplified diagram of the communication system carried by a typical Intelsat satellite serving the Atlantic Ocean region. Figure 3.10b shows the various beams generated by the satellite. The switch matrix has six inputs and six outputs allowing a very large number of variations in connecting the 6-GHz receivers to the 4-GHz transmitters, and also interconnecting the 6/4 and 14/11 GHz sections. This provides Intelsat with a great deal of flexibility in setting up links through the satellite.

When more than one signal shares a transponder using *frequency division multiple access* (FDMA) the power amplifier must be run below its maximum output power to maintain linearity and reduce *intermodulation* products. The degree to which the transmitter output power is reduced below its peak output is known as *output backoff*. In FDMA systems, 1–7 dB of output backoff is typically used, depending on the number of accesses to the transponder and the extent to which the characteristics of the HPA have been linearized. Backoff results in a lower downlink CNR at the earth station with FDMA when multiple accesses to each transponder are required. *Time division multiple access* (TDMA) can theoretically be used to increase the output power of transponders by limiting the transponder to a single access. However, some TDMA systems are hybrid FDMA-TDMA schemes known as multi-frequency time division multiple access (MF-TDMA), in which several TDMA signals share the transponder bandwidth using FDMA. Linearity of the HPA remains an issue for MF-TDMA systems.

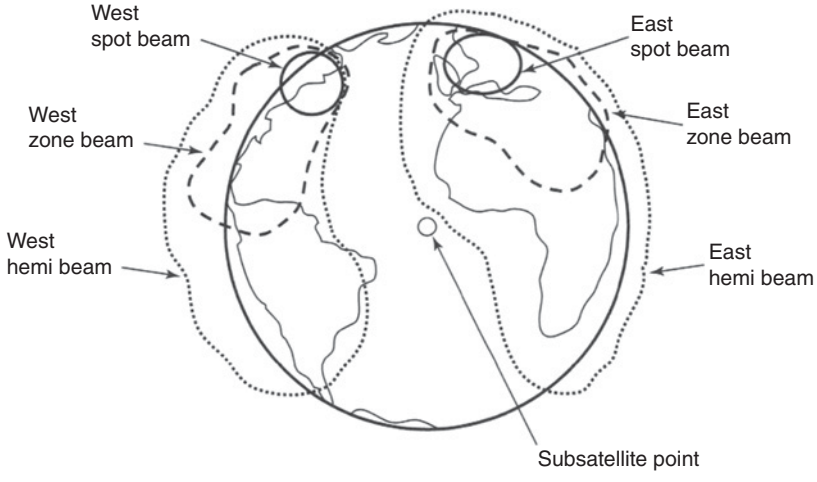
Figure 3.11 shows a typical single conversion bent pipe transponder of the type used on many satellites for the 6/4 GHz band. The signal from the uplink antenna is amplified in a wideband low noise amplifier (LNA) and then applied to a band pass filter that covers the entire 6 GHz uplink band. The signal is then down converted to the 4 GHz band with a mixer and local oscillator (LO). The local oscillator has a frequency of 2225 MHz and is followed by a 4 GHz band pass filter. The frequency conversion process generates two signals known as sum and difference components. The sum signal frequency is equal to the sum of the input frequency and the LO frequency and the difference frequency is equal to the input frequency minus the LO frequency. The difference frequency is needed in a down converter, so a band pass filter must follow the mixer to block the sum component.

For example, if a signal received by the transponder in Figure 3.11 has a carrier frequency of 6.100 GHz and a bandwidth of 36 MHz, the output of the mixer will have two components: a difference component centered at 6.100–2.225 GHz = 3.875 GHz, and a sum component centered at 6.100 + 2.225 GHz = 8.325 GHz. The BPF following the down conversion mixer will be centered at 3.875 MHz to accept the difference component and block the sum component, with a bandwidth of 36 MHz. The 6 GHz receiver in Figure 3.11 typically covers 500 MHz, or a larger bandwidth, depending on which portion of the 6/4 GHz band is used. Multiple carriers can be sent to the satellite from one earth station, or from many earth stations, with a typical carrier frequency spacing of 40 MHz as illustrated in Figure 3.9b.

The output stage of the transponder, the downlink transmitter, typically has two amplifiers in series, a low power amplifier (LPA) and a high power amplifier. The output of the down converter stage is typically 10 mW (10 dBm or – 20 dBW) and must be amplified to 50 or 100 watts, or whatever output power level is desired. For an output of 100 watts, the LPA plus HPA must have a combined gain of 40 dB, which can be achieved with a 30 dB gain LPA and 20 dB gain HPA fed through an electronically controlled attenuator. The attenuator is controlled by commands from earth via the TTC&M system to compensate for loss of gain as the HPA ages. The LPA is a solid state amplifier



(a)



(b)

Figure 3.10 (a) Simplified diagram showing the communication system of a typical Intelsat satellite serving the Atlantic Ocean region using the 6/4 and 14/11 GHz bands. (b) The satellite generates seven beams which operate in both receive and transmit. There is a 6/4 GHz global beam with a small number of transponders, two 6/4 GHz hemisphere beams and two 6/4 GHz zone beams that carry the bulk of the 6/4 GHz traffic, and two 14/11 GHz spot beams centered on North America and Western Europe. The switch matrix is a 6×6 microwave switch that allows interconnection between beams, a form of on-board processing that works for both analog and digital signals.

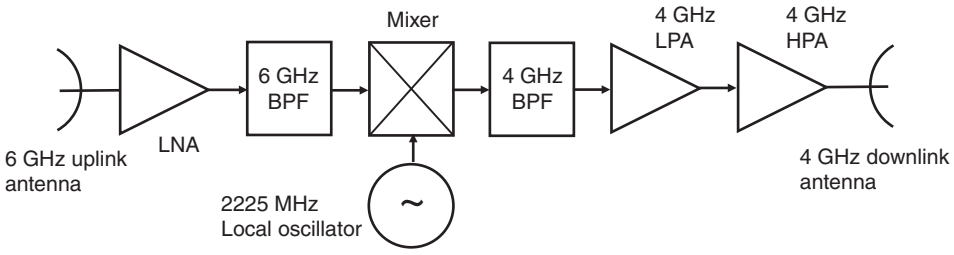


Figure 3.11 Simplified diagram of a bent pipe transponder for the 6/4 GHz band. The mixer and local oscillator form a down converter that changes the 6 GHz frequency of signals received from the uplink to 4 GHz for retransmission to earth. LNA, Low noise amplifier; BPF, Band pass filter; LPA, Low power amplifier; HPA, High power amplifier.

with a maximum output power of a few watts and the HPA for the 4 GHz band is a solid state amplifier capable of a saturated output exceeding 100 watts, but usually operated several decibels below saturation. (See Chapter 4 for a discussion of back off in HPAs.) Traveling wave tube amplifiers are used in transponders of Ka-band satellites, and also where higher output powers are required. For example, the SiriusXM radio broadcast satellites have output powers in the range 800 W to 2.8 kW. Several TWTAs are operated in parallel to achieve this output power.

Reliability is an important consideration in communication satellites, which are required to operate continuously without maintenance for 10 or 15 years. The HPA of a transponder has a lifetime that may be less than the design life of the satellite, which means it may fail before the satellite is due to be replaced. Two HPAs are often used in parallel, with a switch to select which one is operating, to provide *redundancy*. Section 3.7 discusses reliability and the statistical techniques used to analyze the probability of a failure occurring, and how to protect against such failures. Providing a spare HPA in each transponder greatly increases the probability that the satellite will reach the end of its working life with most of its transponders still operational. Transponders can also be arranged so that there are spare HPAs available in the event of a total failure, a technique known as ring redundancy. By including a set of microwave switches at the inputs and outputs of a set of HPAs, a spare amplifier can be switched in to replace one HPA that has failed. A typical arrangement might have 16 HPAs of which 12 are active and four are spares. This is called 4/12 ring redundancy. The two parallel LNAs in Figure 3.9a create 1/2 redundancy (Maral and Bousquet 2002, p. 460).

Transponders for use in the 14/11 and 30/20 GHz bands normally employ a double frequency conversion scheme as illustrated in Figure 3.12. It is easier to make filters, amplifiers, and equalizers at an intermediate frequency such as 1 GHz than at 14 or 11 GHz, so the incoming 14 GHz carrier is translated to an IF around 1 GHz. The amplification and filtering are performed at 1 GHz and a relatively high level carrier is translated back to 11 GHz for amplification by the HPA.

Stringent requirements are placed on the filters used in transponders, since they must provide good rejection of unwanted frequencies, such as intermodulation products, and also have very low amplitude and phase ripple in their pass bands. Frequently a filter will be followed by an *equalizer* that corrects for amplitude and phase variations in the pass band. Phase variation across the pass band produces *group delay distortion,* which is particularly troublesome with high speed phase shift keyed signals.

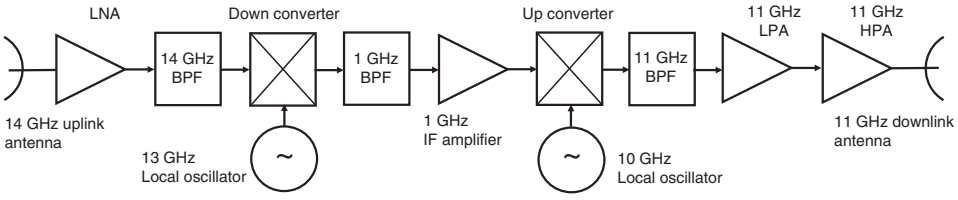


Figure 3.12 Double frequency conversion bent pipe transponder for the 14/11 GHz band. The uplink signals at 14 GHz are downconverted to an intermediate frequency of 1 GHz, amplified, and then up converted to the 11 GHz downlink frequency for retransmission to earth. LNA, Low noise amplifier; BPF, Band pass filter; LPA, Low power amplifier; HPA, High power amplifier; IF, Intermediate frequency.

A considerable increase in the communications capacity of a satellite can be achieved by combining onboard processing with multiple beam technology. A multiple beam satellite generates many narrow spot beams within the coverage footprint of the satellite. The narrow spot beam has a higher gain than a regional beam, creating a higher effective isotropically radiated power (EIRP) that allows higher order modulations to be transmitted. The signals in a regional beam often employ QPSK modulation with half rate FEC, giving a spectral efficiency of one bit per Hz of bandwidth. The higher EIRP of a spot beam provides higher CNR in the earth station receiver so 8-PSK with 5/6 rate FEC can be used, with a spectral efficiency of 2.5 bits per hertz. For example, a DBS-TV satellite with 30 MHz bandwidth transponders can deliver a bit rate of 22.2 Mbps using the DVB-S standard signal format, which uses QPSK modulation and half rate FEC. A spot beam using the DVB-S2 format and 8-PSK with 5/6 rate FEC can transmit 72 Mbps through the same transponder. The narrow beamwidth of spot beams enables RF frequencies to be reused many times, increasing the data transmission capacity of the satellite. Combined with dual polarizations, the ViaSat-I satellite achieves 18 fold frequency reuse in the 30/20 GHz band (ViaSat-I 2017). Spot beams are used on DBS-TV satellites in the United States to transmit high definition television signals that require twice the bit rate of standard definition, and also to transmit local television channels to the specific areas of the country served by those stations. These satellites are discussed in Chapter 10.

Some Ka-band GEO satellites use a phased array antenna that can generate multiple spot beams pointed in different directions on demand. This is useful where a GEO satellite is used to communicate with aircraft flying over oceans to provide internet access and entertainment for the aircraft's passengers. Onboard processing transponders are required for this application because data packets sent to the satellite by the uplink contain information about the location of the aircraft to which signals are to be sent, and that information must be extracted from the signal by the satellite to determine which spot beam to use. Some satellites can move their spot beams a limited amount to optimize the signal at a specific point on earth. The earth subtends an angle of 17° from a GEO satellite, so a phased array antenna on a GEO satellite needs to scan only ±8.5° to point a beam anywhere on earth.

Figure 3.13 illustrates the structure of an OBP transponder. The information needed to select the correct spot beam is contained in the headers of the packets received at the satellite, so the transponder must recover the headers at baseband to extract the spot beam pointing data. A complete digital receiver is required, followed by a conventional transmitter, and a control system that selects which beam receives which packet.

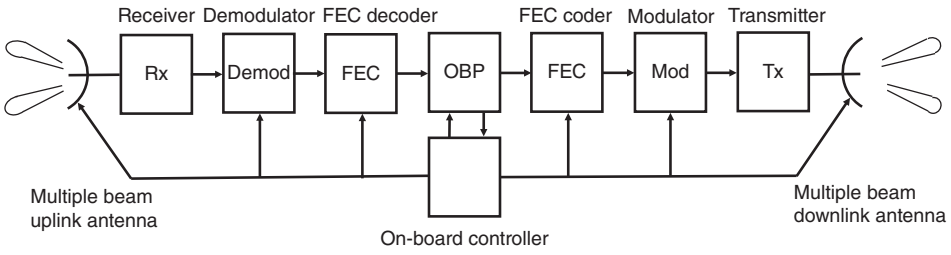


Figure 3.13 Onboard processing (OBP) transponder with multiple beam antennas. The processor extracts the header from each received packet, reads the beam designation, and sets switches to direct the packet to the correct downlink beam. Information in control packets tells the processor which uplink beam to use. Both uplink and downlink can have adaptive coding and modulation. FEC: Forward error correction.

OBP works best with time division multiplexing, where only one signal is present in the transponder at any one time. As packets arrive sequentially at the satellite the header of each packet is read and the packet is routed to the transmitter of the selected spot beam, and the spot beam is pointed at the intended receiving aircraft. The Inmarsat Global Express Ka-band satellites have 89 spot beams, of which 72 can be active at one time and two beams with opposite polarizations can be overlaid to increase capacity to high demand areas (Global Express 2013). Similar techniques are used by LEO internet access satellites – see Chapter 11. Onboard processing can also be used to advantage to switch between the uplink access technique (e.g., MF-TDMA) and the downlink access technique (e.g., TDM) so that small earth stations may access each other directly via the satellite. The processor can provide the data storage needed for a switched-beam system and also can perform error correction independently on the uplink and downlink. The modulation on the uplink and downlink can also be different, depending on the system design and propagation conditions.

Onboard processing transponders have independent receivers and transmitters, allowing the transmitter to operate at a fixed output level. This is called a *regenerative* transponder. When rain affects the uplink reducing the power level of the received uplink signal, the transmitter output power remains the same, unlike a true bent pipe transponder. When the processor has FEC decoding on the uplink, errors can be removed before the packets are transmitted back to the ground. Automatic gain control can be implemented in a bent pipe transponder to maintain a constant transmitter power, but the uplink receiver noise power adds to the earth station receiver noise power and the resulting bit error rate at the receiving earth station is higher than with a regenerative transponder.

3.6 Satellite Antennas

3.6.1 Basic Antenna Types and Relationships

Four main types of antennas are used on satellites. These are

1. *Wire antennas*: monopoles and dipoles.
2. *Horn antennas*.
3. *Reflector antennas*.
4. *Phased array antennas*.

Wire antennas are used primarily at VHF and UHF to provide communications for the TTC&M systems. They are positioned with great care on the body of the satellite in an attempt to provide *omnidirectional* coverage. Most satellites measure only a few wavelengths at VHF frequencies, which makes it difficult to get the required antenna patterns, and there tend to be some orientations of the satellite in which the sensitivity of the TTC&M system is reduced by *nulls* in the antenna pattern.

3.6.2 Antenna Parameters

An *antenna pattern* is a plot of the field strength in the far field of the antenna when the antenna is driven by a transmitter. It is usually measured in *decibels* (dB) below the maximum field strength. The *gain* of an antenna is a measure of the antenna's capability to direct energy in one direction, rather than all around. *Antenna gain* is defined in Chapter 4, Section 4.1, and in Appendix B. At this point, it will be used with the simple definition given above. A useful principle in antenna theory is *reciprocity*. Reciprocity means that an antenna has the same gain and pattern at any given frequency whether it transmits or receives. An antenna pattern measured when receiving is identical to the pattern when transmitting. Appendix B provides a primer on antennas, which supplements the limited information in this section.

Figure 3.14 shows typical satellite antenna patterns for different applications. A GEO satellite that is required to have coverage over the visible earth must have a global beam.

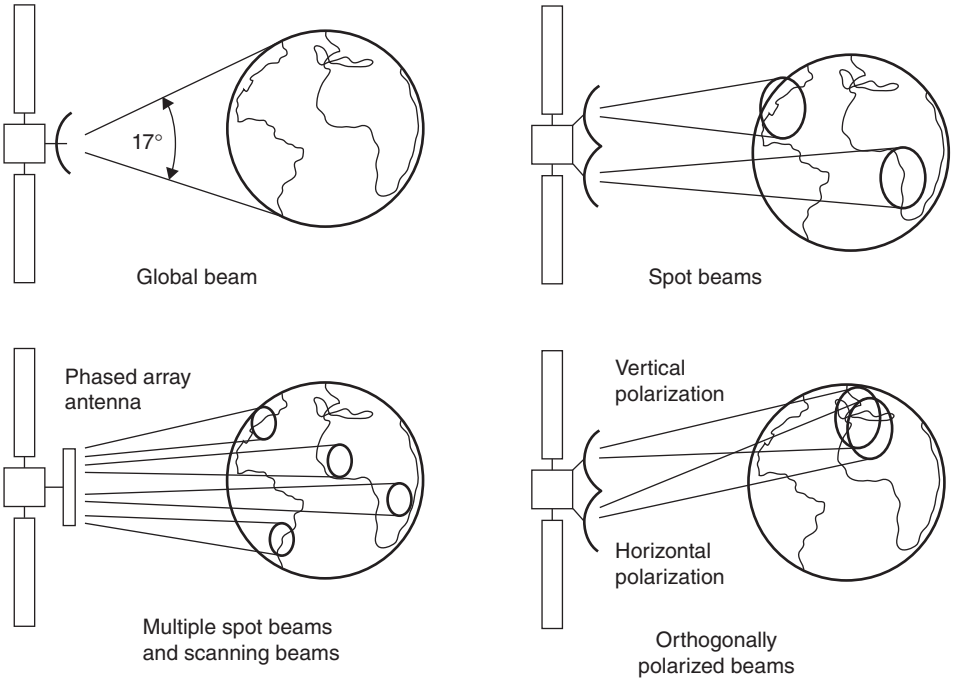


Figure 3.14 Typical satellite antenna patterns and coverage zones. A global beam is 17° wide and can be generated with a horn antenna. Spot beams require reflector antennas and multiple spot beams need a reflector with multiple feeds. Scanning beams typically are generated with a phased array antenna. Orthogonally polarized beams using V and H linear polarizations or left hand and right hand circular polarizations can overlap.

A global beam antenna has a relatively low gain and is only used on GEO satellites that must provide coverage to isolated areas such as islands a long distance from any continent. Regional and spot beams are more widely used, often overlaid as with DBS-TV satellites. Regional beams are carefully shaped to direct radiated power to a specific area. Examples can be seen in Chapter 10 for DBS-TV satellites providing service to the United States, where the shape of the beam is closely tailored to the country's boundaries. Satellites with phased array antennas can move their beams to track moving earth stations, for example aircraft. Orthogonal polarizations, either linear or circular, allow beams to be overlapped

Figure 3.15 shows a typical satellite antenna coverage pattern for Western Europe. The contours are relative signal strength in decibels below the maximum value of 0 dB on the axis of the antenna beam. The contours can also be specified in EIRP, the effective isotropically radiated power in a given direction. EIRP is the product of transmitter power and antenna gain, usually quoted in dBW, decibels greater than one watt. Antenna patterns are frequently specified by their 3 dB beamwidth, the angle between the directions in which the radiated (or received) field falls to half the power

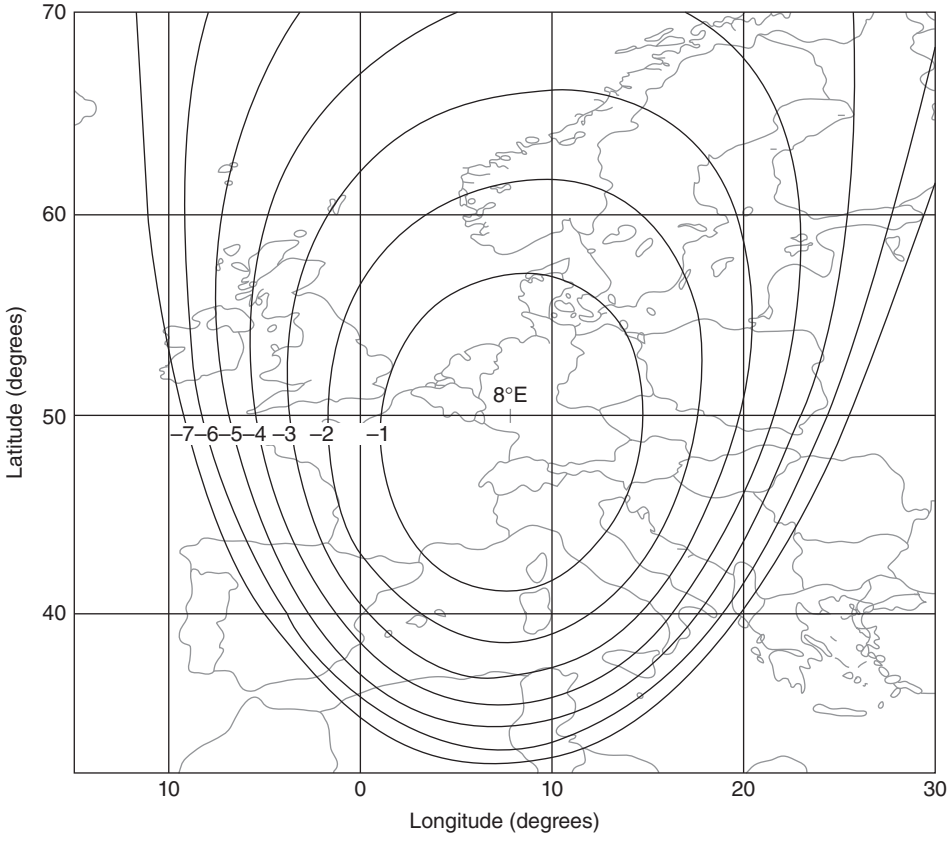


Figure 3.15 Contours of a circular spot beam serving Western Europe from a satellite located at 8° east. The dB values on the contours are relative to the maximum gain of the antenna, which is set to 0 dB. The center of the beam is in western Germany. The contours are spread out at northern latitudes by the curvature of the earth.

relative to the direction of maximum field strength. However, a satellite antenna is used to provide coverage of a certain area, called the *footprint* of the satellite or a *zone* within the footprint, and it is more useful to have contours of satellite EIRP. The beam in Figure 3.15 has a circular cross section; the contours are spread out at higher latitudes by the curvature of the earth.

Horn antennas are used at microwave frequencies when relatively wide beams are required, as for global coverage. A horn is a flared section of waveguide that provides an aperture several wavelengths wide and a good match between the waveguide impedance and free space. Horns are also used as feeds for reflectors, either singly or in clusters. Horns and reflectors are examples of *aperture antennas* that launch a wave into free space from a waveguide. It is difficult to obtain gains much greater than 23 dB or beamwidths narrower than about 10° with horn antennas. For higher gains or narrow beamwidths a reflector antenna or array must be used.

Reflector antennas are usually illuminated by one or more horns and provide a larger aperture than can be achieved with a horn alone. For maximum gain, it is necessary to generate a plane wave in the aperture of the reflector. This is achieved by choosing a reflector profile that has equal path lengths from the feed to the aperture, so that all the energy radiated by the feed and reflected by the reflector reaches the aperture with the same phase angle and creates a uniform phase front. One reflector shape that achieves this with a point source of radiation is the paraboloid, with a feed placed at its focus. The paraboloid is the basic shape for most reflector antennas, and is commonly used for earth station antennas. Satellite antennas often use modified paraboloidal reflector profiles to tailor the beam pattern to a particular coverage zone. Phased array antennas are also used on satellites to create multiple beams from a single aperture, and have been used by Iridium and Globalstar to generate up to 48 beams from a single aperture for their LEO mobile telephone systems (Iridium Next 2018; Globalstar 2017).

3.6.3 Estimating Gain and Beamwidth

Some basic relationships in aperture antennas can be used to determine the approximate size of satellite and earth station antennas, as well as the antenna gain. More accurate calculations are needed to determine the exact gain, efficiency, and pattern of an antenna, and the interested reader should refer to one of the many excellent texts in this field for details (Stutzman and Thiele 2013; Rudge et al. 1983; Silver 1989).

The following approximate relationships will be used here to guide the selection of antennas for a communications satellite.

An aperture antenna has a gain G given by

$$G = \eta_A 4\pi A/\lambda^2 \tag{3.1}$$

where A is the area of the antenna aperture in square meters, λ is the operating wavelength in meters, and η_A is the *aperture efficiency* of the antenna. The aperture efficiency η_A is not easily determined, but is typically in the range 55–70% for reflector antennas with single feeds, lower for antennas with shaped beams. Horn antennas tend to have higher efficiencies than reflector antennas, typically in the range 65–80%. If the aperture is circular, as is often the case, Eq. 3.1 can be written as

$$G = \eta_A(\pi D/\lambda)^2 \tag{3.2}$$

where D is the diameter of the circular aperture. D and λ must have the same units, typically meters.

The beamwidth of an antenna is related to the aperture dimension in the plane in which the pattern is measured. A useful approximation is that the 3 dB beamwidth in a given plane for an antenna with dimension D in that plane is

$$\theta_{3dB} \approx 75\,\lambda/D \text{ degrees} \tag{3.3}$$

where $\theta_{3\text{dB}}$ is the beamwidth between half power points of the antenna pattern and D is the aperture dimension in the same units as the wavelength λ. The beamwidth of a horn antenna may depart from Eq. (3.2) quite radically. For example, a small rectangular horn will produce a narrower beam than suggested by Eq. (3.2) in its E plane and a wider beamwidth in the H plane.

Since both Eqs. 3.2 and 3.3 contain antenna dimension parameters, the gain and beamwidth of an aperture antenna are related. For antennas with $\eta_A \approx 60\%$, the gain is approximately

$$G \approx \frac{33{,}000}{(\theta_{3\ dB})^2} \tag{3.4}$$

where $\theta_{3\text{ dB}}$ is in degrees and G is not in decibels. If the beam has different beamwidths in orthogonal planes, $\theta_{3\text{ dB}}$ should be replaced by the product of the two 3 dB beamwidths. Values of the constant in Eq. 3.3 vary between different sources, with a range 28 000–35 000. The value 33 000 is typical for reflector antennas used in satellite communication systems.

Example 3.4 Global Beam Antenna

The earth subtends an angle of 17° when viewed from geostationary orbit.

What are the dimensions and gain of a horn antenna that will provide global coverage at 4 GHz?

Answer

We can specify a horn to give a circularly symmetric beam with a 3 dB beamwidth of 17° by rearranging Eq. 3.2

$$D/\lambda = 75/\theta_{3\text{ dB}} = 4.4$$

At 4 GHz, $\lambda = 0.075$ m, so $D = 0.33$ m (just over 1 ft). If we use a circular horn excited in the TE_{11} mode, the beamwidths in the E and H planes will not be equal and we may be forced to make the aperture slightly smaller to guarantee coverage in the E plane. *A corrugated horn* designed to support the HE hybrid mode has a circularly symmetric beam and could be used in this application. Waveguide horns are generally used for global beam coverage. Reflector antennas are not efficient when the aperture diameter is less than 8 λ.

Using Eq. 3.3, the gain of the horn is approximately 100, or 20 dB, at the center of the beam. However, in designing our communication system we will have to use the edge of beam gain figure of 17 dB, since those earth stations close to the earth's horizon, as viewed from the satellite, are close to the −3 dB contour of the transmitted beam.

Example 3.5 Regional Coverage Antenna

The continental United States (48 contiguous states) subtends angles of approximately $5° \times 2.5°$ in the E–W and N–S directions when viewed from geostationary orbit.

What dimension must a reflector antenna have to illuminate half this area with a circular beam 2.5° in diameter at 11 GHz?

Can a reflector be used to produce a $5° \times 2.5°$ beam?

What is the gain of the antenna?

Answer

Using Eq. 3.2, we have for a 2.5° circular beam

$$D/\lambda = 75/2.5 = 30$$

and with $\lambda = 0.0272$ m, the antenna diameter $D = 0.82$ m (2.7 ft). The gain of this antenna, from Eq. 3.4 is approximately 34 dB.

To generate a beam with different beamwidths in orthogonal planes we need an aperture with different dimensions in the two planes. To produce a beam with beamwidths of $5° \times 2.5°$ requires a rectangular (or elliptical) aperture $15\,\lambda \times 30\,\lambda$. In order to illuminate such a reflector, a feed horn with unequal beamwidths is required, since the reflector must intercept most of the radiation from the feed for it to have an acceptable efficiency. Rectangular, or more commonly elliptical outline reflectors are used to generate unequal beamwidths. When orthogonal polarizations are to be transmitted or received, it is better to use a circular reflector with a distorted profile to broaden the beam in one plane, or a feed cluster to provide the appropriate amplitude and phase distribution across the reflector.

3.6.4 Satellite Antennas in Practice

The antennas of a communications satellite are often a limiting element in the complete system. In an ideal satellite, there would be one antenna beam for each earth station, completely isolated from all other beams, for transmit and receive. However, if two earth stations are 300 km apart on the earth's surface and the satellite is in geostationary orbit, their angular separation at the satellite is 0.5°. For $\theta_{3\,\mathrm{dB}}$ to be 0.5°, D/λ must be 150, which requires an aperture diameter of 11.3 m at 4 GHz. Antennas this large have been flown on satellites (ATS-6 deployed a 2.5 GHz, 10 m diameter antenna, for example, see Figure 3.16), and large unfurled antennas are used to create multiple spot beams from GEO satellites serving mobile users. At 30 GHz, an antenna with $D/\lambda = 150$ is 1.5 m wide, and spot beam antennas can readily be flown on 14/11 GHz and 30/20 GHz satellites, making multiple beam satellites feasible. A phased array feed or a cluster of feed horns illuminating a reflector can be used to create many 0.5° beams to serve the coverage zone of the satellite. Multiple beam antennas are discussed in Appendix B on antennas, and in Chapter 10 for DBS-TV satellites.

To provide a separate beam for each earth station also requires one antenna feed per earth station when a multiple-feed antenna with a single reflector is used. A compromise between one beam per station and one beam for all stations has been used in many satellites by using zone-coverage beams and orthogonal polarizations within the same beam to provide more channels per satellite. Figure 3.10b shows the coverage zones provided by a typical Intelsat satellite. The largest reflector on the satellite transmits at 4 GHz and produces the peanut shaped patterns for the zone beams, which are designed to serve populated areas such as North America and Western Europe where much telecommunications traffic is generated. This reflector antenna also produces the hemi beams. The smaller antennas are used to transmit and receive the 14/11 GHz spot beams, which concentrate service on the east coast of North America and Western Europe. In addition, there are horn antennas providing global beam coverage at 6 and 4 GHz.

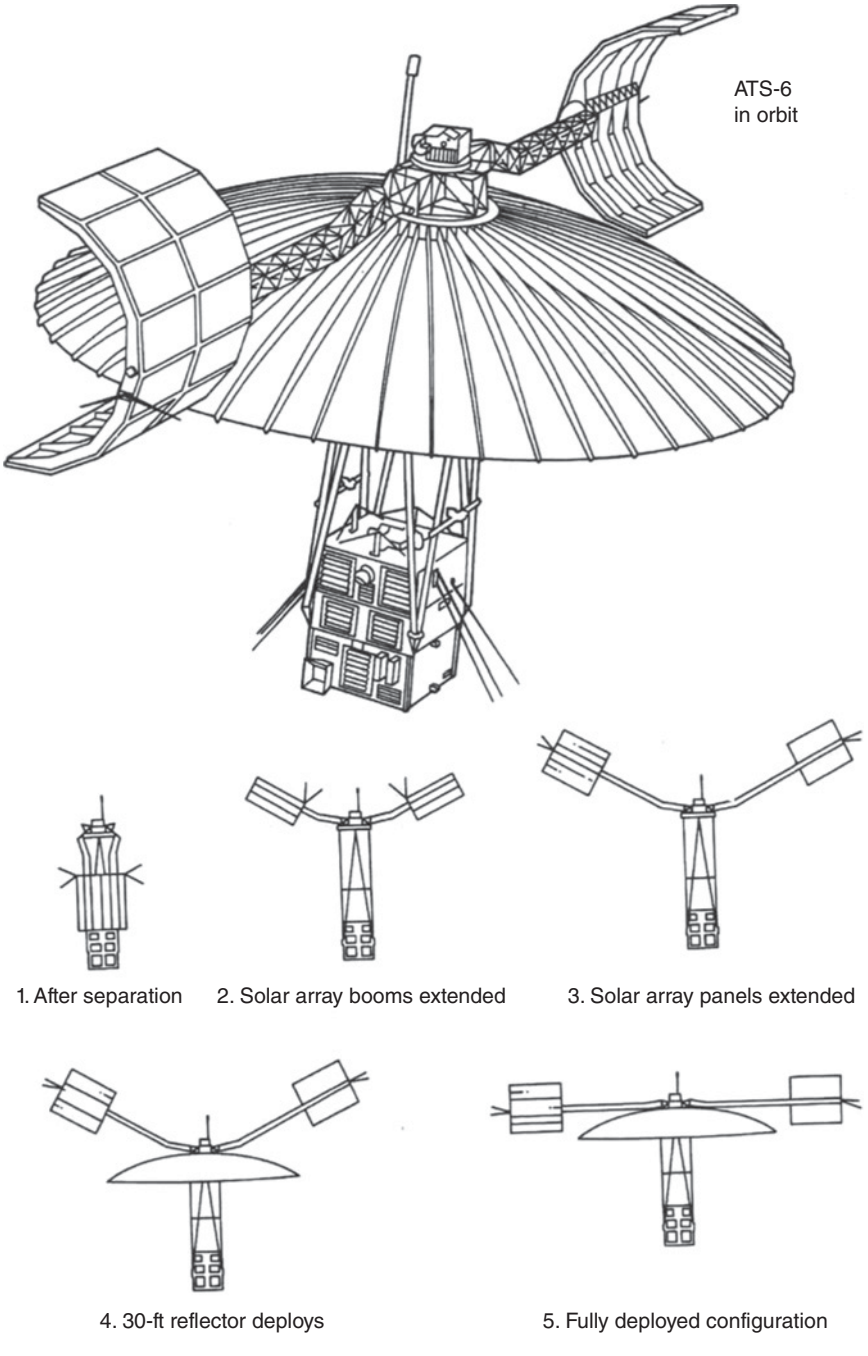


Figure 3.16 Deployment sequence of the 10 m diameter antenna on the ATS-6 satellite. For launch, the reflector folded down like an umbrella with the curved solar cells on top. Source: NASA.

The requirements of narrow antenna beams with high gain over a small coverage zone leads to large antenna structures on the satellite. Frequently, the antennas in their operating configuration are too large to fit within the shroud dimensions of the launch vehicle, and must be folded down during the launch phase. Once in orbit, the antennas can be deployed. In many larger satellites, the antennas use offset paraboloidal reflectors with clusters of feeds to provide carefully controlled beam shapes. The feeds mount on the body of the satellite, close to the communications subsystem, and the reflector is mounted on a hinged arm.

Figure 3.16 shows the deployment sequence used for the 30 ft antenna carried by NASA's ATS-6 satellite. The antenna was built as a series of petals that folded over each other to make a compact unit during launch, which then unfurled in orbit. The solar arrays folded down over the antenna, and were deployed first. Springs or pyrotechnic devices can be used to provide the energy for deployment of antennas or solar array, with a locking device to ensure correct positioning after deployment. Similar unfurlable antennas are used on GEO satellites that provide satellite telephone service at L-band using multiple narrow beams. The deployment mechanism is similar to an umbrella.

One interesting idea is the inflatable antenna, several examples of which have been flown experimentally (Inflatable antennas 1999). The antenna can be squeezed into a small space for launch and inflated from a pressurized gas bottle when the satellite is in orbit. Once inflated, a foam material emitted along with the inflation gas hardens to make a rigid structure. Plastic materials can be sprayed with a metallic coating made up of very small particles of aluminum to create a reflecting surface for electromagnetic (EM) waves. Inflatable antennas can be made very large without a significant weight penalty, and are therefore attractive for any satellite requiring multiple narrow beams. Some inflatable antennas are illustrated in Chapter 8.

3.7 Equipment Reliability and Space Qualification

Large GEO communications satellites are designed to provide operational lifetimes of up to 15 years. Once a satellite is in geostationary orbit, there is little possibility of repairing components that fail or adding more fuel for station keeping. The components that make up the satellite must therefore have very high reliability in the hostile environment of outer space, and a strategy must be devised that allows some components to fail without causing the entire communication capacity of the satellite to be lost. Two separate approaches are used: *space qualification* of every part of the satellite to ensure that it has a long life expectancy in orbit and *redundancy* of the most critical components to provide continued operation when one component fails.

3.7.1 Space Qualification

Outer space, at geostationary orbit distances, is a harsh environment. There is a total vacuum and the sun irradiates the satellite with 1.36 kW of heat and light on each square meter of exposed surface. Where surfaces are in shadow, heat is lost to the infinite sink of space and surface temperature will fall toward absolute zero. Electronic equipment cannot operate at such extremes of temperature and must be housed within the satellite body and heated or cooled so that its temperature stays within the range 0°–75° C. This

requires a thermal control system that manages heat flow throughout a GEO satellite as the sun moves around the satellite once every 24 hours. Thermal problems are equally severe for a LEO satellite that moves from sunlight to shadow every 100 minutes.

The first stage in ensuring high reliability in a satellite is by selection and screening of every component used. Past operational and test experience of components indicates which components can be expected to have good reliability. Only components that have been shown to have high reliability under outer space conditions will be selected. Each component is then tested individually (or as a subsystem) to ensure that it meets its specification. This process is known as *quality control* or *quality assurance* and is vital in building any equipment that is to be reliable. Once individual components and subsystems have been space qualified, the complete satellite must be tested as a system to ensure that its many systems are reliable.

When a satellite is designed, three prototype models are often built and tested. The *mechanical model* contains all the structural and mechanical parts that will be included in the satellite and is tested to ensure that all moving parts operate correctly in a vacuum, over a wide temperature range. It is also subjected to vibration and shock testing to simulate vibration levels and G forces likely to be encountered on launch. The *thermal model* contains all the electronics packages and other components that must be maintained at the correct temperature. Often, the thermal, vacuum, and vibration tests of the entire satellite will be combined in a thermal vacuum chamber for what is known in the industry as a *shake and bake* test. The antennas are usually included on the thermal model to check for distortion of reflectors and displacement or bending of support structures. In orbit, an antenna may cycle in temperature from above 100° C to below −100° C as the sun moves around the satellite. The *electrical model* contains all the electronic parts of the satellite and is tested for correct electrical performance under total vacuum and a wide range of temperatures. The antennas of the electrical model must provide the correct beamwidth, gain, and polarization properties.

Testing carried out on the prototype models is designed to overstress the system and induce failure in any weak components: temperature cycling can be carried out to 10% beyond expected extremes; structural loads and G forces 50% above those expected in flight can be applied. Electrical equipment will be subjected to excess voltage and current drain to test for good electronic and thermal reliability. The prototype models used in these tests will not usually be flown. A separate flight model (or several models) will be built and subjected to the same tests as the prototype, but without the extremes of temperature, stress, or voltage. Preflight testing of flight models, while exhaustive, is designed more to cause failure of parts, rather than to check that they will operate under worst-case conditions.

Space qualification is an expensive process, and one of the factors that makes large GEO satellites expensive. Some low earth orbit satellites have been built successfully using less expensive techniques and relying on lower performance in orbit. LEO satellite systems require large numbers of satellites that are generally less expensive than large GEO satellites. The Iridium system, for example, was designed with 66 operational satellites in its constellation to provide continuous worldwide coverage, with at least eight spare satellites in orbit at any time. If one operational satellite fails, a spare is moved in to take its place. This allowed Iridium satellites to be built with a higher probability of failure than a GEO satellite. Many cubesat satellites have also been built using low cost techniques. Most of the components on the cubesat satellites are not space qualified, but are selected based on past experience of which components have survived well on previous missions.

Many of the electronic and mechanical components that are used in satellites are known to have limited lifetimes, or a finite probability of failure. If failure of one of these components will jeopardize the mission or reduce the communication capacity of the satellite, a backup, or *redundant*, unit will be provided. The design of the system must be such that when one unit fails, the backup can automatically take over or be switched into operation by command from the ground. For example, redundancy is always provided for traveling wave tube amplifiers used in the transponders of a communications satellite, as these are vacuum tubes that are known to have a limited lifetime.

The success of the testing and space qualification procedures used by NASA has been well illustrated by the lifetime achieved by many of its scientific satellites. Satellite designed for a specific mission lasting one or two years have frequently operated successfully for up to 25 years. Sufficient reliability was designed into the satellite to guarantee the mission lifetime such that the actual lifetime has been much greater. In the next section we will look at how reliability can be quantified.

3.7.2 Reliability

We need to be able to calculate the reliability of a satellite subsystem for two reasons: we want to know what the probability is that the subsystem will still be working after a given time period, and we need to provide redundant components or subsystems where the probability of a failure is too great to be accepted. The owner of a communications satellite expects to be able to use a predetermined percentage of its communications capacity for a given length of time. Amortization of purchase and launch costs is calculated on the basis of an expected lifetime. The manufacturers of satellites must provide their customers with predictions (or guarantees) of the reliability of the satellite and subsystems: to do this requires the use of *reliability theory*. Reliability theory is a mathematical attempt to predict the future and is therefore less certain than other mathematical techniques that operate in absolute terms; it is statistical in nature. The application of reliability theory has enabled satellite engineers to build satellites that perform as expected, at acceptable construction costs. It should be noted, however, that the cost of a satellite is very high compared to other equipment with a comparable number of components: a large GEO satellite costs around US$125M to build, close to the cost of a Boeing 777 jet airliner. The cost is acceptable because of the high revenue earning capability of the satellite.

The reliability of a component can be expressed in terms of the probability of failure after time *t*, $P_F(t)$. For most electronic equipment, probability of failure is higher at the beginning of life – the burn-in period – than at some later time. As the component ages, failure becomes more likely, leading to the *bathtub curve* shown in Figure 3.17.

The bathtub curve is familiar to owners of automobiles. A new car may have defects when it is delivered, and errors in manufacturing may lead to components failing soon after purchase. This is one of the reasons that manufacturers offer warranties for the initial life of their products. Once these defects have been overcome, by repair or replacement, reliability improves for a number of years until mechanical parts start to wear and failures occur. In an automobile, preventive maintenance can be carried out to replace parts that are known to wear most quickly. For example, spark plugs and drive belts may be replaced every 100 000 miles. The skill in owning an automobile is to judge the time

at which the vehicle is starting up the end-of-life portion of the bathtub curve. That is the optimum time to dispose of the vehicle, and the worst possible time to buy one.

Preventive maintenance is generally not possible with a geostationary satellite, although there are proposals for small maneuverable spacecraft that can reach GEO orbit to inspect and service GEO satellites (Effective Space Solutions 2017). Refueling GEO satellites and replacing failed electronic components could extend their operational lifetime. Inspection of satellites in orbit may become necessary if there is believed to be a risk of destructive devices being located next to orbiting satellites by a hostile power. It is relatively easy to launch a small bomb that can be raised to GEO orbit and maneuvered next to a satellite. The bomb can be detonated remotely from earth by a controlling earth station. Destruction of all satellites serving the Americas and Europe, for example, would severely limit the capabilities of western armed forces, which rely heavily on satellite communications. Cutting the transoceanic fiber optic cables would then relegate international communications to the high frequency (HF) radio band.

Components for satellites are selected only after extensive testing. The aim of the testing is to determine reliability, causes of failure, and expected lifetime. The result is a plot similar to Figure 3.17. Testing is carried out under rigorous conditions, representing the worst operating conditions likely to be encountered in space, and may be designed to accelerate failure in order to shorten the testing duration needed to determine reliability. Units that are exposed to the vacuum of space are tested in a vacuum chamber, and components subjected to sunlight are tested under equivalent radiant heat conditions. The initial period of reduced reliability can be eliminated by a burn-in period before a component is installed in the satellite. Semiconductors and integrated circuits that are required to have high reliability are subjected to burn-in periods from 100 to 1000 hours, often at a high temperature and excess voltage to induce failures in any suspect devices and to get beyond the initial low reliability part of the bathtub curve.

Spare devices such as solid state high power amplifiers (SSPAs) and TWTAs that have been subjected to testing but not flown on satellites provide a valuable resource for the study of failures in orbit. When a device fails on a satellite, the spare device can be subjected to similar conditions in the laboratory to determine the cause of the failure, using telemetry data from the satellite to determine actual voltages, currents, and temperature on board the satellite at the time of failure.

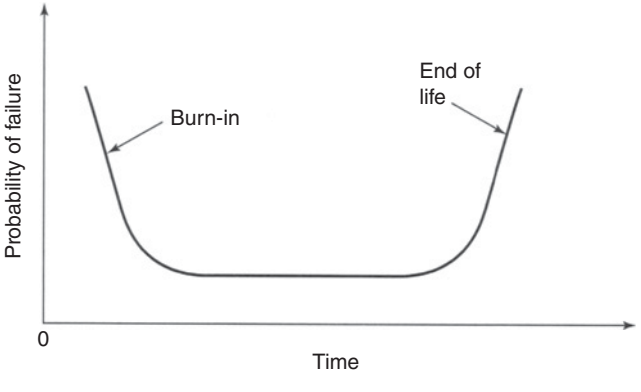


Figure 3.17 Bathtub curve for probability of failure. The burn-in period is also referred to as infant mortality. Once the burn-in period is passed, reliability remains constant until parts start to wear out.

The reliability of a device or subsystem is defined as

$$R(t) = \frac{N_s(t)}{N_o} = \frac{\text{Number of surviving components at time } t}{\text{Number of components at start of test period}} \tag{3.5}$$

The numbers of components that failed in time t is $N_f(t)$ where

$$N_f(t) = N_o - N_s(t) \tag{3.6}$$

From the engineering viewpoint, what we need to know is the probability of any one of the N_o components failing: this is related to the *mean time between failures* (MTBF). Suppose we continue testing devices until all of them fail. The ith device fails after time t_i where

$$\text{MTBF} = m = \frac{1}{N_o}\sum_{i=1}^{N_o} t_i \tag{3.7}$$

The *average* failure rate λ, is the reciprocal of the MTBF, m. If we assume that λ is a constant, then

$$\lambda = \frac{\text{Number of failures in a given time}}{\text{Number of surviving components}}$$

$$\lambda = \frac{1}{N_s}\frac{\Delta N_f}{\Delta t} = \frac{1}{N_s}\frac{dN_f}{dt} = \frac{1}{\text{MTBF}} \tag{3.8}$$

Failure rate λ is often given as the average failure rate per 10^9 hours. The rate of failure, dN_f/dt, is the negative of the rate of survival dN_s/dt, so we can redefine λ as

$$\lambda = -\frac{1}{N_s}\frac{dN_s}{dt} \tag{3.9}$$

By definition from Eq. 3.4, the reliability R is N_s/N_o, so

$$\lambda = \frac{-1}{N_o R}\frac{d}{dt}(N_o R) = \frac{-1}{R}\frac{dR}{dt} \tag{3.10}$$

A solution of Eq. 3.10 is

$$R = e^{-\lambda t} \tag{3.11}$$

Thus the reliability of a device decreases exponentially with time, with zero reliability after infinite time, that is, certain failure. However, end of useful life is usually taken to be the time t_l, at which R falls to 0.37 (1/e), which is when

$$t_l = \frac{1}{\lambda} = m \tag{3.12}$$

The probability of a device failing, therefore, has an exponential relationship to the MTBF and is represented by the right hand end of the bathtub curve.

3.7.3 Redundancy

The equations in the preceding section allow us to calculate the reliability of a given device when we know its MTBF. In a satellite, many devices are used, each with a different MTBF, and failure of one device may cause catastrophic failure of a complete

subsystem. If we incorporate redundant devices, the subsystem can continue to function correctly. We can define three different situations for which we want to compute subsystem reliability: series connection, used in solar cells arrays, parallel connection, used to provide redundancy of the HPAs in satellite transponders, and a switched connection, often used to provide parallel paths with multiple transponders. These are illustrated in Figure 3.18; also shown is a hybrid arrangement, a series/parallel connection, widely used in electronic equipment,

The switched connection arrangement shown in Figure 3.18d is also referred to as *ring redundancy* since any component can be switched in for any other. Switches S_1 and S_2 are a little more complicated than as shown, affording the choice of multiple paths in an M for N ring redundancy configuration. The important point to note is that the active devices (R_1, R_2, … R_n) have sufficient bandwidth, power output range, and so on, to be able to handle any of the channels that might be switched through to them. Most TWTAs and SSPAs are wideband, large power range devices.

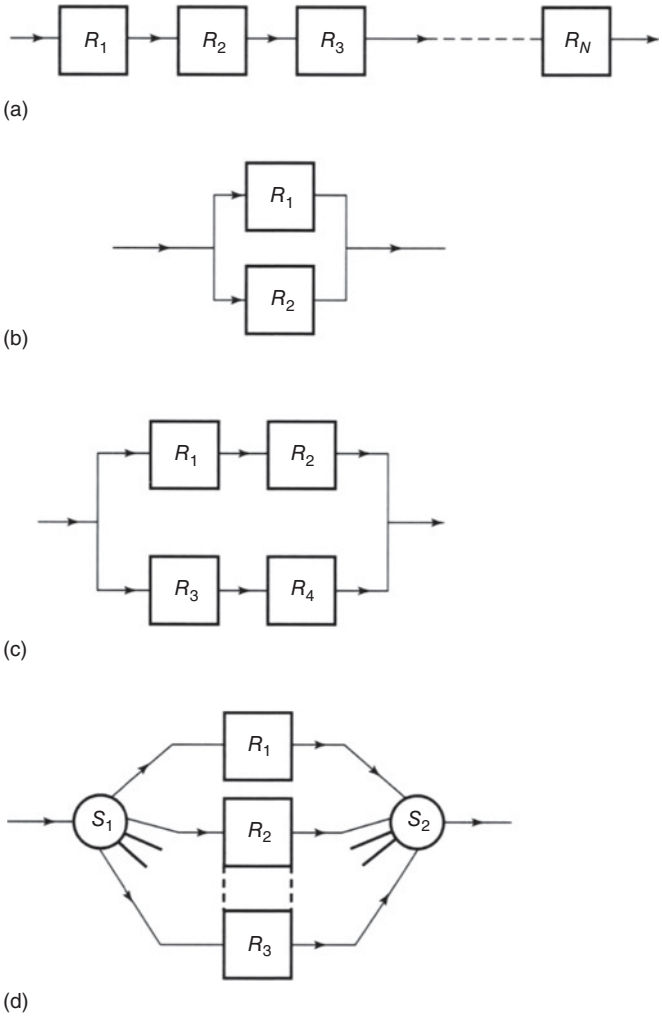


Figure 3.18 Redundancy connections. Each block *R* is a device with known reliability. (a) Series connection. (b) Parallel connection. (c) Series/Parallel connection. (d) Switched connection.

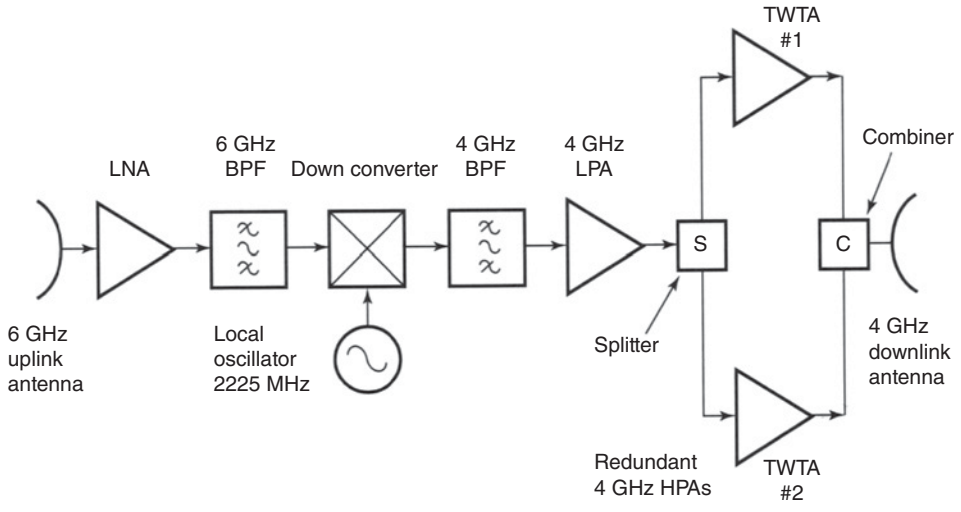


Figure 3.19 Redundant TWTA configuration in a 6/4 GHz transponder. The TWTAs are connected in parallel, but at launch TWTA #2 is switched off. If the first TWTA fails, TWTA #2 is switched on. BPF, Band pass filter; LPA, Low power amplifier; TWTA, Traveling wave tube amplifier.

An example of parallel redundancy for the HPA of a 6/4 GHz bent pipe transponder is shown in Figure 3.19. The transponder translates incoming signals in the 6 GHz uplink band by 2225 MHz and retransmits them in the 4 GHz band. The high power output stage of the transponder has two parallel TWT amplifiers. One TWTA will be switched off, but must present a matched load when both off and on. If one TWTA fails, the other is switched on either automatically, or by command from earth. The TWT is a thermionic device with a heated cathode and a high voltage power supply. In common with other thermionic devices such as cathode ray tubes and magnetrons, they have a relatively short MTBF. Although the MTBF may be 50 000 hours (5.7 years), this is the period after which 50% of such devices will have failed, on average.

The parallel connection of two TWTs, as shown in Figure 3.19 raises the reliability of the amplifier stage to 0.60 at the MTBF period, assuming zero probability of a short circuit. A lifetime of 50 000 hours is less than the typical design lifetime of a large GEO communications satellite. To further improve the reliability of the transponders, a second redundant transponder may be provided with switching between the two systems. Note that a combination of parallel and switched redundancy is used to combat failures that are catastrophic to one transponder channel and to the complete communication system.

3.8 Summary

Satellites that carry communications relays must provide a stable platform in orbit. Large GEO satellites have payload design lives that exceed 10 years and sufficient fuel to provide a maneuvering lifetime that typically exceeds 15 years. The satellite must carry a number of subsystems to support its communications mission. The attitude and orbital control system keeps the satellite in the correct orbit and on station, and pointing in the correct direction. The TTC&M system allows an earth station to control the subsystems in the satellite and to monitor their health. The power system provides the

electrical energy needed to run the satellite (housekeeping) and the communications system. Solar cells generate the electrical power, a power conditioning unit controls its distribution, and batteries provide power during launch, eclipses with GEO satellites, and darkness with LEO satellites.

Satellites often employ frequency reuse, either by using the same frequencies again in spatially separated beams or by using the same frequencies in orthogonal polarizations within the same beam. Frequently both reuse techniques are used simultaneously. Frequency reuse allows the same RF spectrum to be used more than once to increase the satellite's capacity, up to twenty times in some large GEO satellites.

Antennas are a limiting factor in all radio communication systems. Very complex antennas have been developed for satellites to provide multiple beams and orthogonal polarizations from a single antenna. Reflector antennas with clustered feeds and phased array antennas are used to generate shaped and multiple beams. Reliability is an important issue in satellites. Redundancy can be used to provide additional receivers and HPAs that can take over when a unit fails.

Exercises

3.1 The telemetry system of a geostationary communications satellite samples 150 sensors on the spacecraft in sequence. Each sample is transmitted to earth as 10 eight-bit words in a TDM frame. An additional 200 bits are added to the frame for synchronization and status information. The data are then transmitted at a rate of 2.0 kilobit per second using binary phase shift keying (BPSK) modulation of a low power carrier.

- **a.** How long does it take to send a complete set of samples to earth from the satellite?
- **b.** Including the propagation delay, what is the longest time the earth station operator must wait between a change in a parameter occurring at the spacecraft and the new value of that parameter being received via the telemetry link? (Assume a path length of 37 000 km.)

3.2 A three-axis stabilized satellite has two solar arrays. Each array has an area of $5\,m^2$. At the beginning of life each of the solar sails generate 4 kW of electrical power, which falls to 3.6 kW at the end of the satellite's useful lifetime. 1 kW of power is needed for housekeeping (running the satellite systems) and 3 kW is devoted to the telecommunication system.

- **a.** Calculate the efficiency of the solar cells at beginning of life. Assume an incident solar power of $1.36\,kW/m^2$, and normal incidence of sunlight on the sails.
- **b.** Calculate the efficiency of the solar cells at end of life.
- **c.** If one of the solar sails fails and produces no power, how much power is available to the telecommunication system.

3.3 A very low earth orbit satellite orbits the earth in 92 minutes and is in darkness for half of each orbit. The satellite requires 2.5 kW of electrical power, which must be supplied by batteries when the satellite is in darkness. While the satellite is in sunlight, its solar cells must supply sufficient power to run the satellite and charge the batteries. Charging the batteries requires 1.25 times the power that is needed to operate the satellite.

a. How much power must the solar cells supply while the satellite is in sunlight?
b. The solar cells supply current at 50 V. How much current do they deliver?
c. The satellite is in darkness for 46 minutes of each orbit. What capacity must the batteries have, in ampere hours?

3.4 A DBS-TV receiving antenna has a circular aperture with a diameter of 0.6 m and operates at a center frequency of 12.2 GHz. The antenna has an aperture efficiency of 70%.
a. Calculate the wavelength at 12.2 GHz.
b. Calculate the gain of the antenna in decibels using the most accurate formula.
c. Estimate the beamwidth of the antenna in degrees.
d. Estimate the gain of the antenna in decibels using an approximation based on the beamwidth you found in part (c). If this figure does not agree with your result in part (b) explain why.

3.5 The DBS-TV antenna in Question 4 is replaced with an antenna with an elliptical reflector with aperture dimensions 1.0 by 0.6 m.
a. Estimate the beamwidths of the antenna in degrees in each principal plane (the major and minor axis directions).
b. Estimate the gain of the antenna using an approximate formula.
c. The area of an ellipse is given by $A = \pi\ a\ b$ where 2a and 2b are the dimensions of the ellipse on its major and minor axes (1.0 and 0.6 m in the problem). Calculate the gain of the antenna assuming an aperture efficiency of 70%.

3.6 A geostationary satellite provides service to a region, which can be covered by the beam of an antenna on the satellite with a beamwidth of 2.3°. The satellite carries transponders for Ku-band and Ka-band, with separate antennas for transmit and receive. For center frequencies of 14.5/11.0 and 29.5/19.5 GHz, determine the diameters of the four antennas on the satellite.
a. Find the diameters of the two transmitting antennas. Specify the diameter and calculate the gain at each frequency.
b. Find the diameters of the two receiving antennas. Specify the diameter and calculate the gain at each frequency.
c. If any two antennas have the same diameter, explain why.

3.7 A DBS-TV geostationary satellite provides communications within Western Europe at Ku band. The antennas on the satellite have beamwidths of 3° in the E–W direction and 1.5° in the N–S direction. The downlink antenna on the satellite used for broadcasting TV signals operates at a center frequency of 12.25 GHz. The uplink to the satellite operates at 17.5 GHz with a separate satellite antenna.
a. Estimate the gain of the transmitting antenna. Find its dimensions in the N-S and E-W directions.
b. Estimate the gain of the receiving antenna. Find its dimensions in the N-S and E-W directions.

3.8 An earth station antenna has a circular aperture with a diameter of 6.5 m and an aperture efficiency of 68% at 18.6 GHz.
a. Calculate the gain of this antenna and estimate its beamwidth.

b. The aperture efficiency of the antenna is 64% at a frequency of 29.0 GHz. Calculate the gain of this antenna and estimate its beamwidth at 29 GHz.

3.9 A constellation of LEO satellites has an altitude of 400 km. Each satellite has two multiple beam antennas that generate 52 beams for communication with small user terminals on earth. One antenna is used to transmit at 39.0 GHz and the other antenna receives at 43.0 GHz.

a. Using simple geometry, find the coverage angle of the satellite antenna when the lowest elevation angle for an earth station is 40°. (Hint: Draw a diagram of the earth and the satellite and use the law of sines to solve the angles in a triangle. Use an earth radius value of 6378 km.)

b. The footprint of the satellite antenna is a circle on the earth's surface, encompassing the 52 beams. Calculate the diameter of the circle over the surface of the earth, in km.

c. Assume that all 52 beams from the satellite antennas have equal beamwidths and beams touch at their −3 dB contours. Nine beams can be fitted across the diameter of the circle you calculated in part (b). What is the beamwidth of an individual beam? How wide is the beam directly below the satellite, in km? Calculate the gain of the beam at the transmitting and at the receiving frequency.

d. A phased array with a square outline is used to generate the 52 beams of the transmitting antenna. The element spacing is 0.6 wavelengths. Estimate the number of elements that are required in the phased array.

3.10 The earth is 146.9 million km from the sun, and receives light with an intensity of 1.36 kW/m^2. Mars is 227.9 million km from the sun.

a. Calculate the intensity of sunlight for a satellite that is in the vicinity of Mars.

b. If the satellite has solar cells with an efficiency of 29%, what area is required to generate 1 kW of electrical power?

References

AIAA Journal (2018). *Journal of the American Institute of Aeronautics, 1963–2018*. New York, NY: American Institute of Aeronautics.

Astra 3B (2018). https://www.ses.com/our-coverage/satellites/337 (accessed 12 July 2018).

Boeing 376 (n.d.). https://www.boeing.com/history/products/376-satellite.page (accessed 10 July 2018).

Effective Space Solutions (2017). https://www.space.com/37205-satellite-service-repair-spacecraft-2020.htm (accessed 31 July 2018).

ETSI EN 302 307 V1.2.1 (2009). https://www.etsi.org/deliver/etsi_en/302300_302399/302307/01.02.01_60/en_302307v010201p.pdf (accessed 12 July 2018).

Fortesque, P., Swinerd, G., Stark, J. et al. (eds.) (2011). *Spacecraft Systems Engineering: Fourth Edition*. Hoboken NJ: Wiley.

GEO Satellite Positioning with GPS (2005). https://pdfs.semanticscholar.org/0c55/df19199794c78447d531a103014df77b3771.pdf (accessed 21 July 2018).

Global Express (2013). http://www.inmarsat.com/wp-content/uploads/2013/10/Inmarsat_APC_2013_05_George_Nicola.pdf (accessed 13 July 2018).

Globalstar (2017). http://www.globalstar.com/en/index.php?cid=8600. (accessed 12 February 2018).
Hydrazine and Schmidt, E.W. (1984). *Hydrazine and Its Derivatives*. New York: Wiley.
IEEE Trans AP-S (2018). *IEEE Transactions on Aerospace and Electronic Systems*, AS-1 through AS-54, 1963–2018. Piscataway, NJ: Institute of Electronic and Electrical Engineers.
Inflatable Antennas (1999). Inflatable structures taking to flight, *Aviation Week*, January 25, 1999, pp 60–62.
Intelsat 603 (2016). https://en.wikipedia.org/wiki/Intelsat_VI (accessed 22 July 2018).
Ion Thruster (2004). https://www.nasa.gov/centers/glenn/pdf/105819main_FS-2004-11-021.pdf (accessed 11 July 2011).
Ion Thruster (2008). https://www.aps.org/units/dfd/meetings/upload/Gallimore_APSDFD08.pdf (accessed 11 July 2011).
Ion Thruster (2018). https://en.wikipedia.org/wiki/Ion_thruster (accessed 11 July 2011).
Iridium Next (2018). https://www.iridiumnext.com (accessed 4 September 2018).
Maral, G. and Bousquet, M. (2002). *Satellite Communication Systems*, 4e. Chichester, UK: Wiley.
Reaction Wheel (2018). https://en.wikipedia.org/wiki/Reaction_wheel (accessed 10 July 2018).
Rudge, A.W., Milne, K., Olver, A.D., and Knight, P. (1983). *Handbook of Antenna Design*, IEE Electromagnetic Wave Series No. 15, vol. 1, Stevenage, Herts, UK.
Saft (2017). https://www.saftbatteries.com/press-releases/200th-satellite-equipped-saft-lithium-ion-batteries-set-launch (accessed 11 July 2011).
Satellites on Stamps (2012). https://commons.wikimedia.org/wiki/Category:Satellites_on_stamps (accessed 11 July 2011).
Silver, S. (ed.) (1989). *Microwave Antenna Theory and Design*. London, UK: IET, Reprint of Volume 12 of the Radiation Lab series published in 1949.
Solar Panels (2018). https://en.wikipedia.org/wiki/Solar_panels_on_spacecraft (accessed 11 July 2011).
Space Exploration Stamps (2015). https://commons.wikimedia.org/wiki/Category:Space_exploration_on_stamps (accessed 10 July 2018).
Space Revenue (2016). www.spaceindustry.com.au/Documents/Paper%20FINAL-5.pdf (accessed 11 July 2011).
Space Station Solar Arrays (2017). https://www.nasa.gov/mission_pages/station/structure/elements/solar_arrays-about.html (accessed 11 July 2011).
Stutzman, W.L. and G. A. Thiele. (2013). *Antenna Theory and Design*, 3rd edition, Hoboken, NJ, Wiley.
TDRS Fleet (2017). https://www.nasa.gov/directorates/heo/scan/services/networks/tdrs_fleet (accessed 18 July 2018).
V-band (2016). https://en.wikipedia.org/wiki/V_band#Satellite_constellations (accessed 12 July 2018).
Viasat (2017). https://www.viasat.com/about (accessed 22 February 2018).
Wild Blue 1 (2016). https://www.satbeams.com/satellites?norad=29643 (accessed 12 July 2018).

4

Satellite Link Design

The design of a satellite communication system is a complex process requiring compromises between many factors to achieve the best performance at an acceptable cost. We will first consider geostationary satellite systems, since geostationary earth orbit (GEO) satellites currently carry the majority of the world's satellite traffic.

4.1 Introduction

Launching satellites into orbit is a costly proposition. The cost to build and launch a large GEO satellite in the early days was about US$25 000 per kg. By 2018 the development of new launch vehicles by Space X® had lowered the cost of launching a satellite into geostationary transfer orbit using the Falcon 9 rocket to around US$7500 per kg for satellites up to 8300 kg (de Selding 2016). For heavier satellites, the cost of a Delta 4 launch was US$16,400 per kg in 2014 (Kanipe 2014). The cost to construct a large GEO satellite is estimated to be in the range US$50–US$400M depending on its mass and complexity, so a 3000 kg GEO satellite built for US$50M and launched at a cost of US$7500 per kg would cost a total of US$73M. However, system cost is much higher because a spare in-orbit satellite is needed for redundancy, and a network of earth stations, and a control center must be built. Low earth orbit (LEO) communication satellites are generally smaller than geostationary satellites, with mass in the 300–1500 kg range. Launch costs to LEO are approximately one third of the cost to GEO, so a single LEO satellite could be put into orbit for US$25M. Individual LEO satellites do not provide continuous communication and must be launched as a constellation of at least 24 satellites, making system cost well over US$1B.

Current proposals (2018) exist for the launching of thousands of LEO satellites to provide worldwide internet access using Ku-, Ka-, and V-band frequencies. The target price for these satellites is US$0.5M each, with a launch cost of US$0.5M. One proposal calls for several constellations totaling 12 000 satellites with a projected cost of US$12B (SpaceX 2016).

Weight, or more specifically mass is the most critical factor in the design of any satellite, since the heavier the satellite the higher the cost to build and launch it, and the capital cost of the satellite and launch must be recovered over its lifetime by selling services from orbit that include communications, navigation, ground mapping, and surveillance. The overall dimensions of the satellite are critical because the spacecraft must fit within the confines of the launch vehicle. When stowed for launch, the diameter of a

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

commercial spacecraft must be less than the space available within the launch vehicle shroud. Most large GEO satellites use deployable solar panels and antennas, but the antenna reflectors require accurate surfaces, and while they are hinged back against the spacecraft body for launch, they cannot be collapsed like a mesh or inflatable antenna. This limits the maximum aperture dimension to about 3.5 m, although some of the largest launch vehicles can accommodate spacecraft up to 5 m. As in most radio systems, antennas are a limiting factor in the capacity and performance of the communication system.

The mass of a GEO satellite is driven principally by two factors: the number and output power of the transponders on the satellite and the weight of station-keeping fuel. Reflector and phased array antennas can also be bulky and heavy. As much as half the total weight of satellites intended to remain in service for 15 years may be fuel when an apogee kick motor is used for injection into geostationary orbit and gas jets are used for station keeping over the lifetime of the satellite. Satellites with electric propulsion systems offer a significant weight advantage and are now widely used in GEO. High power transponders require lots of electrical power, which can only be generated by solar cells. Increasing the total output power of the transponders raises the demand for electrical power, the dimensions of the solar cells, and the weight of batteries that must be provided to maintain operation during eclipses, all adding more weight to the satellite.

The information carrying capacity of any radio communication link is determined by the RF power at the receiver input. Large antennas are needed to receive weak signals, and the signals from satellites in geostationary orbit are invariably weak. Early satellites were small and light, carrying small antennas and transponders with low output powers, which resulted in very weak signals at the earth's surface. The earth stations required for communication with these satellites were large and expensive, with 25 m to 30 m diameter antennas and high power transmitters. The trend over the 50 years that GEO satellites have been in operation has been toward larger satellites with high output powers and larger antennas leading to the ability to use smaller earth stations, exemplified by very small aperture terminal (VSAT) networks and direct broadcast satellite television (DBS-TV) receiving terminals. Direct to home satellite television (DTH-TV) broadcasting requires millions of receiving terminals, so these are made small and low cost, with the result that the DBS-TV satellites are large and expensive. The designer of a satellite communication system must work to minimize the capital cost of the entire system and must also ensure that sufficient revenue can be earned from the system to recover the large capital cost of building and launching satellites.

Three other factors influence system design: the choice of frequency band, atmospheric propagation effects, and multiple access technique. These factors are all related, with the frequency band often being determined by what is permitted for the specific service proposed. Tables 4.1 and 4.2 tabulate the most important frequencies allocated for satellite communications. The major communication bands are the 6/4, 14/11, and 30/20 GHz bands. (The uplink frequency is quoted first, by convention.) However, over much of the geostationary orbit there is already a satellite using both 6/4 and 14/11 GHz every two degrees of longitude. This is the minimum spacing used for satellites in GEO to avoid interference from uplink earth stations. Additional satellites can only be accommodated if they use another frequency band, such as 30/20 GHz.

Table 4.1 Major frequency allocations for fixed, mobile, and broadcast satellites

Fixed satellite service (FSS) GHz		Mobile satellite service GHz		Broadcast satellite service (BSS) GHz	
Uplink	**Downlink**	**Uplink**	**Downlink**	**Uplink**	**Downlink**
			312–315 MHz		
			387–390 MHz		
		455–460 MHz			
					1.452–1.492 (DAB)
			1.518–1.530		
			1.535–1.559		
		1.610–1.675			
		1.980–2.100			
			2.120–2.170		
					2.320–2.345 (DAB)
					2.250–2.535
			2.485–2.500		
					2.655–2.670
	2.670–2.690				
	3.400–3.500				
	3.600–4.200				
	4.500–4.800				
5.725–5.850					
5.850–7.075					
7.250–7.750					
7.900–8.400					
	10.7–11.7				11.7–12.7
	11.7–12.2 (II)				
12.75–13.25					
13.75–14.8					
				17.3–17.8	
17.8–18.1 (II)					
18.11–21.2					
	17.8–20.2				
	21.4–22.0 (I, III)				
24.75–25.25					
27.50–31.0					
		29.5–29.9			
	38.0–42.0				
42.5–43.5					
		43.5–47.0			
47.2–50.2					
50.4–51.4					
	71.0–76.0				
81.0–86.0					

Table 4.2 Major frequency bands for inter-satellite links (ISLs) and navigational satellites

	Navigation satellites	
ISLs GHz	**Uplinks GHz**	**Downlinks GHz**
		399.9–400.05 MHz
		1.164–1.215
		1.212–1.240
	1.240–1.300	
		1.559–1.610
	1.610–1.626	
		2.483–2.500
	5.000–5.010	
		5.010–5.030
	14.3–14.4	
22.55–23.55		
24.65–24.75	24.65–24.75	
25.25–25.5		
25.5–27.0		
32.3–33.0		
	43.5–47.0	
54.2–58.2		
59.0–71.0	66.0–71.0	
	95.0–100.0	

Rain in the atmosphere attenuates radio signals. The effect is more severe as the frequency increases, with little attenuation at 4 and 6 GHz, but significant attenuation above 10 GHz. Attenuation through rain (in decibels) increases roughly as the square of frequency, so a satellite uplink operating at 30 GHz suffers about four times as much attenuation in rain as an uplink at 14 GHz.

Low earth orbit and medium earth orbit (MEO) satellite systems have similar constraints to GEO satellite systems, but for continuous coverage of earth require more satellites that each serve a smaller area of the earth's surface. Although the satellites are much closer to the earth than GEO satellites and therefore produce stronger signals on the earth's surface, this advantage is lost in mobile systems when the earth terminals have low gain omnidirectional antennas. When the earth terminal is fixed, as in an internet access system, a phased array antenna that has electronic beam steering to track LEO or MEO satellites can be used; however, the cost of this type of earth station is much higher than with a conventional fixed reflector antenna used with a GEO satellite. Communication capacity of a radio link, in terms of megabits per second (Mbps) is directly proportional to antenna gain at each end of the link. The omnidirectional antenna of a handheld mobile terminal working with a LEO satellite system has a nominal gain of 0 dB. A fixed phased array antenna tracking the same satellites can have a gain of 30 dB, which can allow an increase in the bit rate of the link by a factor of one thousand. As an example, mobile terminals for the Iridium LEO satellites using L-band can receive data at 100 kbps. Fixed earth terminals working with proposed Ka-band LEO satellites can receive data at 360 Mbps. LEO and MEO satellites use multiple spot beam antennas to

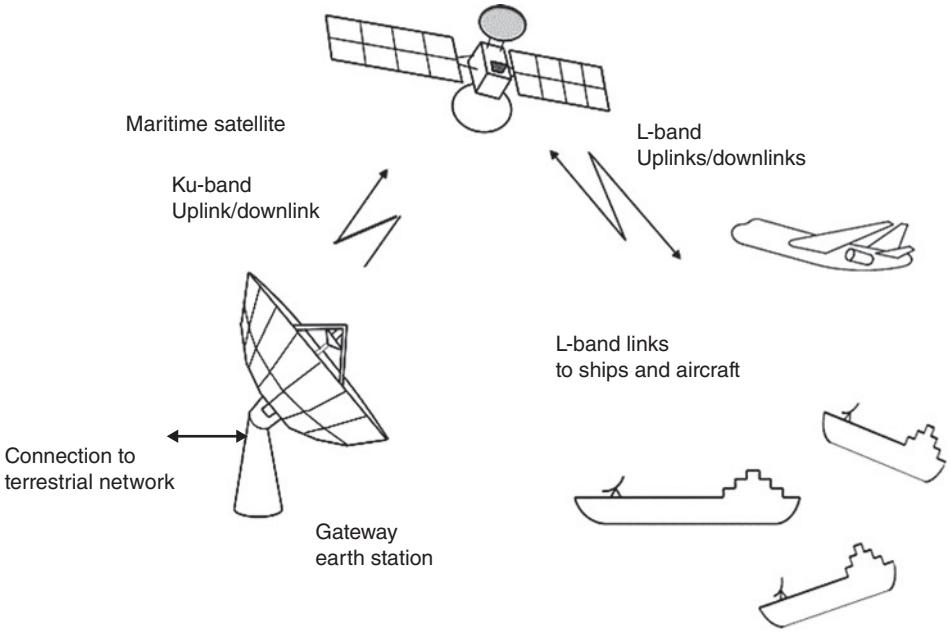


Figure 4.1 Illustration of a maritime satellite system using a GEO satellite. Ships are equipped with a steerable, gyro stabilized antennas. Aircraft have phased array antennas.

increase the gain of the satellite antenna beams relative to regional coverage beams, and also to provide frequency reuse.

Mobile satellite terminals typically operate with low gain antennas at the mobile unit, and at as low an RF frequency as can be obtained. The link between the satellite and the major earth station (often called a *hub* or *gateway* station) is usually in a different frequency band, as it is a fixed link. Figure 4.1 shows an illustration of a maritime satellite communication system using a GEO satellite and L-band links to mobiles, with Ku-band links to a fixed gateway station.

All communication links are designed to meet certain performance objectives, usually a *bit error rate* (BER, the probability that a received bit is in error) in a digital link or a *signal to noise ratio* (SNR) where the signal is audio or video, measured in the baseband channel. The baseband channel is where an information carrying signal is generated or received; for example, a TV camera generates a baseband video signal and a TV receiver delivers a baseband video signal (in digital form) to the screen to form the images that the viewer watches. Digital data are generated by computers at baseband, and BER is measured at baseband. The baseband channel BER or SNR is determined by the *carrier to noise ratio* (CNR) at the input to the demodulator in the receiver. In most satellite communications applications, the CNR at the demodulator input must be greater than 0 dB for the BER or SNR objective to be achieved. Typically, digital links operating at CNRs below 11 dB must use error correction techniques to improve the BER delivered to the user.

The CNR is calculated at the input of the receiver, at the output terminals (or *output port*) of the receiving antenna. RF noise received along with the signal, and noise generated by the receiver are combined into an equivalent noise power at the input to

the receiver, and a noiseless receiver model is used. In a noiseless receiver, the CNR is constant at all points in the RF and intermediate frequency (IF) chain, so the CNR at the demodulator is equal to the CNR at the receiver input. In a satellite link there are two signal paths: an *uplink* from the earth station to the satellite, and a *downlink* from the satellite to the earth station. The *overall* CNR at the earth station receiver depends on both links, and both must therefore achieve the required performance for a specified percentage of time. Path attenuation in the earth's atmosphere may become excessive in heavy rain, causing the CNR to fall below the minimum permitted value, especially when the 30/20 GHz Ka-band or higher frequency band is used, leading to a link *outage*. Satellite links are designed to meet specific performance objectives under *clear sky* conditions, for example, a BER no higher than 10^{-8}. (The terms *clear sky* and *clear air* are both used here to describe a path through the atmosphere that is free of clouds and rain.) When rain affects a link, the BER will increase as the CNR in the receiver falls until some maximum threshold is reached, for example, BER > 10^{-6}, at which the link is regarded as unusable and is considered to be in an *outage* state. The sum of all outages over a specified time period, typically a month or a year, determines the *availability* of the link. Chapter 7 discusses the concepts of availability and outages caused by propagation disturbances in more detail.

Designing a satellite system therefore requires knowledge of the required performance of the uplink and downlink, the propagation characteristics and rain attenuation for the frequency band being used at the earth station locations, and the parameters of the satellite and the earth stations. Additional constraints may be imposed by the need to conserve RF bandwidth and to avoid interference with other users. Sometimes, all of this information is not available and the designer must estimate values and produce tables of system performance based on assumed scenarios. It is usually impossible to design a complete satellite communication system at the first attempt. A trial design must first be generated, and then refined until a workable compromise is achieved. This chapter sets out the basic procedures for the design of satellite communication links, and includes design examples for a direct broadcast satellite television system using a GEO satellite, a Ku-band video distribution system, and a LEO satellite system for personal communication.

Table 4.1 shows the major frequency bands allocated to the fixed, mobile, and broadcasting satellite services for frequency bands up to 100 GHz. There are many additional frequency allocations, and also many restrictions on how the frequencies can be used. The International Telecommunications Union (ITU) determines how radio frequencies are to be used internationally through a series of World Radio Conferences (WRCs) and publishes tables of international frequency allocations.

Individual countries often have further restrictions. Some frequency bands are also divided between civil and government use. Most countries publish their own frequency allocations tables on the web; for example, the US Federal Communication Commission (FCC) publishes an online table of frequency allocations (FCC Online Table of Frequency Allocations 2017). Subsequent chapters that discuss specific applications for satellite communication systems have further details on the frequency bands allocated to those services.

The ITU divides up the earth's surface into three regions for the purpose of allocating radio frequencies. Regions I, II, and III are regions of the earth's surface defined in the ITU's Radio Regulations (2017). Region I covers Europe, Africa, and northern Asia. Region II covers North and South America, and Region III covers the remainder of

Asia. Where Table 4.1 has (I), (II), or (III) after a frequency band, this indicates that the specified frequencies are available only in those geographic regions. Broadcast satellite frequencies are for TV broadcasting, except for the L- and S-band frequencies marked digital audio broadcasting (DAB), which are used for radio broadcasting. For radio broadcast systems, L-band frequencies (1.452–1.492 GHz) are allocated to Africa and Asia, and S-band frequencies (2.320–2.345 GHz) are allocated to North America.

Table 4.2 shows the major frequencies used for inter-satellite links (ISLs) and navigational satellites. The Iridium LEO system of 66 satellites has ISLs between satellites in adjacent orbits and between adjacent satellites in the same orbit using the 22.55–23.55 GHz band. This band includes the resonance frequency of water vapor, which causes attenuation on space-earth paths. The 59.0–71.0 GHz band contains the resonance frequencies of oxygen, which cause severe attenuation in the earth's atmosphere, and is therefore useful for ISLs since there is no atmosphere in the region of space where satellites orbit. For more information on intersatellite frequency allocations see Recommendation ITU-R S1591 (2002).

Navigation satellites are included in the general category of *radiolocation*, which includes space based radars used for earth surveillance. To date, most Global Navigation Satellite Systems (GNNS) satellites such as global positioning system (GPS) and Galileo have used downlink frequencies in L-band between 1.16 and 1.61 GHz. Space based radars are used for both civil and military surveillance. Civil applications include meteorology, by mapping of cloud top heights, mapping of crops for agriculture, and mapping of oceans for sea ice and wave heights. Military applications include locating vehicles, ships, and aircraft, as well as buildings and construction sites.

4.2 Transmission Theory

The calculation of the power received by an earth station from a satellite transmitter is fundamental to the understanding of satellite communications. In this section, we discuss two approaches to this calculation: the use of *flux density* and the *link equation*.

Consider a transmitting source, in free space, radiating a total power P_t watts uniformly in all directions as shown in Figure 4.2. Such a source is called *isotropic*; it is an idealization that cannot be realized physically because it could not create transverse electromagnetic (EM) waves. At a distance R meters from the hypothetical isotropic source transmitting RF power P_t watts, the flux density crossing the surface of a sphere with radius R m is given by

$$F = \frac{P_r}{4\pi R^2} \text{ W/m}^2 \tag{4.1}$$

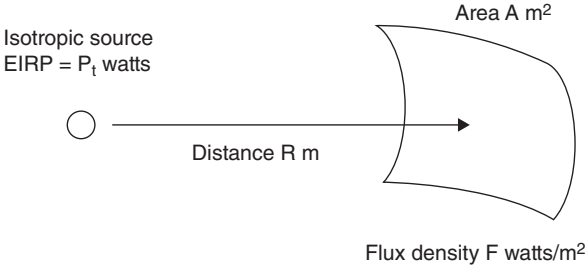


Figure 4.2 Calculation of flux density from an isotropic source with EIRP P_t watts. The flux density is measured over a 1 m^2 section of a sphere at a distance R meters from the source.

All real antennas are directional and radiate more power in some directions than in others. Any real antenna has a gain $G(\theta)$, defined as the ratio of power per unit solid angle radiated in a direction θ to the average power radiated per unit solid angle (Silver 1949, p. 2).

$$G(\theta) = \frac{P(\theta)}{P_o/4\pi} \text{ W/m}^2 \tag{4.2}$$

where

$P(\theta)$ is the power radiated per unit solid angle by the antenna
P_o is the total power radiated by the antenna
$G(\theta)$ is the gain of the antenna at an angle θ

The reference for the angle θ is usually taken to be the direction in which maximum power is radiated, called the *boresight* direction of the antenna or the *antenna electrical axis*. The gain of the antenna is then the value of $G(\theta)$ at angle $\theta = 0°$, and is a measure of the increase in flux density radiated by the antenna over that with an ideal isotropic antenna radiating the same total power. See Appendix B for more details of antennas and their properties.

For a transmitter with output P_t watts driving a lossless antenna with gain G_t, the flux density in the direction of the antenna boresight at distance R meters is

$$F = \frac{P_t G_t}{4\pi R^2} \text{ W/m}^2 \tag{4.3}$$

The product $P_t\ G_t$ is often called the *effective isotropically radiated power* (EIRP), and describes the combination of transmitter power and antenna gain in terms of an equivalent isotropic source with power $P_t\ G_t$ watts, radiating uniformly in all directions.

If we had an ideal receiving antenna with an aperture area of A m^2, as shown in Figure 4.3, we would collect power P_r watts given by

$$P_r = F \times A \text{ watts} \tag{4.4}$$

A practical antenna with a physical aperture area of A_r m^2 will not deliver the power given in Eq. (4.4). Some of the energy incident on the aperture is reflected away from the antenna, referred to as scattering, and some is absorbed by lossy components. This reduction in efficiency is described by using an *effective aperture* A_e where

$$A_e = \eta_A A_r \text{m}^2 \tag{4.5}$$

and η_A is the *aperture efficiency* of the antenna (Stutzman and Thiele 2013, p. 363). The aperture efficiency η_A accounts for all the losses between the incident wavefront and the antenna output port: these include *illumination efficiency* or *aperture taper efficiency* of

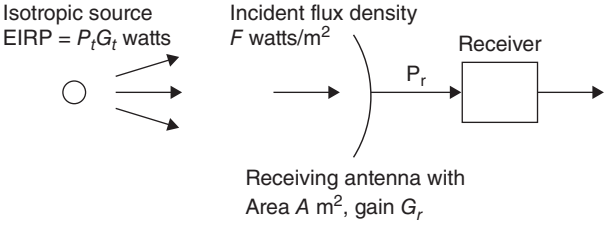


Figure 4.3 Calculation of received power by an antenna with gain G_r from a source with EIRP $P_t\ G_t$ watts. F is the flux density incident on the receiving antenna. P_r is the power delivered to the receiver.

the antenna, which is related to the energy distribution produced by the feed across the aperture, and also losses due to spillover, blockage, phase errors, diffraction effects, polarization, and mismatch losses. For parabolodial reflector antennas, η_A is typically in the range 50–75%, lower for small antennas and higher for large Cassegrain and Gregorian antennas. Horn antennas can have efficiencies approaching 80%. (See Appendix B for an explanation of the aperture efficiency of antennas.)

Thus the power received by a real antenna with a physical receiving area A_r and effective aperture area A_e m^2 at a distance R from the transmitter is

$$P_r = \frac{P_t G_t A_e}{4\pi R^2} \text{ watts} \tag{4.6}$$

Note that this equation is essentially independent of frequency if G_t and A_e are constant within a given band; the power received at an earth station depends only on the EIRP of the satellite, the effective area of the earth station antenna, and the distance R.

A fundamental relationship in antenna theory is that the gain and area of an antenna are related by (Stutzman and Thiele 2013, p. 363)

$$G = 4\pi A_e/\lambda^2 \tag{4.7}$$

where λ is the wavelength (in meters for A_e in square meters) at the frequency of operation.

Substituting for A_e in Eq. (4.6) gives

$$P_r = \frac{P_t G_t G_r}{(4\pi R/\lambda)^2} \text{ watts} \tag{4.8}$$

This expression is known as the *link equation*, and it is essential in the calculation of power received in any radio link. The frequency (as wavelength, λ) appears in this equation for received power because we have used the receiving antenna gain, instead of effective area. The term $(4\pi R/\lambda)^2$ is known as the *path loss*, L_p. It is not a loss in the sense of power being absorbed; it accounts for the way energy spreads out as an EM wave travels away from a transmitting source in three-dimensional (3-D) space.

Collecting the various factors, we can write

$$P_r = \frac{\text{EIRP} \times \text{Receiving antenna gain}}{\text{Path Loss}} \text{ watts} \tag{4.9}$$

In communication systems, decibel quantities are commonly used to simplify equations like (4.9). In decibel terms, we have

$$P_r = \text{EIRP} + G_r - L_p \text{ dBW} \tag{4.10}$$

where

$\text{EIRP} = 10\ log_{10}\ (P_t\ G_t)$ dBW
$G_r = 10\ log_{10}\ (4\pi A_e/\lambda^2)$ dB

Path loss L_p is given by

$$L_p = 10\ \log_{10}\ [(4\pi R/\lambda)^2] = 20\ \log_{10}\ (4\pi R/\lambda) \text{ dB} \tag{4.11}$$

If you are unfamiliar with decibels, read Appendix A, which discusses how decibels are used in the analysis of radio communication systems. Equation (4.10) represents an idealized case, in which there are no additional losses in the link. It describes

transmission between two ideal antennas in otherwise empty space. In practice, we will need to take account of a more complex situation in which we have losses in the atmosphere due to attenuation by oxygen, water vapor, and rain, losses in the antennas at each end of the link, and possible reduction in antenna gain due to mispointing. All of these factors are taken into account by the *system margin* but need to be calculated to ensure that the margin allowed is adequate. More generally, Eq. (4.10) can be written

$$P_r = \text{EIRP} + G_r - L_p - L_a - L_{ta} - L_{ra} \text{ dBW} \tag{4.12}$$

where

L_a = attenuation in the atmosphere
L_{ta} = losses associated with the transmitting antenna
L_{ra} = losses associated with the receiving antenna

The conditions in Eq. (4.12) are illustrated in Figure 4.4. The expression dBW means decibels greater or less than 1 W (0 dBW). The units dBW and dBm (dB greater or less than 1 W and 1 mW) are widely used in communications engineering. EIRP, being the product of transmitter power and antenna gain is normally quoted in dBW.

Note that once a value has been calculated in decibels, it can readily be scaled if one parameter is changed. For example, if we calculated G_r for an antenna to be 48 dB at a frequency of 4 GHz, and wanted to know the gain at 6 GHz, we can multiply G_r by $(6/4)^2$.

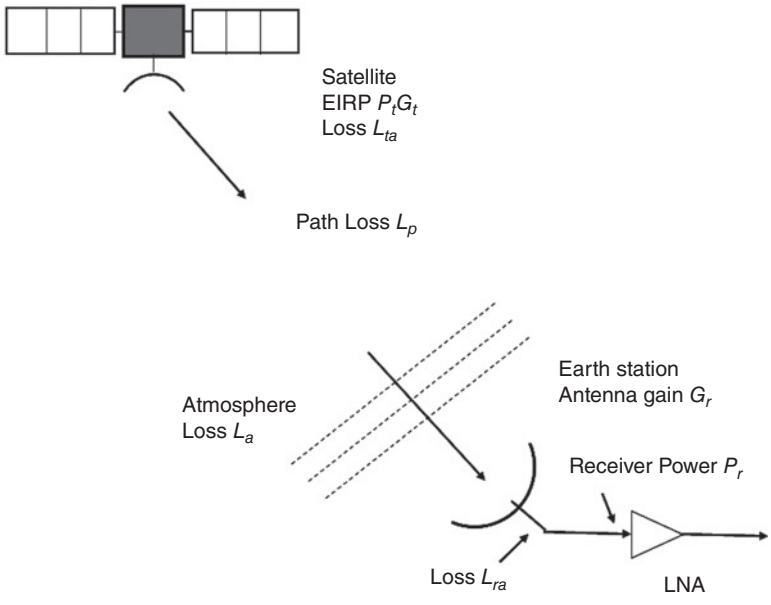


Figure 4.4 Calculation of received power from a satellite with EIRP P_t G_t watts including losses. Loss L_{ta} is an off-axis loss deducted from the satellite antenna on-axis gain when calculating the antenna gain in the direction of the receiving earth station. Atmospheric loss L_a includes clear air loss caused by gases in the atmosphere and any additional loss from clouds and rain. Receiving antenna losses L_{ra} include ohmic losses in the waveguide between the antenna feed and the LNA, and an off-axis loss if the receiving antenna does not point directly at the satellite.

Using decibels, we simply add 20 log (6/4) (or 20 log (3) − 20 log (2)) = 9.5 − 6 = 3.5 dB. Thus the gain of our antenna at 6 GHz is 51.3 dB.

Appendix A gives more information on the use of decibels in communications engineering.

Example 4.1

A satellite at a distance of 40 000 km from a point on the earth's surface radiates a power of 10 W from an antenna with a gain of 17 dB in the direction of the observer. Find the flux density at the receiving point, and the power received by an earth station antenna at this point with an effective area of 10 m^2.

Answer
Using Eq. (4.3)

$$F = P_t G_t/(4\pi R^2) = 10 \times 50/[4\pi \times (4 \times 10^7)^2] = 2.49 \times 10^{-14}$$

The power received with an effective collecting area of 10 m^2 is therefore

$$P_r = 2.49 \times 10^{-13} \text{ W}$$

The calculation is more easily handled using decibels. Noting that 10 $\log_{10}$ $4\pi \approx 11.0$ dB

$$\begin{aligned} F \text{ in dB units} &= 10 \log_{10}(P_t G_t) - 20 \log_{10}(R) - 11.0 \\ &= 27.0 - 152.0 - 11.0 \\ &= -136.0 \text{ dBW/m}^2 \end{aligned}$$

Then

$$P_r = F \text{ dBW/m}^2 + A_e \text{ dBm}^2$$
$$P_r = -136.0 + 10.0 = -126 \text{ dBW or } -96 \text{ dBm}$$

Here we have put the antenna effective area into decibels greater than 1 m^2 (10 m^2 = 10 dB greater than 1 m^2) and also given the answer in dBW and dBm, decibels above 1 watt and 1 milliwatt.

Example 4.2

The satellite in Example 4.1 operates at a frequency of 11 GHz. The receiving antenna has a gain of 52.3 dB. Find the received power at the earth station in dBW and dBm. It is common practice to quote transmit power in dBW and received power in dBm.

Answer
Using Eq. (4.10) and working in decibels

$$P_r = \text{EIRP} + G_r - \text{path loss dBW}$$

$$\text{EIRP} = 27.0 \text{ dBW}$$

$$G_r = 52.3 \text{ dB}$$

$$\text{Path loss } L_p = (4\pi R/\lambda)^2 = 20 \log_{10}(4\pi R/\lambda) \text{ dB}$$

$$= 20 \log_{10}[(4\pi \times 4 \times 10^7)/(2.727 \times 10^{-2})] = 205.3 \text{ dB}$$

$$P_r = 27.0 + 52.3 - 205.3 = -126.0 \text{ dBW}$$

The received power in dBm units is numerically 30 dB greater than in dBW.

Hence

$$P_r = 126.0 + 30 = -96.0 \text{ dBm}$$

We have the same answer as in Example 4.1 because the figure of 52.3 dB is the gain of a 10 m^2 aperture at a frequency of 11 GHz.

Equation (4.12) is commonly used for calculation of received power in a microwave link and is set out as a link power budget in tabular form using decibels. This allows the system designer to adjust parameters such as transmitter power or antenna gain and quickly recalculate the received power.

The received power, P_r calculated by Eqs. (4.6) and (4.8) is commonly referred to as carrier power, C. This is because satellite links typically use phase modulation for digital transmission where the amplitude of the carrier is not changed when the data is modulated onto the carrier, so received carrier power C watts is always equal to received power P_r watts.

4.3 System Noise Temperature and G/T Ratio

Noise temperature is a useful concept in communications receivers, since it provides a way of determining how much thermal noise is generated by active and passive devices in the receiving system. At microwave frequencies, a black body with a physical temperature, T_p degrees kelvin generates electrical noise over a wide bandwidth.

The noise power is given by (Rappaport 2002, p. 612)

$$P_n = kT_pB_n \text{ watts} \tag{4.13}$$

where

k = Boltzmann's constant = 1.39 x 10^{-23} J/K = $-$ 228.6 dBW/K/Hz
T_p = physical temperature of source in kelvin degrees
B_n = noise bandwidth in which the noise power is measured, in hertz

P_n is the available noise power (in watts) and will be delivered only to a load that is impedance matched to the noise source. The term $k\,T_p$ is a noise power spectral density, in watts per hertz. The density is constant for all radio frequencies up to 300 GHz, but we need a way to describe the noise produced by the components of a low noise receiver. This can conveniently be done by equating the component to a black body radiator with an equivalent noise temperature T_n kelvins. A device with a noise temperature of T_n kelvins (symbol K, not °K) produces at its output the same noise power as a black body at a physical temperature T_n degrees kelvin followed by a noiseless amplifier with the same gain as the actual device. The description of a low noise component by an equivalent noise source at the input of a noiseless amplifier is very useful because we can add noise temperatures to determine the total noise power in a receiver, as shown in the following analysis. Note that the unit of noise temperature is kelvins, not degrees kelvin, a distinction sometimes lost on the suppliers of consumer satellite broadcast receiving equipment.

In satellite communication systems we are always working with very weak signals (because of the large distances involved) and must make the noise level as low as possible to meet the CNR requirements. This is done by making the bandwidth in the receiver,

usually set by the IF amplifier stages, to be just large enough to allow the signal (carrier and sidebands) to pass unrestricted, while keeping the noise power to the lowest value possible. The bandwidth used in Eq. (4.12) should be the equivalent noise bandwidth. Frequently we do not know the equivalent noise bandwidth and use the 3 dB bandwidth of our receiving system instead. The error introduced by using the 3 dB bandwidth is small when the filter characteristic of the receiver has steep sides, as is always the case in radio communication systems to avoid interference problems.

Amplifier noise temperatures from 25 to 250 K can be achieved without physical cooling for receivers in the frequency bands up to Ka-band when GaAsFET amplifiers are employed. GaAsFET amplifiers can be built to operate at room temperature with typical noise temperatures of 25 K at 4 GHz and 65 K at 11 GHz. Noise temperature increases with frequency, and a low noise amplifier (LNA) for a 20 GHz receiver might have a noise temperature of 100 K. One might ask how an amplifier can have a noise temperature that is lower than its physical temperature. Noise temperature simply relates the noise produced by an amplifier to the thermal noise from a matched load at the same physical temperature placed at the input to the amplifier. If the amplifier produced no noise at all, its noise temperature would be 0 K. If the amplifier produces less noise than a matched load at the same physical temperature, its noise temperature will be lower than its physical temperature.

To determine the performance of a receiving system we need to be able to find the total thermal noise power against which the signal must be demodulated. We do this by determining the *system noise temperature*, T_s. T_s is the noise temperature of a noise source located at the input of a noiseless receiver, which gives the same noise power as the original receiver, measured at the output of the receiver, and usually includes noise from the antenna and the atmosphere.

If the overall end-to-end gain of the receiver is G_{rx} (G_{rx} is a ratio, not in decibels) and its narrowest bandwidth is B_n Hz, the noise power at the demodulator input is

$$P_{no} = kT_sB_nG_{rx} \text{ watts} \tag{4.14a}$$

where G_{rx} is the gain of the receiver from RF input to demodulator input.

The noise power referred to the input of the receiver is P_n where

$$P_n = kT_sB_n \text{ watts} \tag{4.14b}$$

Let the antenna deliver a signal power P_r watts to the receiver RF input. The signal power at the demodulator input is $P_r\ G_{rx}$ watts, representing the power contained in the carrier and sidebands after amplification and frequency conversion within the receiver. Hence, the carrier to noise ratio (CNR) at the demodulator is given by

$$\frac{C}{N} = \frac{P_rG_{rx}}{kT_sB_nG_{rx}} = \frac{P_r}{kT_sB_n} \tag{4.15}$$

The gain of the receiver cancels out in Eq. (4.15), so we can calculate CNRs for our receiving stations at the antenna output port. This is convenient, because a link budget will find P_r at this point. Using a single parameter to encompass all of the sources of noise in a receiving terminal is very useful because it replaces several sources of noise in the receiver by a single system noise temperature, T_s.

4.3.1 Earth Station Receivers

Figure 4.5a shows a simplified communications receiver with an RF amplifier and single frequency conversion, from its RF input to the IF output. This is the form that has been used for most radio receivers before the introduction of the digital radios, known as the *superhet* (short for *super heterodyne*). The superhet receiver has three main subsystems: a *front end* (RF amplifier, *mixer* and *local oscillator* [LO]) an *IF amplifier* (IF amplifiers and filters), a *demodulator,* and a baseband section.

The band pass filter (BPF) that follows the LNA is called an *image rejection* filter; its purpose is to block noise in the frequency band $f_{lo} - f_{if}$ from entering the mixer. When low side injection of the local oscillator is used, the image band is at $f_r - 2f_{if}$. Any mixer and local oscillator form a frequency conversion stage that can down convert two input signals, one at $f_{lo} + f_{if}$, a signal that is one IF distance above the local oscillator frequency, and a second signal at $f_{lo} - f_{if}$, a signal that is one IF distance below the local oscillator called an *image*. The image rejection filter that follows the LNA blocks any signals and noise in the band $f_{lo} - f_{if}$ from reaching the mixer. Noise is the main concern in a satellite communications receiver; if the image rejection filter is omitted and a second band of noise reaches the mixer, the noise power at the mixer output is doubled and the CNR falls by 3 dB. The image rejection filter is placed after the LNA to avoid any loss that the filter would introduce if placed ahead of the LNA, as this would increase the system noise temperature of the receiver. Loss after the LNA simply reduces the gain of the LNA and filter stage slightly without increasing the receiver system noise temperature. When multiple signals are being received, for example in a frequency division multiple

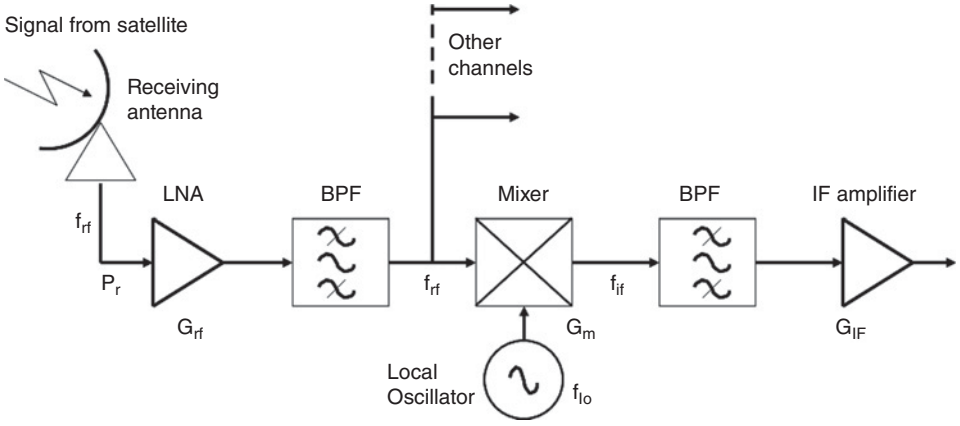


Figure 4.5a Simplified receiver with single frequency conversion.
The received signal at frequency f_{rf} is first amplified by the LNA, then selected by the first BPF, which acts as an image rejection filter blocking noise in a band around frequency $f_{lo} + f_{if}$. The mixer multiplies the received signal by the local oscillator signal to yield two new frequencies $f_{rf} + f_{lo}$ and $f_{rf} - f_{lo}$; the second bandpass filter selects the lower sideband signal at frequency $f_{rf} - f_{lo}$. The IF amplifier amplifies the IF signal to a level that allows it to be sent over a coaxial cable to the indoor unit. The components shown in Figure 4.5a comprise the outdoor unit in a typical DBS-TV or VSAT receiving system.
LNA: Low noise amplifier; BPF: Band pass filter; IF: Intermediate frequency; P_r: Received power at antenna output; G_{rf}: Gain of LNA; G_m: Gain of mixer; G_{if}: Gain of IF amplifier. Frequency f_{rf} is the RF frequency of the received signal, f_{lo} is the local oscillator frequency, and f_{if} is the intermediate amplifier frequency.

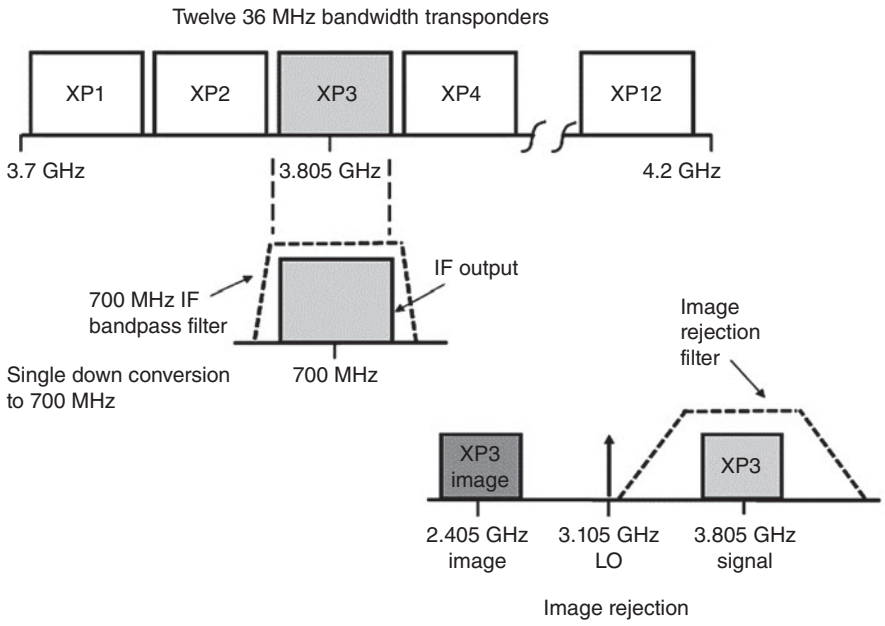


Figure 4.5b Frequency plan for single conversion C-band receiver.

access (FDMA) receiver, the first IF frequency is set high enough that none of the signals appears at the mixer output as an image frequency.

Figure 4.5b shows how the receiver in Figure 4.5a can select the signals from transponder #3, centered at a frequency of 3805 MHz, of a 12 transponder C-band satellite and down convert the transponder output to an IF frequency of 700 MHz using a local oscillator set to 3105 MHz. Parallel channels with different local oscillator frequencies convert the other 11 transponder frequencies to the common IF frequency of 700 MHz. The use of a common IF frequency allows all the blocks in the 12 parallel channels that follow the IF portion of the receiver to be identical. Also shown is the image frequency for the wanted channel XP3.

The RF amplifier in a satellite communications receiver must generate as little noise as possible, so it is called a *low noise amplifier* (LNA). The mixer and local oscillator form a frequency conversion stage that down converts the RF signal to a fixed IF, where the signal can be amplified and filtered accurately. A unit that combines the LNA and down converter is known as a *low noise block converter* (LNB), or sometimes an LNC.

Many earth station receivers use the *double superhet configuration* shown in Figure 4.6a, which has two stages of frequency conversion. The front end of the receiver, the *outdoor unit* is mounted behind the antenna feed and converts the incoming RF signals to a first IF, typically in the range 900–2500 MHz. This allows the receiver to accept all the signals transmitted from a satellite in a 500 MHz bandwidth at C-band or Ku-band, for example. The RF amplifier has a high gain and the mixer is followed by a stage of IF amplification. The 900–2500 MHz IF signal is sent over a coaxial cable to a set-top receiver, the *indoor unit* that contains another down converter and a tunable local oscillator. The local oscillator is tuned to convert the incoming signal from a selected transponder to a second IF frequency. The second IF amplifier has a bandwidth

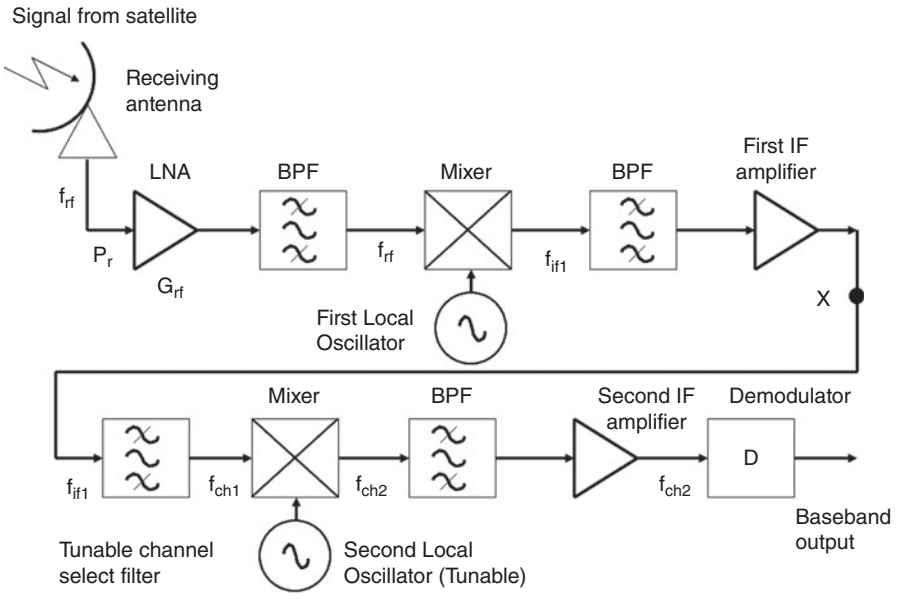


Figure 4.6a Double conversion superhet receiver using the same principle as the single conversion receiver in Figure 4.5a, but with two down conversions of the signal to a second intermediate frequency f_{if2}. The second intermediate frequency f_{if2} may be $f_{if1} - f_{lo2}$ or $f_{if1} + f_{lo2}$.
LNA: Low noise amplifier; BPF: Band pass filter; IF: Intermediate frequency; D: Demodulator. Additional channels connect at point X to extract the signals from other transponders with tunable channel selection filters and different second local oscillator frequencies.

matched to the spectrum of the wanted signal. Simple direct broadcast satellite TV receivers at Ku-band use this approach, with a typical second IF filter bandwidth of 20 MHz. More complex receivers designed to receive multiple signals may have several IF amplifiers at different frequencies in the range 900–5000 MHz. The double frequency conversion configuration is also used in Ku- and Ka-band transponders.

Figure 4.6b shows an example of the frequency plan for a double conversion receiver, which receives signals from transponder #3 of a Ku-band satellite at a frequency of 11.805 GHz. The first IF stage in the receiver uses a local oscillator at 10.8 GHz to down convert all 12 transponders to the first IF band 900–1400 MHz. A tunable bandpass filter with a bandwidth of 36 MHz is set to a center frequency of 1005 MHz to select the signals from transponder #3, and the second local oscillator at 935 MHz down converts this signal to a common second IF frequency of 70 MHz. Signals from the other 11 transponders are selected by identical second IF stages connected at point X in Figure 4.6a using tunable bandpass filters set to the first IF frequencies of each signal, and different local oscillator settings.

As discussed in Chapter 5, digital signal processing (DSP) can be employed to replace many of the blocks shown in Figures 4.5a and 4.6a. Hardware filters require inductors, which are physically large devices consisting of coils of wire, often wound on a core of magnetic material such as ferrite. It is not possible to include inductors in integrated circuits, so small devices such as cellular phones and GPS receivers use digital filtering instead. The design approach is known generically as a digital radio, in which many

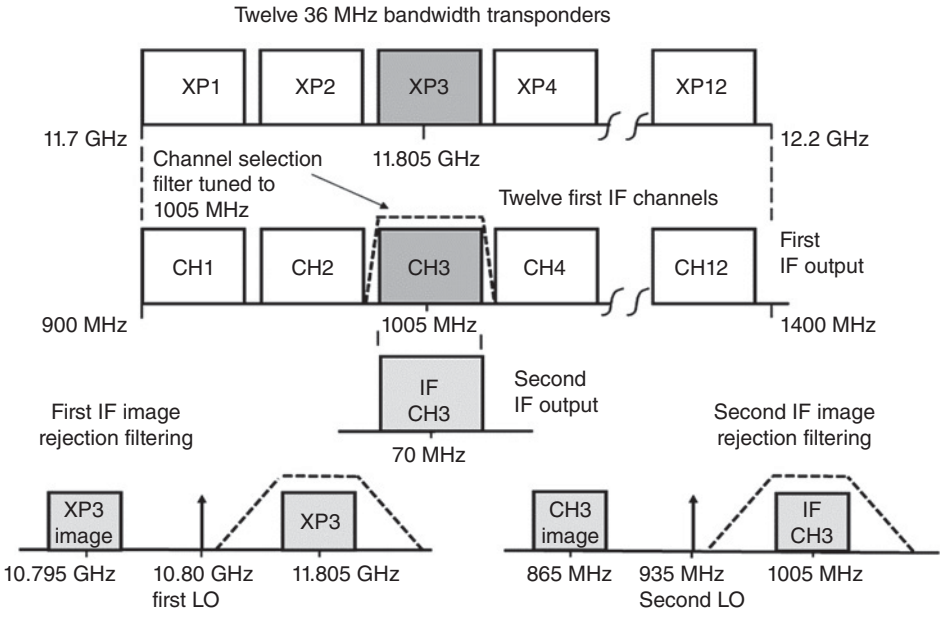


Figure 4.6b Frequency plan for a double conversion Ku-band receiver. The entire 500 MHz band of signals received from the satellite is down converted to the first IF frequency covering 900–1400 MHz, using a first local oscillator at 10.80 GHz. Channel XP3 is selected with a bandpass filter centered at 1005 MHz. The 1005 MHz signal is down converted to the second IF at 70 MHz using a local oscillator at 935 MHz. The bandpass filters centered at 11.95 GHz and 1005 MHz are image rejection filters, as illustrated at the bottom of the figure.

blocks of the hardware receivers shown in Figures 4.5a, 4.5b, 4.6a, and 4.6b are implemented by digital signal processing (DSP), either in application specific integrated circuits (ASICs), field programmable gate arrays (FPGAs), or by a microprocessor. Fast analog to digital converters can be used to digitize the RF signal directly after the LNA in a GPS or mobile receiver operating in L-band, or after the first IF amplifier in Ku- and Ka-band receivers employing a single down conversion to an L-band IF. Similar techniques can be used to create digital transmitters, allowing hand held devices such as cellular phones to fit in a pocket. Many digital radio receivers employ software control of some of their internal functions, such as filter center frequency and bandwidth, and local oscillator frequency, as well as the digital signal processing applied to the baseband signals. These are known as *software radios* and are increasing being used in satellite communication systems.

In general, it is difficult to make good narrowband filters with a ratio of bandwidth to center frequency less than 1%. The inverse of this ratio is the Q factor of the filter; it is easier to build hardware filters with Q factors of 50 or lower, and this also applies to digital filters. If we want a 36 MHz bandwidth filter for a signal received from a satellite at 4 GHz, it is difficult to implement the filter at the RF frequency of 4 GHz, where the filter bandwidth is less than 1% of the center frequency and Q exceeds 100. We would instead down convert the 4 GHz signal to an IF around 700 MHz, as in the example in

Figure 4.5b, where the 36 MHz bandwidth is 8.6% of the IF frequency and the Q of the required filter is 11.6. This is the advantage of the superhet receiver design: very accurate filters can be used by converting the signal to a convenient IF. With Ku-band and Ka-band satellites, a double conversion superhet receiver is used with two down conversion stages, as illustrated in Figure 4.6a. The Q factors of the bandpass filters in Figure 4.6a are 9.4 for the first IF and 14.4 for the second IF.

The down conversion in the front end is achieved by multiplying the received signal and the local oscillator frequency in a non-linear device – a mixer. Multiplication of two RF signals creates products at their sum and difference frequencies; the frequency of the local oscillator (LO) is usually set to $f_{signal} - f_{if}$. This is called *low side injection*. The receiver could also receive another RF signal at a frequency $f_{lo} - f_{if}$, which would produce an output from the mixer at f_{if}. This is called an *image frequency*. It is blocked by a bandpass filter in the RF amplifier that is wide enough to pass the wanted range of signal frequencies but has high attenuation for the image frequency.

One further advantage of the superhet receiver design is that tuning of the receiver can be done with the local oscillator. The IF is at a fixed frequency, and the local oscillator frequency is varied to select the wanted signal. The front end is followed by an IF amplifier stage, which contains bandpass filters that exactly match the spectrum of the received signal. In many small earth stations, known as VSATs, the LNA, LO, and first IF amplifier and filters are all included in a single package called a *low noise block converter* (LNB or LNC) located immediately behind the antenna feed. This configuration is used in all DBS-TV receiving systems.

4.3.2 Calculation of System Noise Temperature

The equivalent circuits in Figure 4.7a can be used to represent a receiver for the purpose of noise analysis. The noisy devices in the receiver are replaced by equivalent noiseless blocks with the same gain and noise generators at the input to each block such that the block produces the same noise at its output as the device it replaces. The entire receiver is then reduced to a single equivalent noiseless block with the same end-to-end gain as the actual receiver and a single noise source at its input with temperature Ts, called the *system noise temperature*.

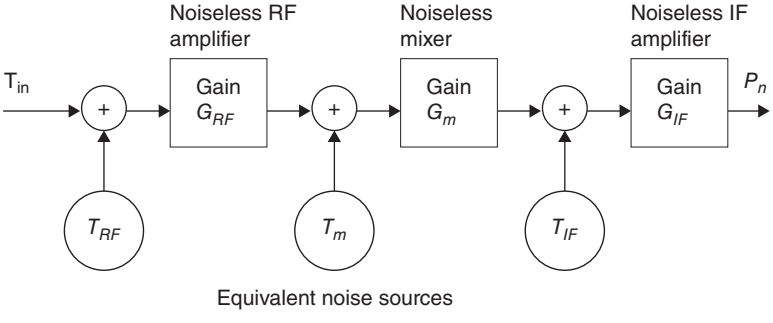


Figure 4.7a Noise model of receiver. T_{in} is the noise temperature of the sky and antenna, T_{rf} is the noise temperature of the LNA, G_{rf} is the gain of the LNA, T_m is the noise temperature of the mixer, G_m is the gain of the mixer, T_{if} is the noise temperature of the IF amplifier, and G_{if} is the gain of the IF amplifier. P_n is the noise power at the output of the receiver.

The total noise power at the output of the IF amplifier of the receiver in Figure 4.7a is given by

$$P_n = G_{IF}\, kT_{IF}\, B_n + G_{IF}\, G_m\, kT_m\, B_n + G_{IF}\, G_m\, G_{RF}\, kB_n(T_{RF} + T_{in}) \text{ watts} \quad (4.16)$$

where G_{RF}, G_m, and G_{IF} are respectively the gains of the RF amplifier, mixer, and IF amplifier, and T_{RF}, T_m, and T_{IF} are their equivalent noise temperatures. T_{in} is the noise temperature of the antenna, measured at its output port. Antennas do not usually generate noise unless they have ohmic loss. T_{in} accounts for noise radiated into the antenna from the signal path through the atmosphere and also any noise radiated from the earth into the sidelobes of the antenna pattern. Initial calculations of system noise temperature are usually made by assuming clear sky conditions and using an assumed value for attenuation on the signal path. Calculation of an exact antenna temperature requires convolution of the 3-D antenna pattern with a model of the 3-D temperature profile of the earth and sky, and is rarely attempted.

Any part of the signal path that incurs a loss by absorption of signal energy results in the generation of thermal noise. This is called an *ohmic loss* to distinguish it from other types of loss such as path loss. The loss in the atmosphere in clear sky conditions (no rain present) is caused mainly by absorption of microwave signal energy by oxygen and water vapor molecules. Because these molecules absorb microwave energy they also radiate thermal noise that is received by the earth station antenna. Thermal noise generated by the atmosphere is characterized by sky noise temperature.

Figure 4.7a shows a model of a noiseless receiver in which each block in the receiver is replaced by a noiseless block followed by an equivalent noise source at the output of the block. The noise source in each case has a noise temperature that results in the same output noise power as the noisy device, measured in the receiver noise bandwidth. In Figure 4.7b, the noise sources are combined into a single equivalent system noise source with a noise temperature T_s at the input of the receiver, and the receiver is represented as a single noiseless block that has the same end to end gain as the receiver in Figure 4.7a. Figure 4.7c is an alternative configuration to Figure 4.7b with a single equivalent noise source with noise temperature T_{no} at the output of the receiver.

Equation (4.16) can be rewritten as

$$\begin{aligned} P_n &= G_{IF}\, G_m\, G_{RF} \left[\frac{k\, T_{IF}\, B_n}{(G_m\, G_{RF})} + \frac{k\, T_m\, B_n}{G_{RF}} + k\, B_n\, (T_{RF} + T_{in}) \right] \\ &= G_m G_{IF}\, G_{RF}\, k\, B_n \left[T_{RF} + T_{in} + \frac{T_m}{G_{RF}} + \frac{T_{IF}}{G_m\, G_{RF}} \right] \text{ watts} \end{aligned} \quad (4.17)$$

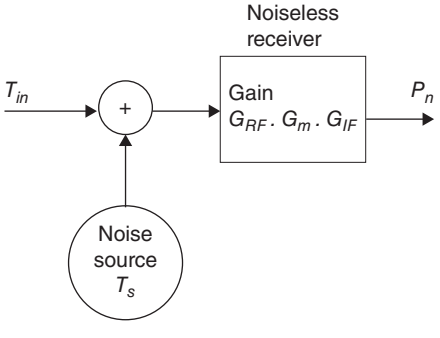


Figure 4.7b Noise model of receiver with a single noise source T_s, the system noise temperature, at the input to a noiseless receiver with identical gain to the receiver in Figure 4.7a.

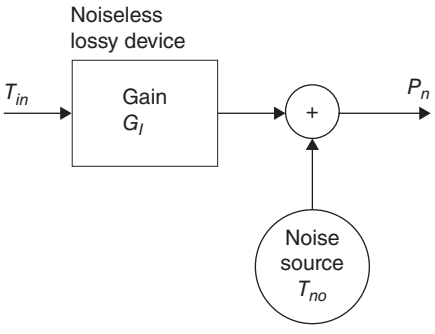


Figure 4.7c Noise model of receiver with a single noise source T_{no}, the system noise temperature, at the input to a noiseless receiver with identical gain to the receiver in Figure 4.7a.

The single source of noise shown in Figure 4.7b with noise temperature T_s generates the same noise power P_n at its output as the model in Figure 4.7a

$$P_n = G_m\, G_{IF}\, G_{RF}\, kT_s\, B_n \text{ watts} \tag{4.18}$$

The noise power at the output of the noise model in Figure 4.7b will be the same as the noise power at the output of the noise model in Figure 4.7a if

$$k\, B_n\, T_s = k\, B_n \left[T_{RF} + T_{in} + \frac{T_m}{G_{RF}} + \frac{T_{IF}}{G_m G_{RF}} \right] \text{ watts} \tag{4.19}$$

Hence the equivalent noise source in Figure 4.7b has a system noise temperature T_s where

$$T_s = \left[T_{RF} + T_{in} + \frac{T_m}{G_{RF}} + \frac{T_{IF}}{G_m G_{RF}} \right] \text{K} \tag{4.20}$$

Succeeding stages of the receiver contribute less and less noise to the total system noise temperature. Frequently, when the RF amplifier in the receiver front end has a high gain, the noise contributed by the IF amplifier and later stages can be ignored and the system noise temperature is simply the sum of the antenna noise temperature and the LNA noise temperature, so $T_s = T_{antenna} + T_{LNA}$. Note that the values for component gains in Eq. (4.20) must be linear ratios, not in decibels.

The noise model shown in Figure 4.7b replaces all the individual sources of noise in the receiver by a single noise source at the receiver input. This assumes that all the noise comes in from the antenna or is internally generated in the receiver. In some circumstances, we need to use a different model to deal with noise that reaches the receiver after passing through a lossy medium. Waveguide and rain losses are two examples. When raindrops cause attenuation, they radiate additional noise whose level depends on the attenuation. We can model the noise emission as a noise source placed at the output of the atmosphere, which is the antenna aperture. The noise model for an *equivalent output noise source* is shown in Figure 4.7c, and produces a noise temperature T_{no} given by

$$T_{no} = T_p(1 - G_l) \text{ K} \tag{4.21}$$

where G_l is the linear gain (less than unity, not in decibels) of the attenuating device or medium, and T_p is the physical temperature in degrees kelvin of the device or medium.

For an attenuation of A dB, the value of G_l is given by

$$G_l = 10^{-A/10} \tag{4.22}$$

Table 4.3 Gain and noise temperature values for 4 GHz receiver example

T_{in}	25 K	
T_{RF}	50 K	
T_m	500 K	
T_{IF}	1000 K	
G_{RF}	23 dB	(ratio 200)
G_{IF}	30 dB	(ratio 1000)

Example 4.3

Suppose we have a 4 GHz receiver with the gains and noise temperatures in Table 4.3.

Calculate the system noise temperature assuming that the mixer has a gain $G_m = 0$ dB. Recalculate the system noise temperature when the mixer has a 10 dB loss. How can the noise temperature of the receiver be minimized when the mixer has a loss of 10 dB?

Answer

The system noise temperature is given by Eq. (4.20)

$$T_s = [25 + 50 + (500/200) + (1000/200)] = 82.5 \text{ K}$$

If the mixer had a loss, as is usually the case, the effect of the IF amplifier would be greater. For a mixer with a loss of 10 dB, $G_m = -10$ dB and the linear value is $G_m = 0.1$ as a ratio. Then

$$T_s = [25 + 50 + (500/200) + (1000/20)] = 127.5 \text{ K}$$

The lowest system noise temperatures are obtained by using a high gain LNA. Suppose we increase the LNA gain in this example to $G_{RF} = 50$ dB, giving a ratio $G_{RF} = 10^5$. Then

$$T_s = [25 + 50 + (500/10^5) + (1000/10^4)] = 75.1 \text{ K}$$

The high gain of the RF LNA has made the system noise temperature almost as low as it can go. The minimum value of T_s is given by $T_{s\,min}$ where in this example

$$T_{s\,min} = T_{in} + T_{rf} = 75 \text{ K}$$

The mixer and IF amplifier contribute almost nothing to the system noise temperature. LNAs for use in satellite receivers usually have gains in the range 40–55 dB with the result that system noise temperature can be equated to $T_{in} + T_{rf}$.

Example 4.4

The system illustrated in Example 4.3, Table 4.3, has an LNA with a gain of 50 dB. A section of lossy waveguide with an attenuation of 2 dB is inserted between the antenna and the RF amplifier. Find the new system noise temperature for a waveguide temperature of 300°K.

Answer

The waveguide loss of 2 dB (ratio 1.58) can be treated as a gain, G_l that is less than unity: $G_l = 1/1.58 = 0.631$. The lossy waveguide attenuates the incoming noise and adds noise

generated by its own ohmic loss. The equivalent noise generator placed at the output of the section of waveguide that represents the noise generated by the waveguide has a noise temperature T_{wg}, where

$$T_{wg} = T_p(1 - G_l) = 300(1 - 0.631) = 110.7 \text{ K}$$

The waveguide attenuates the noise from the antenna, so $T_{in} = 0.631 \times 25 = 15.8$ K
The new system noise temperature, referred to the input of the LNA, is

$$T_s = 15.8 + 110.7 + 50 + (500/10^5) + (1000/10^4) = 176.6 \text{ K}$$

The system noise temperature is $10 \log_{10} (176.6/75) = 3.7$ dB higher than the original receiver configuration without the 2 dB waveguide loss. In addition, we have lost 2 dB of signal power so the receiver output CNR is reduced by 5.7 dB. Avoiding losses between the antenna and LNA is critical in a low noise receiver, which is why the LNA is mounted immediately behind the antenna feed in virtually all satellite communication receivers. Antennas for GPS receivers typically include an LNA in the antenna base, powered by a DC voltage across the conductors of the coaxial cable connecting the antenna to the GPS receiver. However, the RF filter in a GPS receiver is typically located ahead of the LNA to block interference that could saturate the amplifier. Any loss in the RF filter and the corresponding increase in system noise temperature is accepted in exchange for the reduction in interference.

We can refer the system noise temperature to the antenna output port by dividing the above result by G_l. This transfers the noise source from the LNA input to the waveguide input.

$$T_s = 176.6/0.631 = 280 \text{ K}$$

The new system noise temperature is 5.7 dB higher than the system noise temperature without the lossy waveguide, but there is no longer a loss of signal, so we have the same result for the reduction in CNR.

Note that when the system noise temperature is low, each 0.1 dB of attenuation ahead of the RF amplifier will add approximately 6.6 K to the system noise temperature. Using the formula in Example 4.2 with $T_p = 290\,^{\circ}$K, $G_l = -0.1$ dB $= 0.977$ as a ratio gives

$$T_{no} = 290 \times 0.023 = 6.6 \text{ K}$$

This is the reason for placing the front end of the receiver at the output of the antenna feed. Waveguide losses ahead of the LNA can have a disastrous effect on the system noise temperature of low noise receiving systems.

The value of T_{in} in Examples 4.3 and 4.4 was set to 25 K. This corresponds to an atmospheric path attenuation of approximately $0.1 \times 25/6.6 = 0.4$ dB, using the above formula and rounding to the nearest tenth of a dB, assuming a noiseless antenna. Note that in the analysis of communication systems, results in decibels are usually quoted to the nearest tenth of a dB. Including an additional decimal place implies that all calculations are correct to 0.01 dB, which is never the case because of the assumptions made at the beginning of the calculation.

Table 4.4 Comparison of noise temperature and noise figure

Noise temperature (K)	0	20	40	60	80	100	120	150	200	290
Noise figure (dB)	0	0.29	0.56	0.82	1.06	1.29	1.50	1.81	2.28	3.0
Noise temperature (K)	400	600	800	1000	1500	2000	3000	5000	10 000	
Noise figure (dB)	3.8	4.9	5.8	6.5	7.9	9.0	10.5	12.6	15.5	

4.3.3 Noise Figure and Noise Temperature

Noise figure is frequently used to specify the noise generated within a device. The operational noise figure is defined by the following formula (Krauss et al. 1980, p. 26)

$$NF = (SNR)_{\text{in}}/(\text{SNR})_{\text{out}} \tag{4.23}$$

where $(\text{SNR})_{\text{in}}$ is the SNR at the input to the device and $(\text{SNR})_{\text{out}}$ is the SNR at the output of the device. Because noise temperature is more useful in satellite communication systems, it is best to convert noise figure to noise temperature, T_{n}. The relationship is

$$T_{\text{n}} = T_{\text{o}}(NF - 1)\ \text{K} \tag{4.24}$$

where the noise figure is a linear ratio, not in decibels and where T_{o} is the reference temperature used to calculate the standard noise figure – usually 290 K. NF is frequently given in decibels and must be converted to a ratio before being used in Eq. (4.24).

Table 4.4 gives a comparison between noise figure and noise temperature over the range encountered in typical systems.

4.3.4 G/T Ratio for Earth Stations

The link equation can be rewritten in terms of CNR at the earth station

$$\frac{C}{N} = \left[\frac{P_t G_t G_r}{kT_s B_n}\right]\left[\frac{\lambda}{4\pi R}\right]^2 = \left[\frac{P_t G_r}{kB_n}\right]\left[\frac{\lambda}{4\pi R}\right]^2\left[\frac{G_r}{T_s}\right] \tag{4.25}$$

Thus $CNR \propto G_{\text{r}}/T_{\text{s}}$, and the terms in the square brackets are all constants for a given satellite system. The ratio $G_{\text{r}}/T_{\text{s}}$, which is usually quoted as simply G/T in decibels with units dBK^{-1}, can be used to specify the quality of a receiving earth station or a satellite receiving system, since increasing $G_{\text{r}}/T_{\text{s}}$ increases the received CNR.

Satellite terminals may be quoted as having a negative *G/T,* which is below $0\ \text{dBK}^{-1}$. This simply means that the numerical value of G_{r} is smaller than the numerical value of T_{s}.

Example 4.5 Earth Station G/T Ratio

An earth station antenna has a diameter of 30 m with an aperture efficiency of 68% and is used to receive a signal at 4150 MHz. At this frequency, the system noise temperature is 60 K when the antenna points at the satellite at an elevation angle of 28°. What is the

earth station *G/T* ratio under these conditions? If heavy rain causes the sky temperature to increase so that the system noise temperature rises to 88 K, what is the new G/T value?

Answer
First calculate the antenna gain. For a circular aperture

$$G_r = \eta_A 4\pi A/\lambda^2 = \eta_A(\pi D/\lambda)^2$$

At 4150 MHz, λ = 0.0723 m. Then

$$G = 0.68 \times (\pi 30/0.0723)^2 = 1.16 \times 10^6 \text{ or } 60.6 \text{ dB}$$

Converting T_s into dBK

$$T_s = 10 \log_{10} 60 = 17.8 \text{ dBK}$$

$$\text{G/T} = 60.6 - 17.8 = 42.8 \text{ dBK}$$

If $T_s = 88$ K in heavy rain

$$\frac{G}{T} = 60.6 - 19.4 = 41.2 \text{ dB/K}$$

4.4 Design of Downlinks

The design of any satellite communication is based on two objectives: meeting a minimum CNR for a specified percentage of time, and carrying the maximum revenue earning traffic at minimum cost. There is an old saying that "an engineer is a person who can do for a dollar what any fool can do for one hundred dollars." This applies to satellite communication systems. Any satellite link can be designed with very large antennas to achieve high CNRs under all conditions, but the cost will be very high. The art of good system design is to reach the best compromise of system parameters that meets the specification at the lowest cost. For example, if a satellite link is designed with sufficient margin to overcome a 20 dB rain fade rather than a 3 dB fade, an earth station antenna with seven times the diameter is required.

All satellite communications links are affected by rain attenuation. In the 6/4 GHz band the effect of rain on the link is small. In the 14/11 GHz Ku-band, and even more so in the 30/20 GHz Ka-band and higher frequency bands, rain attenuation becomes all important. Satellite links are typically designed to achieve reliabilities of 99.5–99.99%, averaged over a long period of time, typically a year. That means the CNR in the receiver will fall below the minimum permissible value for proper operation of the link for between 0.5% and 0.01% of the specified time; the link is then said to suffer an *outage*. The time period over which the percentage of time is measured can be a month, sometimes the *worst month* in attenuation terms, or a year. Attenuation due to heavy rain is a variable phenomenon, both with time and place. Chapter 7 discusses the prediction of path attenuation and provides ways to estimate the likely occurrence of outages on a given link. In this chapter we will simply assume certain rain attenuation statistics to use in examples of link design.

C-band links can be designed to achieve 99.99% reliability because the rain attenuation rarely exceeds 1 or 2 dB. The time corresponding to 0.01% of a year is 52 minutes; at this level of probability the rain attenuation statistics are usually not stable and wide fluctuations occur from year to year. Outages occur in heavy rain, usually in thunderstorms, and thunderstorm occurrence varies widely from year to year. A link designed to

have outages totaling 52 minutes each year may well have outages of several hours one year and none the next. Outage times of 0.1–0.5% of a year (8–40 hours) are often tolerated in Ku-band links used for DTH-TV. The allowable outage time for a link depends in part on the traffic carried. Telephone traffic needs real-time channels that are maintained for the duration of a call, so C-band or Ku-band terminals used for voice channels need sufficient link margin that outage times are small. Links between major computer networks also require very high availabilities. Satellite links used for internet access can tolerate more frequent outages because they do not require continuous transmission.

Rain attenuation can be overcome in many cases by using a small clear sky link margin and changing the forward error correction (FEC) coding rate and modulation when rain attenuation occurs. This technique is called adaptive coding and modulation (ACM), and has been adopted for many systems designed for internet access by satellite. The data rate on the link slows down when there is rain in the link, but since this occurs for only a small percentage of the time is tolerated in preference to an outage. Most of the time, clear sky prevails and high speed data transfer is the norm. Chapter 11 on internet access by satellite discusses the ACM technique in detail.

GEO Ka-band satellite links are currently becoming widespread in aeronautical services for aircraft crossing the oceans to provide internet access for passengers. Aircraft avoid flying through heavy rain and thunderstorms, and long distance flights are at altitudes above clouds and rain, except for thunderstorms. Many non-geostationary orbit (NGSO) and LEO systems also propose to use Ka-band links (see Chapter 9). DTH-TV (*satellite TV*) transmissions generally achieve better than 99.7% availability over a one year period – see Chapter 10 for details.

4.4.1 Link Budgets

Calculation of CNR in a receiver is simplified by the use of *link budgets*. A link budget is a tabular method for evaluating the received power and noise power in a radio link, and is similar to a monetary budget, where received power is regarded as equivalent to income and losses are equivalent to expenditure.

Link budgets invariably use decibel units for all quantities so that signal and noise powers can be calculated by addition and subtraction. Since it is usually impossible to design a satellite link accurately at the first attempt, link budgets make the task much easier because, once a link budget has been established, it is easy to change any of the parameters and recalculate the result. Table 4.5a shows a typical link budget for a C-band downlink using a global beam on a GEO satellite and a 9 m receiving earth station antenna.

The link budget must be calculated for an individual transponder, and must be repeated for each of the individual links. In a two-way satellite communication link there will be four separate links, each requiring a calculation of CNR. When a *bent pipe transponder* is used the uplink and downlink CNRs must be combined to give an *overall CNR*. In this section we will calculate the CNR for a single link. Later examples in this chapter demonstrate the evaluation of a complete satellite communication system.

Link budgets are usually calculated for a *worst case*, the one in which the link will have the lowest CNR. Factors that contribute to a worst case scenario include: an earth station located at the edge of the satellite coverage zone or spot beam where the received signal is typically 3 dB lower than in the center of the beam because of the satellite antenna pattern (*footprint*); maximum path length from the satellite to the earth station; a low elevation angle at the earth station giving the highest atmospheric path attenuation in

clear air; and maximum rain attenuation on the link causing loss of received signal power and an increase in receiving system noise temperature. The edge of the coverage pattern of the satellite antenna and the longest path usually go together. However, when a satellite has a multiple beam antenna, this will not be the case. Earth station antennas are assumed to be pointed directly at the satellite, and therefore operate at their on-axis gain. If the antenna is mispointed, a loss factor is included in the link budget to account for the reduction in antenna gain.

The calculation of CNR in a satellite link is based on the two equations for received signal power and receiver noise power that were presented in Sections 4.1 and 4.2. Equation (4.12) gives the received carrier power in dB watts as

$$P_r = \text{EIRP} + G_r - L_p - L_a - L_{ta} - L_{ra} \text{ dBW} \tag{4.26}$$

From Eq. (4.13), a receiving terminal with a system noise temperature T_s K and a noise bandwidth B_n Hz has a noise power P_n watts referred to the input of the LNA where

$$P_n = k\, T_s\, B_n \text{ watts} \tag{4.27}$$

The receiving system noise power is usually written in decibel units as

$$N = k + T_s + B_n \text{ dBW} \tag{4.28}$$

where k is Boltzmann's constant (−228.6 dBW/K/Hz), T_s is the system noise temperature in dBK, and B_n is the noise bandwidth of the receiver in dBHz. Note that because we are working in units of power, all decibel conversions are made as 10 $\log_{10}$ (T_s) or 10 $\log_{10}$ (B_n). The 20 $\log_{10}$ factor used in the calculation of path loss results from the use of the squared power in the $(4\pi R/\lambda)^2$ term in the path loss equation. Electrical engineering students typically first meet decibels in the calculation of amplifier gain, as $G = 20 \log_{10} (V_{out}/V_{in})$, which is strictly a misuse of decibels. A number in decibels is a power ratio; using 20 $\log_{10}$ (…) for amplifier gain assumes that the input and output impedances of the amplifier are the same. In communications, conversions to decibels are always 10 $\log_{10}$ (…) unless the quantity is squared, when we can write 20 $\log_{10}$ (…) rather than 10 $\log_{10}$ $(\ldots)^2$.

Example 4.6 Link Budget for C-Band Downlink With Earth Coverage Beam

The satellite used in this example is in GEO and carries 24 C-band transponders, each with a bandwidth of 36 MHz. The downlink band is 3.7–4.2 GHz and the satellite uses dual orthogonal circular polarizations to double the number of available channels, thus providing an effective RF bandwidth of 864 MHz. Figure 4.8 illustrates a GEO satellite located at 30°W longitude serving the Atlantic Ocean region.

The satellite provides coverage of the visible earth, which subtends an angle of approximately 17° from a satellite in geostationary orbit, by using a global beam antenna. Antenna beamwidth and gain are linked together by the relationship

$$G \approx \sqrt{30{,}000/(\text{beamwidth in degrees})^2}$$

where G is a ratio (not in decibels). The on-axis gain of the global beam antenna is approximately 20 dB. However, we must make a worst case assumption in the link budget calculation, which is for an earth station at the edge of the coverage zone of the satellite where the effective gain of the antenna is 3 dB lower, at 17 dB. The edge of the coverage zone does not necessarily have to be at the −3 dB contour of the satellite antenna

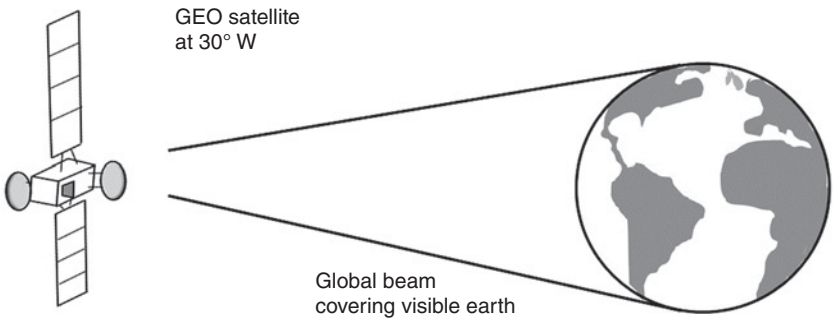


Figure 4.8 GEO satellite at 30° west longitude with global beam antenna serving the Atlantic Ocean region. Note that most of the energy transmitted by the satellite falls into the ocean; only a small fraction reaches populated areas.

footprint. Coverage extends beyond the −3 dB contour into the sidelobes of the satellite antenna pattern. A larger receiving antenna can be used to compensate for the loss of signal power when operating outside the −3 dB contour.

The CNR for the downlink is calculated in clear air conditions and also in heavy rain. The satellite can connect earth stations in North and South America to earth stations in Europe and Africa using a global beam that covers the visible earth as seen from the satellite. However, most of the signal radiated by the satellite ends up in the ocean and only a small part is available for communications between continents. A satellite antenna with a gain of 20 dB has an effective aperture diameter of 5.6 wavelengths given by

$$G = \eta_{\mathrm{A}}\left(\frac{\pi D}{\lambda}\right)^2$$

which gives $D = 0.42$ m at a frequency of 4 GHz. If the satellite antenna's aperture efficiency is 65%, the physical diameter is 0.52 m. The calculation of CNR is made at a mid-band frequency of 4 GHz. Appendix B explains the properties of antennas.

The saturated output power of the transponder is 80 W = 19 dBW. Reducing the output power of an amplifier from its maximum value helps to linearize the channel, so we will assume an output *backoff* of 2 dB, which means the power transmitted by the transponder is now 17 dBW.

Hence the on-axis EIRP of the transponder and antenna is

$$P_{\mathrm{t}}G_{\mathrm{t}} = 17 + 20 = 37 \text{ dBW}$$

The transmitted signal is a single 30 MHz bandwidth channel carrying a digital signal in this example.

The maximum path length for a GEO satellite link at the edge of coverage is 40 000 km, which gives a path loss of 196.5 dB at 4 GHz ($\lambda = 0.075$ m). We must make an allowance in the link budget for some losses that will inevitably occur on the link. At C-band, propagation losses are small, but the slant path through the atmosphere will suffer a typical attenuation of 0.2 dB in clear air. We will allow an additional 0.5 dB *margin* in the link design to account for miscellaneous losses, such as antenna mispointing, polarization mismatch, and antenna degradation, to ensure that the link budget is realistic.

Table 4.5a summarizes the parameters of the link and presents a link budget for the downlink from the satellite to a receiving earth station.

Table 4.5a Example for a C-band GEO satellite downlink budget in clear air

C-band satellite parameters		
Transponder saturated output power 80 W	$P_{t\,sat}$	19 dBW
Antenna gain, on axis	G_t	20 dB
Transponder bandwidth	B_{transp}	36 MHz
Downlink frequency band		3.7–4.2 GHz
Digital signal noise bandwidth	B_n	30 MHz
Minimum permitted overall CNR in receiver	$(CNR)_{o\,min}$	14.0 dB
Receiving C-band earth station		
Downlink frequency		4.00 GHz
Antenna gain, on axis, 4 GHz	G_r	49.7 dB
Receiver IF bandwidth	B_n	30 MHz
Receiving system noise temperature	T_s	45 K
Downlink power budget		
Satellite transponder output power, 80 W	P_t	19.0 dBW
Transponder output backoff	B_{out}	−2.0 dB
Satellite antenna gain, on axis	G_t	20.0 dB
Earth station antenna gain	G_r	49.7 dB
Free space path loss at 4 GHz	L_p	−196.5 dB
Edge of beam loss for satellite antenna	L_{ant}	−3.0 dB
Clear sky atmospheric loss	L_a	−0.2 dB
Other losses (*margin*)	L_{misc}	−0.5 dB
Received power at earth station	P_r	−113.5 dBW
Downlink noise power budget in clear air		
Boltzmann's constant	k	−228.6 dBW/K/Hz
System noise temperature, 58 K	T_s	17.6 dBK
Noise bandwidth, 30 MHz	B_n	74.8 dBHz
Receiver noise power	N	−136.2 dBW

The system noise temperature is 58 K because the clear air attenuation of 0.2 dB creates an antenna noise temperature of 13 K, which adds to the LNA noise temperature of 45 K. Hence the CNR in the receiver in clear air is

$$\text{CNR} = P_r - N = -113.5\ \text{dBW} - (-136.2\ \text{dBW}) = 22.7\ \text{dB}$$

The receiving earth station has a gain of 49.7 dB at 4 GHz, and a receiving system noise temperature of 58 K in clear sky conditions. The G/T ratio for this earth station is

$$\frac{G}{T} = 49.7 - 10\ \log_{10}\ 58 = 32.0\ \text{dBK}^{-1}$$

The earth station receiver CNR is first calculated for clear sky conditions, with no rain in the slant path. The CNR is then recalculated taking account of the effects of rain. The minimum permitted overall CNR for this link is 14.0 dB giving a maximum BER of 10^{-6}

with quadrature phase shift keying (QPSK) modulation and no FEC. A CNR of 22.7 dB in clear air with QPSK modulation gives a BER well below 10^{-16} (Chapter 5 shows how this calculation is made). At a bit rate of 60 Mbps the theoretical time between bit errors is longer than 3×10^8 seconds or 9.5 years. The link is said to be *essentially error free*. Since clear sky conditions at 4 GHz prevail in most geographical locations for more than 97% of any given year, the link operates error free for most of the year.

Example 4.7 C-Band Link CNR in Heavy Rain

The results for the receiving terminal CNR in clear air are used as the starting point for the calculation of CNR with rain in the slant path to the terminal. Table 4.5a shows that we have a downlink CNR of 22.7 dB in clear air, giving a *link margin* of 8.7 dB over the minimum CNR allowed of 14.0 dB. This link margin is available in clear air conditions, but will be reduced when there is rain in the slant path.

Heavy rain in the slant path can cause up to 1 dB of attenuation at 4 GHz when the satellite has a low elevation angle and the slant path through the rain is long, which reduces the received power by 1 dB and increases the noise temperature of the receiving system. Using the output noise model discussed in the previous section with a medium temperature of 273 K, and a total path loss for clear air plus rain of 1.2 dB (ratio of 1.32), the sky noise temperature in rain is

$$T_{\text{sky}} = 273 \times (1 - 1/1.32) = 66 \text{ K}$$

In clear air the sky noise temperature is approximately 13 K, the result of 0.2 dB of clear air attenuation. The system noise temperature with rain in the downlink path is $T_{s\ rain}$ where

$$T_{s\ rain} = 45 + 66 = 111 \text{ K}$$

The clear air system noise temperature is 58 K. The increase in system noise temperature results in a corresponding increase in receiver noise power given by ΔN where

$$\Delta N = 10\ \log_{10}(111/58) = 2.8 \text{ dB}$$

Note that the CNR in the C-band earth station receiver is affected much more by the increase in sky noise temperature than by the rain attenuation. In making this calculation, it is important to remember that clear air attenuation, 0.2 dB, in this case, is always present and must be added to the rain attenuation to give the total path attenuation before calculating the system noise temperature in rain. (You still want to be able to breathe when it rains.)

We can now adjust the link budget very easily to account for heavy rain in the slant path without having to recalculate the CNR from the beginning. The received carrier power is reduced by 1 dB because of the rain attenuation and the system noise temperature is increased by 2.8 dB. Table 4.5b shows the new downlink budget in rain.

CNR in the receiver in heavy rain is

$$\text{CNR} = P_{r\ rain} - N_{\text{rain}} = -114.5 \text{ dBW} - (-132.7 \text{ dBW}) = 18.2 \text{ dB}$$

The CNR in rain has a downlink margin of 4.2 dB over the minimum permissible CNR of 14.0 dB. The excess CNR margin will translate into lower BER, and can be traded off against earth station antenna gain to allow the use of smaller (and therefore lower cost)

antenna. We will examine how the 4.2 dB of link margin can be traded against other parameters in the system.

A reduction in earth station antenna gain of 4.2 dB is a reduction in the gain value, as a ratio, of 2.63. Antenna gain is proportional to diameter squared, so the diameter of the earth station antenna can be reduced by a factor of $\sqrt{2.63} = 1.62$, from 9 to 5.6 m to lower the cost of the earth station.

Example 4.8 4 GHz Downlink With Regional Beam

Global beam antennas are not widely used, although most satellites with international coverage carry them to serve outlying earth stations that are not within the coverage of regional beams. The low gain and broad coverage of a global beam results in poor utilization of transponder power, since most of the transmitted power is lost over the oceans, as seen in Figure 4.8, and global beams have been referred to derisively as *fish warmers*. Regional TV signal distribution is much more common, so the C-band link in Tables 4.5a and 4.5b is more likely to use a regional antenna, serving a continent or a group of countries. The United States, for example, can be covered with a 6° by 3° beam, as illustrated in Figure 4.9. Additional beams may be needed to cover Alaska and Hawaii, or a more complex antenna with a *shaped beam* can be used.

The gain of a typical satellite antenna providing coverage of the 48 contiguous states (CONUS) is 32.2 dB on axis (calculated from $G = 30\,000/[\theta_1 \times \theta_2]$), which is 12.2 dB higher than the on-axis gain of a global beam. Using the link budget in Tables 4.5a and 4.5b, we can trade the extra 12.2 dB gain (ratio 16.7) of a regional coverage satellite antenna for a reduction in earth station antenna dimensions. For the example of a 9.0 m earth station antenna in Example 4.7, we could reduce the antenna diameter by a factor of $\sqrt{16.7} \approx 4.1$ m to a diameter of 2.2 m (approximately 7 ft 3 in.). The cost of antennas increases approximately as the diameter of the antenna to the power 2.7 for antennas larger than 2 m. (See Appendix B for details.) Reducing the diameter of the

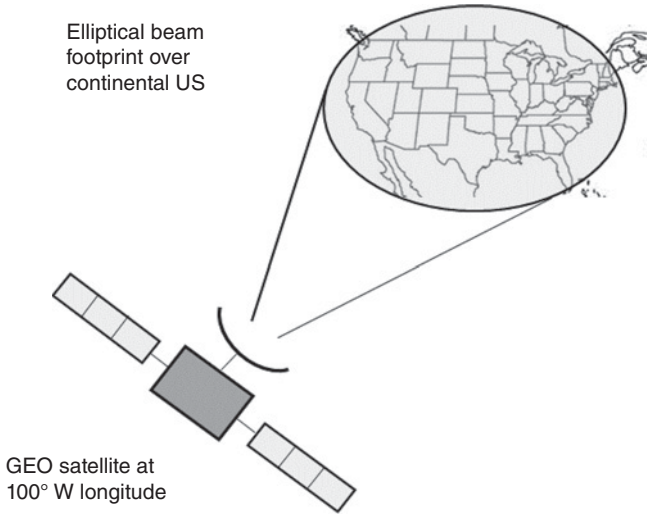


Figure 4.9 GEO satellite at 100°W longitude with elliptical regional beam antenna serving the continental United States.

earth station antenna from 9.0 to 2.2 m reduces its cost by a factor of approximately 45, a very significant cost saving.

Table 4.5b C-band downlink budget in rain

Received power at earth station in clear air	P_{rca}	−113.5 dBW
Rain attenuation	A	1.0 dB
Received power at earth station in rain	$P_{r\ rain}$	−114.5 dBW
Receiver noise power in clear air	N_{ca}	−135.5 dBW
Increase in noise power due to rain	ΔN_{rain}	2.8 dB
Receiver noise power in rain	N_{rain}	−132.7 dBW

In the 1970s and 1980s, television programming was distributed to cable TV *head ends* by C-band regional satellites, and later by Ku-band satellites. The C-band signals were transmitted as one video channel per 36 MHz transponder using frequency modulation (FM) and at first were not encrypted. An industry grew up supplying 8 and 10 ft diameter dishes to home owners, equipped with receivers with a 100 K LNA, allowing reception of cable TV channels without payment. The threshold for successful demodulation of the FM video signal was 11 dB. Comparing these parameters to the example in Tables 4.5a and 4.5b shows that a link margin of approximately 2 dB was available with the C-band home satellite TV system. Eventually the video signals of the popular TV programs were encrypted and users of the satellite receiving systems were forced to pay a subscription to receive those cable TV channels.

The above examples show how the link budget can be used to study different combinations of system parameters. Most satellite link analyses do not yield the wanted result at the first try, and the designer or analyst must use the link budget to adjust system parameters until an acceptable result is achieved. More examples of link budgets are included later in this chapter.

4.5 Ku-Band GEO Satellite Systems

Satellite communication systems developed rapidly through the 1970s and 1980s filling all the available 4 GHz transponders on GEO satellites serving populated areas of the world. Ku-band was the next available frequency band and larger GEO satellites carrying both C-band and Ku-band transponders were launched to meet the expanding demand for video distribution services and long distance communications. Voice traffic largely disappeared from GEO satellites because of the long delays and echo problems that are inevitable on a very long two-way GEO satellite link. The growth of optical fiber links, including under-sea fiber optic cables captured most of the point to point traffic. As seen in Chapter 1, the ability of a satellite to broadcast signals to an entire continent favored their use for distribution of television programming and direct to home television services. VSAT networks, where a single gateway station connects to hundreds or thousands of small earth stations use Ku-band almost exclusively. This concept is also proposed for several of the NGSO systems (see Chapters 9 and 11) where a Ku- or Ka-band link to a tracking phased array antenna on the ground from a tracking phased array on the satellite acts as a gateway for wireless local loops carrying internet traffic.

The 14/12 GHz band offers the advantage over C-band of higher gain earth station antennas and smaller antenna sizes. However, rain affects Ku-band links much more than C-band, requiring link designs to include larger margins to maintain acceptable availability. The antennas on a GEO satellite using Ku-band to provide coverage of a specific region, for example, North America, have the same gain as at C-band, but have smaller diameters.

4.5.1 Direct Broadcast Satellite Television

DBS-TV originally started in Europe in the 1980s using analog FM transmission in Ku-band. It achieved a reasonable measure of success, due in part to the much slower introduction of cable TV systems in Europe than occurred in the United States. In the 1990s, digital transmission became possible, and several systems were developed in the United States in the 12.2–12.7 GHz band allocated to DBS-TV services. In the United States, Directv, a system originally developed by a consortium led by Hughes, has been very successful and had over 19 million customers by year-end 2016, offering 200 television and audio channels. Another DBS-TV provider in the United States, Echostar, (Dish Network) offered similar services and had 14 million customers in year 2016. TV and audio channels are available from DBS-TV providers in a mixture of subscription packages, much like cable TV companies offer, and as pay per view for individual movies and special events. In rural areas of the United States, DBS-TV offers hundreds of television channels in place of the few terrestrial broadcasting stations that are typically available. In city areas, DBS-TV offers an alternative to cable television at a similar cost. The development of low cost Ku-band antennas and receivers, Motion Pictures Experts Group (MPEG) video compression techniques, and high speed digital integrated circuits specifically for DBS television, has made DBS-TV practical. Chapter 10 gives more detail on the design of DBS-TV systems. Figure 4.10 illustrates a DBS-TV system.

In North America the 12.2–12.7 GHz band was set aside for exclusive use by DBS-TV GEO satellite downlinks so that high power transponders could be used on specially designed DBS-TV satellites. Typical transponder output levels are 100–200 W with a maximum flux density at the earth's surface of up to $-100\,\text{dBW/m}^2$. The satellites typically carry 16 transponders, with a typical total transmitted RF power of 2.6 kW. Uplinks to DBS-TV satellites are in the 17 GHz band. DBS-TV satellites are large and heavy, use a three-axis stabilized design, and have a large area of solar cells to generate the power required by the transponders. Typical mass for a DBS-TV satellite in 2017 was 6800 kg at launch, among the largest commercial GEO satellites at that time.

The flux density at the earth surface produced by a DBS-TV 160 W transponder is typically in the range –110 to $-125\,\text{dBW/m}^2$ with an antenna footprint that covers the continental United States, which allows small receiving antennas (dishes) to be used for DBS-TV reception with diameters in the range 0.45–0.9 m. The small dish required for DBS-TV reception played a critical part in the acceptance and success of DBS-TV in the United States. Originally, the analog cable TV signals were not encrypted when transmitted via satellite, so it was possible to receive free TV signals with a 2–3 m C-band dish having a small margin for rain attenuation. The US Congress took the view that the cable TV companies should encrypt their signals if they did not want unauthorized viewing of the program material. Local governments of many cities and towns refused to permit these large dishes in residential areas, although they became popular in rural areas and an estimated 4 million systems were sold in the 1980s.

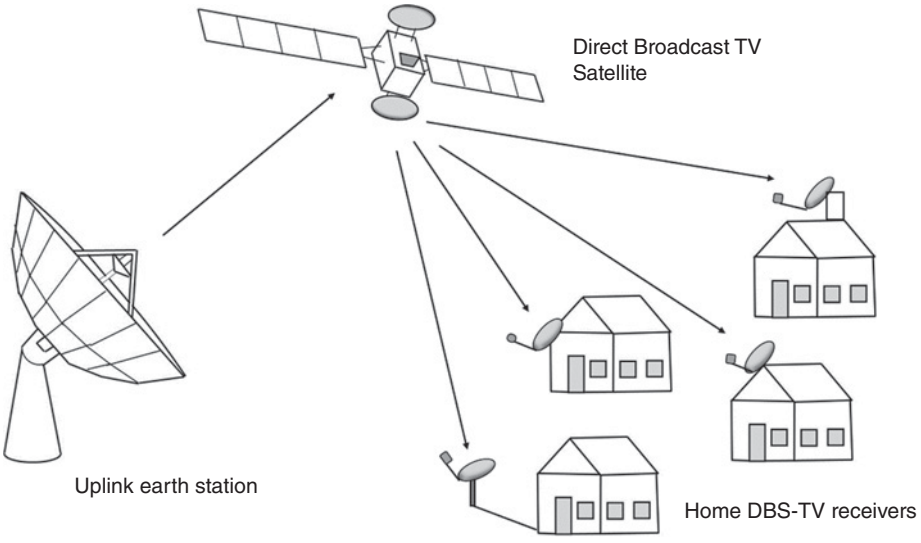


Figure 4.10 Illustration of a direct broadcast satellite television system (DBS-TV). The uplink earth station transmits multiple digital TV signals in compressed form to a number of transponders on the satellite. DBS-TV systems operate in Ku-band and Ka-band.

The US Congress passed laws in the 1990s that prevented local governments from restricting the use of antennas less than 1 m in diameter, opening up a large market for Ku-band DBS-TV services. Home owners' associations are not covered by the restriction, and sometimes do not permit satellite dishes on the roofs of houses in a private housing development. A similar approach was adopted in Europe for DBS receive antennas, and also over much of the globe. Almost anywhere you can travel on Earth, you will see small satellite TV antennas sprouting from roofs, walls, and chimneys.

The high flux density created by powerful transponders makes sharing of the DBS-TV frequency bands impossible, so the 12 GHz DBS-TV band, known as the broadcast satellite service (BSS) band is allocated exclusively for television broadcasting. The small home receiving antenna has a wide beam, typically 4° for a 0.45 m (18 in.) dish, which forces wide spacing of DBS-TV satellites to avoid interference by the signals from adjacent DBS-TV satellites. A 9° spacing in the GEO arc was adopted by the United States, later reduced to 4.5°, which restricts the number of DBS-TV satellites that can be placed in geostationary orbit to serve the United States. Several DBS-TV satellites can be clustered at the same GEO location with separations of one quarter of a degree, with transponders operating at different frequencies and polarizations. This increases the capacity of the GEO orbit. In the 1990s the US FCC successfully auctioned spectrum and orbital locations for DBS-TV satellites, raising hundreds of millions of dollars from companies that foresaw a profitable commercial venture.

The DBS-TV system must provide a received signal power at the small receiving antenna that has an adequate CNR margin in clear sky conditions. Heavy rain will cause attenuation that exceeds the link margin, so occasional *outages* will be experienced, especially during the summer months when thunderstorms and heavy rain are more frequent. The CNR margins used in DBS-TV systems are usually quite small to avoid the need for a large receiving antenna. The selection of a CNR margin is a design

trade-off between the outage level that customers can be expected to tolerate, the maximum allowable diameter of the receiving dish antenna, and the power output from the satellite transponders. Typical designs with receiving antennas in the 0.5–0.9 m range and 100–200 W satellite transponders yield rain attenuation margins of 3–8 dB depending on the location of the receiving terminal within the satellite antenna coverage, and outage times totaling 10–40 hours per year. These link margins are much lower than those found in high capacity communication systems. Availability of the link has been exchanged for a smaller earth terminal antenna and lower equipment costs to the user. The transmitting antennas used on DBS-TV satellites providing service to the United States have footprints that direct more signal energy to the south east of the country where heavy rain, and consequently rain attenuation is most frequent. Satellites with multiple beams can also be used for DBS-TV allowing higher power to be directed to areas where heavy rainfall occurs more frequently, and also to provide higher capacity to densely populated regions. Chapter 10 gives details of the design of DBS-TV satellites and illustrates typical satellite footprints for the United States.

A representative link budget for a GEO DBS-TV system serving the United States with a regional beam is shown in Table 4.6a. The received signal power P_r and the noise power N are included in the link budget along with the receiver CNR and link margin. The threshold CNR value is set at 8.3 dB, corresponding to a system using the digital video broadcast standard for satellites (DVB-S), with QPSK modulation, half rate FEC

Table 4.6a Link budget for Ku-band DBS-TV receiver with regional beam and DVB-S signals

Transponder saturated output power, 160 W	$P_{t\,sat}$	22.0 dBW
Transponder output backoff	B_o	−1.5 dB
Transponder operating output power	P_t	20.5 dBW
Transponder bandwidth	B_{xp}	30 MHz
Satellite antenna beam on-axis gain	G_t	34.3 dB
Path loss at 12.5 GHz, 38 000 m path	L_p	−206.0 dB
Receiving antenna gain, on axis	G_r	33.8 dB
Edge of beam loss	L_{ta}	−3.0 dB
Clear sky atmospheric loss	L_a	−0.4 dB
Miscellaneous losses	L_{misc}	−0.4 dB
Received power, C	P_r	−121.2 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz
LNA noise temperature	T_{LNA}	86 K
System noise temperature, clear sky, 110 K	T_s	20.4 dBK
Receiver noise bandwidth, 20 MHz	B_n	73.0 dBHz
Receiver noise power	N	−135.2 dBW
Clear sky CNR in receiver		14.0 dB
Link margin over 8.3 dB threshold		5.7 dB
Modulation and FEC coding		QPSK rate 1/2
Bit rate on downlink		20 Mbps
Link availability throughout US		Better than 99.7%

coding, and an implementation margin of 1.4 dB. Half rate FEC coding can provide 6 dB of coding gain, and a maximum BER of 10^{-6}. These values were typical for DBS-TV around year 2000. The earth station is located on the −3 dB contour of the downlink beam – as mentioned above, the center of the beam is skewed toward the south eastern part of the United States to counter the higher rain attenuation experienced in that part of the country.

Table 4.6b presents a link budget for the same service using a satellite with spot beams and the DVB-S2 signal format. The DVB-S2 standard has been widely introduced with DBS-TV satellites after 2010, but requires a different receiver at the customers' premises. Spot beams provide higher EIRP toward the receiving area and allow the use of lower transponder power. The DBV-S2 standard offers a wide range of phase shift keying (PSK) modulation and FEC code rates. Higher bit rates can be achieved by using 8-PSK modulation and 2/3 FEC coding rate, which has a threshold CNR of 8.4 dB. The spot beams are often used for delivery of high definition local programming to television markets within the spot beam area. DBS-TV is described in more detail in Chapter 10.

Table 4.6b Link budget for Ku-band DBS-TV receiver with spot beams and DVB-S2 signals

Transponder saturated output power, 50 W	$P_{t\,sat}$	17.0 dBW
Transponder output backoff	B_o	−1.5 dB
Transponder operating output power	P_t	15.5 dBW
Transponder bandwidth	B_{xp}	30 MHz
Satellite antenna spot beam on-axis gain	G_t	45.0 dB
Path loss at 12.5 GHz, 38 000 m path	L_p	−206.0 dB
Receiving antenna gain, on axis	G_r	33.8 dB
Edge of beam loss	L_{ta}	−3.0 dB
Clear sky atmospheric loss	L_a	−0.4 dB
Miscellaneous losses	L_{misc}	−0.4 dB
Received power, C	P_r	−115.5 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz
LNA noise temperature	T_{LNA}	86 K
System noise temperature, clear sky, 110 K	T_s	20.4 dBK
Receiver noise bandwidth, 22 MHz	B_n	73.4 dBHz
Receiver noise power	N	−134.8 dBW
Clear sky CNR in receiver		19.3 dB
Link margin over 8.5 dB threshold		10.8 dB
Modulation and FEC coding		8-PSK rate 2/3
Bit rate on downlink		44 Mbps
Link availability throughout United States		Better than 99.9%

Note that by employing spot beams and using DVB-S2 format signals, the bit rate on the link has more than doubled compared to earlier DBS-TV systems, and the link margin has increased.

In discussing Ku-band rain attenuation, we will use statistics that are representative of many locations in the central and eastern parts of the United States, where typical path

attenuation in rain exceeds 3 dB for 0.2% (15 hours) and 6 dB for 0.01% (52 minutes) of an average year. Such attenuation levels at the given time percentage are typical of many temperate latitude locations such as the west coast of the United States, central United States, Virginia, and other states north of Virginia on the east coast of the United States, Europe, Chile, Uruguay, and New Zealand. (See Chapter 7 for details of rain attenuation probabilities and distributions.) DBS-TV receiving systems are typically designed to have an average annual availability exceeding 99.7%, which corresponds to an outage time of 0.3% of the year, or about 25 hours. For much of the United States, this corresponds to rain attenuation in the slant path of 3 dB and requires a link margin of 5.7 dB when allowance is made for the increase in antenna noise temperature that accompanies 3 dB of rain attenuation.

Direct broadcast TV systems in the United States adopted the DVB-S standard for modulation and FEC by the mid 1990s. The DBV-S standard employs QPSK modulation with a double layer of FEC encoding. Figure 4.11 shows the relationship between BER and CNR for several different combinations of modulation and FEC methods. The effective CNR includes an allowance for implementation margin, which accounts for the non-ideal performance of real satellite communication links. See Chapters 5 and 10 for more details of these relationships. Allowing for a non-ideal receiver, a threshold CNR of approximately 14 dB is required at the input to the QPSK demodulator to maintain the BER below 10^{-6} without FEC. When DVB-S standard coding is applied to the signal, the threshold for BER = 10^{-6} is reduced by 6–8 dB, with a threshold in the 6–8 dB region.

The link budget in Table 4.6a shows how the required clear sky CNR is achieved for a receiver located on the −3 dB contour of the satellite antenna beam. A receiver located in the center of this beam would have a clear sky CNR 3 dB higher, and a corresponding fade

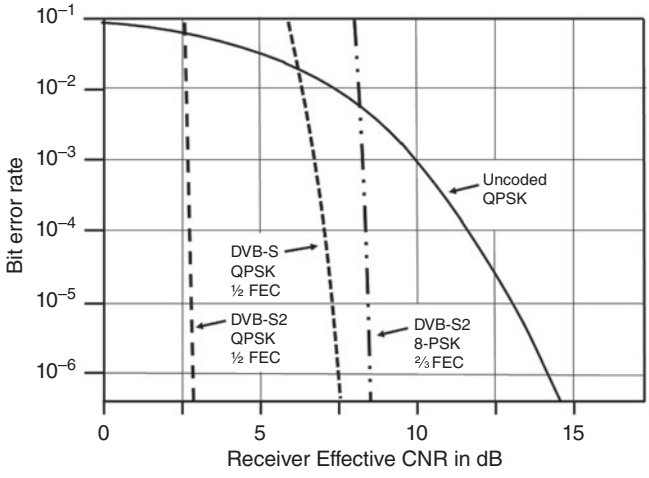


Figure 4.11 Bit error rates for different modulation and coding methods. The DVB-S QPSK with half rate FEC coding curve corresponds to the performance of typical DBS-TV receivers around year 2000 using the DVB-S signal format. Later receivers that take advantage of the DVB-S2 format have a threshold approximately 4 dB lower than the DVB-S receiver. Where sufficient signal is available, modulation can be 8-PSK with 2/3 rate FEC, which can send twice as many bits as DVB-S signals in the same transponder bandwidth. BER: Bit error rate (probability of a bit error); CNR: Carrier to noise ratio; DVB-S: Digital video broadcast standard for satellites. Effective CNR includes an allowance for non-ideal SRRC filtering in the receiver. (Implementation margin is 1.8 dB in this figure.)

margin of 8.7 dB, sufficient to ensure only a few outages each year. However, only one person can live at the center of the beam. Antenna beams have a shape that is parabolic in decibels. Half of the footprint area within the −3 dB contour is enclosed by the −2 dB contour, so half of the DBS-TV terminals have a link margin of 6.7 dB or higher.

In the link budget in Table 4.6a a transponder with saturated output power of 160 W is used, with 1.5 dB backoff. The satellite antenna gain is 34.3 dB on axis, corresponding to a high efficiency antenna with a beam that is shaped to cover the land mass of the United States. The beam is approximately 5.5° wide in the E–W direction and 2.5° in the N–S direction. The resulting coverage zone, taking account of the earth's curvature, is approximately 4000 km E–W and 2000 km N–S. A maximum path length of 38 000 km is used in this example. The receiving antenna is a high efficiency design with a front-fed offset parabolic reflector 0.45 m diameter and a circularly polarized feed. The offset design ensures that the feed system does not block the aperture of the antenna, which increases its efficiency. The gain of the antenna in this example is 33.8 dB at 12.5 GHz with an aperture efficiency of 70%. The receiver is located at the −3 dB contour of the transmitting antenna, and a loss of 0.4 dB for clear sky attenuation at 12 GHz and 0.4 dB for receive antenna mispointing and other losses is included. The result is a received carrier power of −121.2 dBW in clear sky conditions.

The noise power budget of the link is based on a receiver noise bandwidth of 20 MHz. The IF filters in the receiver must be designed to match the symbol rate of the transmitted signal, and to approximate a *square root raised cosine* (SRRC) transfer function. (See Chapter 5 for details of digital transmission techniques.) The noise bandwidth of all SRRC filters is always numerically equal to the symbol rate of the digital transmission. In the DBS-TV system described in Table 4.6a, a QPSK signal with a symbol rate of 20 Msps is assumed, which results in a receiver noise bandwidth of 20 MHz. The 20 Msps QPSK transmission delivers a bit rate of 40 Mbps, but the half rate FEC coding reduces the data rate to 20 Mbps. A 20 Mbps data stream can carry 6 or 7 live compressed digital video signals using MPEG-2 compression, or up to 10 prerecorded and compressed video signals.

DBS-TV receivers can be used outside the −3 dB contour of the satellite beam, but will have a lower link margin and consequently more outages per year, if heavy rain occurs frequently. For example, a receiver on the −5 dB contour of the satellite beam will have a link margin of 3.7 dB, which would allow about 2 dB of rain attenuation before the CNR reaches threshold. If the user is in a relatively dry area, for example, central Canada, the performance of the receiving system may be quite acceptable, or a larger receiving antenna can be used to compensate for the lower received signal level.

In Table 4.6b, the DVB-S2 signal format of 8-PSK with 2/3 rate FEC coding used with spot beams allows the 30 MHz bandwidth transponder to deliver data at 44 Mbps. MPEG-4 compression is used for high definition TV broadcasts and requires roughly 4 Mbps per TV channel, so the transponder can carry up to 11 high definition channels. (The number of channels can vary depending on the configuration of MPEG-4 compression that is used.) Most of the parameters in Tables 4.6a and 4.6b are the same, demonstrating the increase in capacity that can be achieved with spot beams and DVB-S2 signals.

The CNR in the home receiver will fall when rain is in the path between the satellite and the receiving antenna. Much of the reduction in CNR is caused by an increase in the sky noise temperature. The following calculations show how the system noise temperature and $(CNR)_{dn}$ are determined when rain attenuation is present in the downlink.

4.5.2 Effect of Rain on Direct to Home Satellite TV Ku-Band Downlink

The first step is to determine the total path attenuation, A_{total} in dB, which is the sum of the clear sky path attenuation due to atmospheric gaseous absorption, A_{ca} and attenuation due to rain, A_{rain}

$$A_{\text{total}} = A_{\text{ca}} + A_{\text{rain}} \text{ dB} \tag{4.29}$$

The sky noise temperature resulting from a path attenuation A_{total} dB is found from the output noise model of Section 4.3 using an assumed medium temperature of 270 K for the rain.

$$T_{\text{sky}} = 270 \times (1 - 10^{-A/10}) \text{ K} \tag{4.30}$$

The antenna noise temperature may be assumed to be equal to the sky noise temperature, although in practice not all of the incident noise energy from the sky is output by the antenna, and a coupling coefficient, $\eta_{c,}$ of 90–95% is sometimes used when calculating antenna noise temperature in rain. Thus antenna noise temperature may be calculated as

$$T_{\text{A}} = \eta_{\text{c}} \times T_{\text{sky}} \text{ K} \tag{4.31}$$

Almost all satellite receivers use a high gain LNA as the first element in the receiver front end. This makes the contribution of all later parts of the receiver to the system noise temperature negligible. System noise temperature is then given by $T_{\text{s rain}}$ where

$$T_{\text{s rain}} = T_{\text{LNA}} + T_{\text{A rain}} \text{ K} \tag{4.32}$$

In Eq. (4.32), the LNA is assumed to be placed right at the feed horn so that there is no waveguide or coaxial cable run between the feed horn of the antenna and the LNA. We will assume that there are no feed losses. The increase in noise power, ΔN_{rain} dB, caused by the increase in sky noise temperature is given by

$$\Delta N_{\text{rain}} = 10 \log_{10} \left[\frac{kT_{\text{s rain}} B_{\text{n}}}{kT_{\text{sca}} B_{\text{n}}} \right] = 10 \log_{10} \left[\frac{T_{\text{s rain}}}{T_{\text{sca}}} \right] \text{ dB} \tag{4.33}$$

where T_{sca} is the system noise temperature in clear sky conditions.

The received power is reduced by the attenuation caused by the rain in the slant path, so in rain the value of carrier power is reduced from C_{ca} to C_{rain} where

$$C_{\text{rain}} = C_{\text{ca}} - A_{\text{rain}} \text{ dB} \tag{4.34}$$

The resulting $(\text{CNR})_{\text{dn rain}}$ value when rain intersects the downlink is given by

$$(\text{CNR})_{\text{dn rain}} = (\text{CNR})_{\text{dn ca}} - A_{\text{rain}} - \Delta N_{\text{rain}} \text{ dB} \tag{4.35}$$

where $(\text{CNR})_{\text{dn ca}}$ is the downlink CNR in clear sky conditions.

If a linear (bent pipe) transponder is used, the $(\text{CNR})_{\text{up}}$ must be combined with $(\text{CNR})_{\text{dn rain}}$ to yield the overall $(\text{CNR})_o$ ratio for the link. However, the transmitting earth stations for DBS-TV service use large antennas and high power transmitters with

uplink power control (UPC) to ensure that the uplink CNR is always at least 20 dB higher than the downlink CNR. The contribution of satellite noise to overall CNR can be ignored when this is the case. Some digital systems use regenerative transponders that provide constant output power regardless of uplink attenuation provided that the received CNR at the satellite is above the threshold of the onboard processing demodulator. In this case the value of $(\text{CNR})_{\text{dn rain}}$ will be used as the overall $(\text{CNR})_{\text{o}}$ value in rain for the link.

Example 4.9 Calculation of Rain Attenuation Margin

In the example of a DBS-TV system in Table 4.6a, a link margin of 5.7 dB is available before the $(\text{CNR})_{\text{o}}$ threshold of 8.3 dB is reached. This example shows how the link margin can be distributed between rain attenuation and an increase in receiver system noise power caused by an increase in sky noise. It is very difficult to write a set of equations that solve this problem, because the combination of linear and decibel arithmetic leads to a transcendental equation. Instead, an iterative calculation is used to find the exact rain attenuation, which causes the receiver CNR to equal the threshold value. We will begin by calculating the increase in system noise temperature that results from an estimated 3 dB rain attenuation in the downlink path to determine the increase in noise power and thus the value of $(\text{CNR})_{\text{dn rain}}$.

The clear sky attenuation is given in Table 4.6a as 0.4 dB. This must be added to the rain attenuation – the atmosphere does not go away when it rains! Thus total path attenuation is 3.4 dB, and the sky noise temperature in rain will be, from Eq. (4.30)

$$T_{\text{sky rain}} = 270 \times (1 - 10^{-3.4/10}) = 147 \text{ K}$$

We will assume 100% coupling between the sky noise temperature and the antenna temperature in this example. In clear sky conditions the sky noise temperature is

$$T_{\text{ca}} = 270 \times (1 - 10^{-0.04}) = 24 \text{ K}$$

The sky temperature has increased from 24 K in clear sky conditions to 147 K when 3 dB rain attenuation occurs in the downlink. We must calculate the new system noise temperature when rain is present in the slant path. The LNA of the system in Table 4.6a has a noise temperature of 86 K and the clear sky system noise temperature is $T_{\text{sky ca}} = 110$ K. With 3 dB of rain attenuation in the downlink, the system noise temperature, given by Eq. (4.32), increases to

$$T_{\text{s rain}} = T_{\text{LNA}} + T_{\text{A}} = 86 + 147 = 233 \text{ K}$$

The increase in receiver noise power referred to the receiver input is given by Eq. (4.33)

$$\Delta N_{\text{rain}} = 10 \log_{10}(233/110) = 3.3 \text{ dB}$$

From Eq. (4.35)

$$(\text{CNR})_{\text{dn rain}} = 14.0 - 3.0 - 3.3 = 7.8 \text{ dB}$$

The receiver CNR is below the threshold value of 8.3 dB by 0.5 dB, indicating that the maximum downlink rain attenuation is less than the estimated 3.0 dB. The previous procedure needs to be repeated with a revised estimate of the rain attenuation until the closest value of rain attenuation that gives 8.3 dB for $(\text{CNR})_{\text{dn rain}}$ is found; that value is 2.7 dB in this case. In the southeastern United States, in states such as Florida and

Louisiana where very heavy rain occurs more often than in other parts of the United States, link availability may be slightly less than 99.7%. Shaping of the satellite beam to direct more power to these parts of the United States helps to reduce the number of outages experienced in that region.

Receivers located within the −1 dB contour of the satellite antenna beam have shorter path lengths giving 2.3 dB higher CNR than the receiver used in the example shown in Tables 4.6a and 4.6b, so they have a clear air downlink CNR of 16.3 dB and a downlink margin of 8.0 dB. The calculation of the availability of these receivers requires some care, because we cannot just add the extra 2.3 dB of link margin to the rain attenuation. Antenna noise increases with every decibel of extra rain attenuation, reducing the received power level, C, and increasing the system noise power N. The iterative procedure must be used again to find the combination of reduction in C and increase in N that leads to an additional 2.3 dB degradation in the overall CNR value.

We will guess that increasing the rain attenuation from 3 to 4.1 dB gives the required result. With 4.1 dB rain attenuation, the path attenuation is 4.5 dB and system noise temperature is

$$T_{\text{s rain}} = 86 + 270(1 - 0.355) = 260 \text{ K}$$

The increase in noise power from the clear sky condition is

$$\Delta N = 10 \log_{10}\left(\frac{260}{110}\right) = 3.7 \text{ dB}$$

Hence the decrease in $(\text{CNR})_{\text{dn rain}}$ for 4.1 dB of rain attenuation is 4.1 + 3.7 = 7.8 dB, and

$$\text{CNR}_{\text{dn rain}} = 16.3 - 7.8 = 8.5 \text{ dB}$$

so we are above the 8.3 dB threshold by 0.2 dB. Another trial is needed to determine the exact downlink attenuation that reduces the signal to the threshold level. A rain attenuation margin of 4.1 dB at Ku-band would give an availability of 99.85% or better over the central region of the United States. This example demonstrates that the increase in noise temperature of a low noise DBS-TV Ku-band receiving system is a significant factor when rain attenuation is present in the downlink path. Rain attenuation alone cannot be equated to link margin.

It is worth noting that rain causes significant attenuation on Ku-band downlinks for less than 2% of an average year over most of the United States. For 98% of the year these links are operating in near clear sky conditions with downlink CNR values in the 14–17 dB range. With half rate FEC applied to the data stream, the bit error rate will be below 10^{-16}, and the error rate for a live MPEG-2 compressed video stream will be effectively zero. Thus the video transmission is essentially error free for all but a few tens of hours each year.

4.6 Uplink Design

The design of the uplink is easier than the downlink in many cases, since an accurately specified carrier power must be presented at the satellite transponder and it is often feasible to use much higher power transmitters at earth stations than can be used on a satellite, as illustrated in the analysis of DBS-TV in the previous section. However, VSAT systems use earth stations with small antennas and transmitter powers below

5 W, giving low uplink EIRP. Satellite telephone handsets are restricted to transmitting at power levels below one half watt because of the risk of EM radiation hazards to the user. In mobile systems the link from the satellite telephone to the gateway earth station is often the link with the lowest CNR.

The cost of transmitters tends to be high compared with the cost of receiving equipment in satellite communication systems. The major growth in satellite communications has been in point-to-multipoint transmission, as in cable TV distribution and DBS-TV. One high power gateway earth station provides service via a DBS-TV satellite to many low-cost receive-only stations, and the high cost of the transmitting station is only a small part of the total network cost.

The typical bent pipe satellite transponder is a quasilinear amplifier and the received carrier level determines the output level. Where a traveling wave tube is used as the output high-power amplifier (HPA) in the transponder, as is often the case, and FDMA is employed, the HPA must be run with a predetermined backoff to avoid intermodulation (IM) products appearing at the output. The output backoff is typically 1–3 dB when more than one signal is present in the transponder, and is determined by the uplink carrier power level received at the spacecraft. Accurate control of the power transmitted by the earth station is therefore essential, which is easily achieved in a fixed network of earth stations. Where a very large number of earth stations access a single transponder using FDMA, such as in some VSAT networks, transponder output backoff of 5–7 dB may be required to maintain intermodulation products at a sufficiently low level (Maral and Bousquet 2002, p. 426). Even with a single access to the transponder (i.e., only one carrier present) some backoff is normally applied to avoid the PM-AM conversion that occurs when modulated signals are transmitted through a non-linear device (Glover and Grant 1998, p. 250).

Earth station transmitter power is set by the power level required at the input to the transponder. This can be done in one of two ways. Either a specific flux density is required at the satellite, or a specific power level is required at the input to the transponder. Early Intelsat C-band satellites required high flux densities to saturate their transponders, in the range −73.7 to −67.5 dBW/m^2, depending on the transponder gain setting (Dicks and Brown 1975, pp. 73–103). This is a high flux density, which requires a large earth station and a powerful transmitter generating up to 3 kW. Domestic GEO satellites generally require lower flux densities allowing the use of smaller earth station antennas. At C-band, a typical uplink earth station transmits a maximum of 100 W with a 5–9 m antenna, giving a peak flux density at the satellite of −100 dBW m-2.

Although flux density at the satellite is a convenient way to determine earth station transmit EIRP requirements, analysis of the uplink requires calculation of the power level at the input to the transponder so that the uplink CNR can be found. The link equation is used to make this calculation, using either a specified transponder CNR or a required transponder output power level. When a CNR is specified for the transponder, the calculation of required transmit power is straightforward. Let $(\text{CNR})_{\text{up}}$ be the specified CNR in the transponder, measured in a noise bandwidth B_{n} Hz. The bandwidth B_{n} Hz is the bandwidth of the bandpass filter in the IF stage of the earth station receiver for which the uplink signal is intended. Even if B_{n} is much less than the transponder bandwidth, it is important that the uplink CNR be calculated in the bandwidth of the receiver, not the bandwidth of the transponder. Repeating the equations from Section 4.2, the noise power referred to the transponder input is N_{xp} watts. In dB units

$$N_{\text{xp}} = k + T_{\text{xp}} + B_{\text{n}} \text{ dBW} \tag{4.36}$$

where T_{xp} is the system noise temperature of the transponder in dBK and B_n is in units of dBHz.

The power received at the input to the transponder is P_{rxp} where

$$P_{rxp} = P_t + G_t + G_r - L_p - L_{up} \text{ dBW} \tag{4.37}$$

where $P_t + G_t$ is the uplink earth station EIRP in dBW, G_r is the satellite antenna gain in dB in the direction of the uplink earth station and L_p is the path loss in dB. The factor L_{up} accounts for all uplink losses other than path loss. The value of $(CNR)_{up}$ at the LNA input of the satellite receiver is given by

$$(CNR)_{up} = 10 \log_{10} \left(P_r / k\, T_s B_n\right) = P_{rxp} - N_{xp} \text{ dB} \tag{4.38}$$

The earth station transmitter output power P_t is calculated from Eq. (4.23) using the given value of $(CNR)_{up}$ in Eq. (4.38) and the noise power N_{xp} calculated from Eq. (4.36). Note that the received power at the transponder input is also given by

$$P_{rxp} = N_{xp} + (CNR)_{up} \text{ dBW} \tag{4.39}$$

The earth station transmitter output power P_t can also be calculated from the output power of the transponder and transponder gain when these parameters are known and a bent pipe transponder is used. In general

$$P_{rxp} = P_{sat} - BO_o - G_{xp} \text{ dBW} \tag{4.40}$$

where P_{sat} is the saturated power output of the transponder in dBW, BO_o is the output backoff in dB, and G_{xp} is the gain of the transponder in dB.

With small-diameter earth stations, a higher power earth station transmitter is required to achieve a similar satellite EIRP. This has the disadvantage that the interference level at adjacent satellites rises, since the small earth station antenna inevitably has a wider beam. Thus it is not always possible to trade off transmitter power against uplink antenna size. Early GEO satellites were spaced 3° apart in the GEO orbit to minimize interference from adjacent uplinks. The uplink interference problem determines satellite spacing and limits the capacity of the geostationary orbit in any frequency band. The ITU has established a set of specifications for the antenna pattern of antennas transmitting to GEO satellites designed to minimize interference between adjacent satellites. To increase the capacity of the crowded geostationary orbit arc south of the United States, the FCC introduced new regulations in 1983 requiring better control of 6 GHz earth station antenna transmit patterns so that inter-satellite spacing could be reduced to 2°. The same specification was adopted by the ITU-R for the entire geostationary arc (Recommendation ITU-R S 465.5 1993). The requirement was for the transmit antenna pattern to lie below $G(\theta) = 29 - 25 \log_{10} \theta$ dB in the range $1° < \theta < 7°$ from the antenna boresight and $G(\theta) = 32 - 25 \log_{10} \theta$ dB beyond 7°. This specification was later extended to the antenna beam profile illustrated in Table 4.7 and Figure 4.12, with additional specifications for other frequency bands and satellite communication systems (Recommendation ITU-R S.524-6 2000). Further restrictions on antenna patterns have also been recommended by the ITU. See, for example, Recommendation ITU-R S.1328-3 (2001) and ITU Radio Regulations (2001).

An example of one specification is shown in Figure 4.12 for the uplink antennas in the 6 GHz fixed service uplink band in the form of an envelope that the sidelobes of

Table 4.7 ITU recommendation for uplink antenna pattern envelope at 6 GHz

Angle off-axis	Maximum EIRP per 4 kHz
$2.5° \leq \theta \leq 7°$	$(32 - 25 \log \theta)$ dB(W/4 kHz)
$7° < \theta \leq 9.2°$	11 dB(W/4 kHz)
$9.2° < \theta \leq 48°$	$(35 - 25 \log \theta)$ dB(W/4 kHz)
$48° < \theta \leq 180°$	−7 dB(W/4 kHz)

the antenna must not exceed. Other specifications exist for the Ku-band and Ka-band uplinks, and also for VSAT systems. The specifications are written in terms of the maximum permissible EIRP of the uplink antenna in any 4 kHz bandwidth within the transmitted signal.

At frequencies above 10 GHz, for example, 14.6 and 30 GHz, propagation disturbances in the form of fading in rain cause the received power level at the satellite to fall. This lowers the uplink CNR in the transponder, which lowers the overall $(CNR)_o$ ratio in the earth station receiver when a linear (bent pipe) transponder is used on the satellite. Uplink power control (UPC) can be used to combat uplink rain attenuation. The transmitting earth station monitors a beacon signal from the satellite, and watches for reductions in power indicating rain fading on the uplink and downlink. Automatic monitoring and control of transmitted uplink power is used in 14 and 17 GHz uplink earth stations to maintain the uplink CNR in the satellite transponder during up periods of rain attenuation. New generations of Ka-band satellites employ uplink power level detection at the satellite. A control link to each uplink earth station closes the loop.

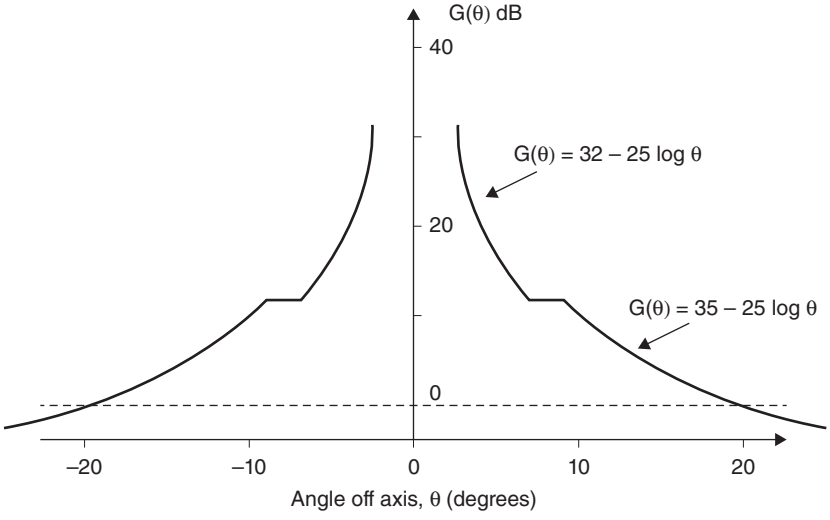


Figure 4.12 ITU specification for 6 GHz fixed service uplink antenna beam profile. $G(\Theta)$ is the gain of the antenna as a function of the angle Θ away from the axis (direction of maximum gain). Similar specifications apply to Ku-band and Ka-band uplink antenna patterns (Recommendation ITU-R S.524-6 2000). Source: Reproduced with permission of ITU.

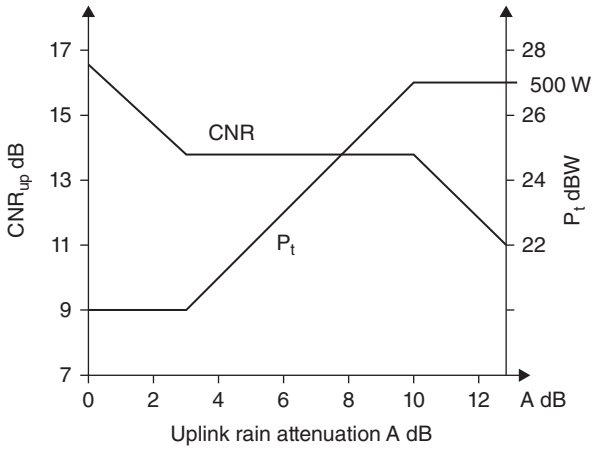


Figure 4.13 Example of uplink power control with 7 dB dynamic range. In this example uplink power control is implemented when the attenuation on the uplink path is estimated to exceed 3 dB. The increase in transmit power matches the uplink path attenuation dB for dB, holding CNR in the satellite transponder constant over the attenuation range 3–10 dB. CNR: Carrier to noise ratio in transponder; P_t: Transmit power in watts or dBW; UPC: Uplink power control.

Since the downlink is always at a different frequency from the uplink, a downlink attenuation of A dB must be scaled to estimate uplink attenuation. The scaling factor used is typically $(f_{up}/f_{down})^a$ where the exponent a is typically between 2.0 and 2.4. For example, an uplink station transmitting at 14.0 GHz to a Ku-band satellite monitors the satellite beacon at 11.45 GHz. The uplink attenuation is therefore given by

$$A_{up} = A_{down} \times \left(f_{up}/f_{down}\right)^a \text{ dB} \tag{4.41}$$

where A_{up} is the estimated uplink rain attenuation and A_{down} is the measured downlink rain attenuation. For a value of $a = 2.2$ and $(f_{up}/f_{down}) = 1.222$, the factor $(f_{up}/f_{down})^a$ is 1.56. Hence a downlink rain attenuation of 3 dB would give an estimated uplink attenuation of 4.7 dB. This uplink attenuation value applies only to rain and does not include gaseous attenuation or tropospheric scintillation, which require different scaling ratios. (See Chapter 7 for details.)

UPC cannot be applied until a certain amount of attenuation has built up in the link. This is typically around 2 dB for the downlink due to measurement inaccuracies, corresponding to about 3 dB for a Ku-band uplink. As rain begins to affect the link between the earth station and satellite, the uplink CNR in the transponder will fall until UPC starts to operate in the earth station transmitter. The transponder CNR will then remain relatively constant until the UPC system reaches the maximum available transmit power. Further attenuation on the uplink will cause the CNR in the transponder to fall. Figure 4.13 shows an example of UPC applied over a range of 7 dB.

It is easy to make errors when calculating received power and noise power in link budgets, especially when using decibel values. Software for the calculation of link budgets can be helpful, and a spreadsheet such as Excel® can be developed to make the calculations. There is a limited range of values for many of the parameters in satellite communication systems. Knowing that range is helpful in spotting errors, especially when software is used to calculate link budgets. When using link budget software, a test calculation using the methods outlined in this chapter should be made by hand to check that the spreadsheet or software is working correctly. Here are some of the ranges of expected values in typical satellite communication systems.

Power transmitted by a satellite. Range is from 100 mW to 1 kW, or −10 to 30 dBW. Values above 30 dBW are unlikely unless the satellite generates an unusually large amount of electrical power. The latest generation of radio broadcasting satellites can generate 2.7 kW of transmit power.

Satellite antenna gain. Lowest gain is 0 dB for a LEO satellite and highest gain is around 64 dB for a GEO satellite with a 5 m antenna transmitting at 30 GHz.

Path loss. For a 30 GHz transmission at a distance of 40 000 km, path loss is 214 dB. A LEO satellite at an altitude of 500 km operating at 1500 MHz has a path loss of 150 dB when directly above the earth station. Values outside this range are unlikely unless the spacecraft is in deep space. A common error is to use kilometers instead of meters for distance, resulting in a 60 dB error in path loss.

Noise bandwidth. Lowest bandwidth is for a single digital data or voice signal at about 5 kHz, although data packets can be transmitted at lower rates. Widest bandwidth for a single high speed digital data signal is unlikely to exceed 500 MHz, giving noise bandwidths in the range 37–87 dBHz.

Receiving system noise power. Earth station noise temperatures for satellite communication systems are in the range 25 K at L and C-band to 500 K at Ka- and V- bands, giving decibel values in the range 14–27 dBK. Mobile terminals with omnidirectional antennas have higher noise temperatures because noise radiated by the earth enters the antenna. Combining bandwidth and noise temperature ranges results in earth station receiver noise powers, which are unlikely to fall outside the range −175 to −121 dBW. Note that the noise temperature of the receiver on a satellite is always higher than 300 K because the satellite antenna looks at the earth, which is a relatively hot body compared to the sky as seen by a directional earth station antenna.

Earth station transmitter powers. A single channel hand held transmitter can operate at power levels as low as 100 mW. Large earth stations with high power transmitters can transmit 10 kW, giving a range of transmit power from −10 to 40 dBW.

4.7 Design for Specified CNR: Combining CNR and C/I Values in Satellite Links

The BER or SNR in the baseband channel of an earth station receiver is determined by the ratio of the carrier power to the noise power in the IF amplifier at the input to the demodulator. The noise present in the IF amplifier comes from many sources. So far in our analysis of uplinks and downlinks we have considered only the receiver thermal noise and noise radiated by atmospheric gases and rain in the slant path. When a complete satellite link is engineered, the noise in the earth station IF amplifier will have contributions from the receiver itself, the receiving antenna, sky noise, the satellite transponder from which it receives the signal, and interference from adjacent satellites and terrestrial transmitters which share the same frequency band.

4.7.1 Combining Uplink and Downlink Carrier to Noise Ratios

When more than one CNR is present in the link, we can add the individual CNRs reciprocally to obtain an *overall CNR*, which we will denote here as $(\text{CNR})_o$. The overall

$(CNR)_o$ is what is measured in the earth station at the output of the IF amplifier and SRRC filter, and is given by

$$(CNR)_o = \frac{1}{1/(CNR)_1 + 1/(CNR)_2 + 1/(CNR)_3 + \cdots} \tag{4.42}$$

This is sometimes referred to as the *reciprocal CNR formula*. The CNR values must be linear ratios, NOT decibel values. Since the noise power in the individual CNRs is referenced to the carrier power at that point, all the C values in Eq. (4.42) are the same. Expanding the formula by cross multiplying gives the overall $(CNR)_o$ as a power ratio, not in decibels

$$(CNR)_o = 1/(N_1/C + N_2/C + N_3/C + \cdots) = C/(N_1 + N_2 + N_3 + \cdots) \tag{4.43}$$

In decibel units:

$$(CNR)_o = C \text{ dBW} - 10 \log_{10} (N_1 + N_2 + N_3 + \cdots \text{ watts}) \text{ dB} \tag{4.44}$$

Note that $(CNR)_{dn}$ cannot be measured at the receiving earth station. The satellite always transmits noise as well as signal, so a CNR measurement at the receiver will always yield $(CNR)_o$, the combination of transponder and earth station CNRs.

To calculate the performance of a satellite link we must determine the uplink $(CNR)_{up}$ ratio in the transponder and the downlink $(CNR)_{dn}$ in the earth station receiver. We must also consider whether there is any interference present, either in the satellite receiver or the earth station receiver. One case of importance is where the transponder is operated in an FDMA mode and *intermodulation products* (IM) are generated by the transponder's non-linear input-output characteristic. If the IM power level in the transponder is known, a C/I value can be found and included in the calculation of $(CNR)_o$ ratio. Interference from adjacent satellites is likely whenever small receiving antennas are used, as with VSATs and DBS-TV receivers. See Chapter 6 for more details on intermodulation in FDMA.

Since CNR values are usually calculated from power and noise budgets, their values are typically in decibels. There are some useful rules of thumb for estimating $(CNR)_o$ from two CNR values:

- If the CNR values are equal, $(CNR)_o$ is 3 dB lower than either value.
- If one CNR value is 10 dB smaller than the other value, $(CNR)_o$ is 0.4 dB lower than the smaller of the CNR values.
- If one CNR value is 20 dB or more greater than the other CNR value, the overall $(CNR)_o$ is equal to the smaller of the two CNR values within the accuracy of decibel calculations (±0.1 dB).

Example 4.10

Thermal noise in an earth station receiver results in a $(CNR)_{dn}$ ratio of 20.0 dB.

A signal is received from a bent pipe transponder with $(CNR)_{up} = 20.0$ dB.

a. What is the value of overall $(CNR)_o$ at the earth station?
b. If the transponder introduces intermodulation products with a carrier to interference ratio C/I = 24 dB, what is the overall $(CNR)_o$ at the receiving earth station?

Answer

a. Using Eq. (4.42) and noting that $(\text{CNR})_{\text{up}} = 20.0$ dB corresponds to a CNR ratio of 100

$$(\text{CNR})_{\text{o}} = \frac{1}{1/(\text{CNR})_{\text{up}} + 1/(\text{CNR})_{\text{dn}}} = \left[\frac{1}{0.01 + 0.01}\right] = 50 \text{ or } 17.0 \text{ dB}$$

b. The intermodulation (C/I) value of 24.0 dB corresponds to a ratio of 250. The overall $(\text{CNR})_{\text{o}}$ value with interference present is then

$$(\text{CNR})_{\text{o}} = \frac{1}{(0.01 + 0.01 + 0.004)} = 41.7 \text{ or } 16.2 \text{ dB}$$

4.7.2 Overall $(\text{CNR})_{\text{o}}$ With Uplink and Downlink Attenuation

Most satellite links are designed with link margins to allow for attenuation that may occur in the link or increases in noise power caused by interference. Interference is often treated as though it were white noise, regardless of whether the interfering signal actually has a uniform power spectral distribution or Gaussian statistics. When the interference has known characteristics, such as a depolarized co-channel or a coherent jamming signal, cancelation techniques can be used to reduce the level of interference. Noise jamming cannot be canceled.

The effect of a change in the uplink CNR has a different impact on overall $(\text{CNR})_{\text{o}}$ depending on the operating mode and gain of the transponder.

There are three different transponder types or operating modes:

$$\begin{aligned}
&\text{Linear transponder: } P_{\text{out}} = P_{\text{in}} + G_{\text{xp}} \text{ dBW} \\
&\text{Non-linear transponder: } \text{P}_{\text{out}} = P_{\text{in}} + G_{\text{xp}} - \Delta G \text{ dBW} \\
&\text{Regenerative transponder: } \text{P}_{\text{out}} = \text{constant dBW}
\end{aligned} \tag{4.45}$$

where P_{in} is the power delivered by the satellite's receiving antenna to the input of the transponder, P_{out} is the power delivered by the transponder HPA to the input of the satellite's transmitting antenna, G_{xp} is the linear gain of the transponder, and all parameters are in decibel units. The parameter ΔG is dependent on P_{in} and accounts for the loss of gain caused by the non-linear saturation characteristics of a transponder, which is driven hard to obtain close to its maximum output power – the gain is effectively falling as the input power level increases. (See Chapter 6 for a detailed discussion of intermodulation effects and non-linearity in transponders.)

The maximum output power from a transponder is called the *saturated output power* and is the nominal transponder power output rating that is usually quoted. The transponder input-output characteristic is highly non-linear when operated at this output power level. When a transponder is operated close to its saturated output power level, digital waveforms are changed, resulting in *intersymbol interference* (ISI), and FDMA operation results in the generation of intermodulation products by multiplication of the individual signals. Transponders are usually operated with *output backoff*, to make the characteristic more nearly linear. The exact amount of output backoff required in any given application depends on the specific characteristics of the transponder and the signals it carries. Typical values of output backoff are 1 dB for a single PSK carrier to 3 dB for FDMA operation with several carriers. The corresponding input backoff values might be 3 and 5 dB, but the individual transponder characteristic

must be known to make an accurate assessment. For convenience in this text, we will frequently assume linear transponder operation when calculating the overall $(CNR)_o$ ratio, even if this may not, in fact, be the case.

Onboard processing satellites have transponders that regenerate the uplink signal for transmission on the downlink and therefore have constant transmit power. The bit errors rates in the transponder (calculated from $(CNR)_{up}$) and the earth station (calculated from $(CNR)_{dn}$) add together to give an overall BER. Frequently, $(CNR)_{up}$ is much higher than $(CNR)_{dn}$ so overall bit error rate is dominated by errors occurring at the receiving earth station.

4.7.3 Uplink and Downlink Attenuation in Rain

Rain attenuation affects uplinks and downlinks differently. We usually assume that rain attenuation is occurring on either the uplink or the downlink, but not on both at the same time. This is usually true for earth stations that are well separated geographically, but not if they are close together (<20 km). Heavy rain occurs with a somewhat random geographic distribution for less than 1% of the time, so the probability of significant attenuation occurring on both the uplink and downlink simultaneously is small when the transmitting and receiving earth stations are separated by more than 20 km. In the following analysis of uplink and downlink attenuation effects, it will be assumed that one link is attenuated and the other is operating in clear air.

4.7.3.1 Uplink Attenuation and $(CNR)_{up}$

The transponder receiver noise temperature does not change significantly when rain is present in the uplink path to the satellite. The satellite receiving antenna beam is always sufficiently wide that it "sees" a large area of the (warm) earth's surface and local noise temperature variations due to thunderstorm clouds, for example, are insignificant. The noise temperature of the earth seen by a GEO satellite varies according to latitude, with the tropics significantly warmer than northern latitudes (NASA earth observations 2013). The corresponding system noise temperature for transponders on a GEO satellite in C-band through V-band is in the range 300–800 K, the higher values applying to the higher frequency bands. There is effectively no increase in uplink noise power when heavy rain is present in the uplink because of averaging of radiation temperature over the footprint of the satellite antenna.

Rain attenuation on the uplink path to the satellite reduces the power at the satellite receiver input, and thus reduces $(CNR)_{up}$ in direct proportion to the attenuation on the slant path. If the transponder is operating in a linear mode, the output power will be reduced by the same amount, which will cause $(CNR)_{dn}$ to fall by an amount equal to the attenuation on the uplink. When both $(CNR)_{up}$ and $(CNR)_{dn}$ are reduced by A_{up} dB, the value of $(CNR)_o$ is reduced by exactly the same amount, A_{up} dB. Hence for the case of a linear transponder and rain attenuation in the uplink of A_{up} dB

$$(CNR)_{o\ uplink\ rain} = (CNR)_{o\ clear\ air} - A_{up}\ \text{dB Linear transponder} \tag{4.46}$$

If the transponder is non-linear, the reduction in input power caused by uplink attenuation of A_{up} dB results in a smaller reduction in output power, by an amount ΔG.

$$(CNR)_{o\ uplink\ rain} = (CNR)_{o\ clear\ air} - A_{up} + \Delta G\ \text{dB Non-linear transponder} \tag{4.47}$$

If the transponder is digital and regenerative, or incorporates an automatic gain control (AGC) system to maintain a constant output power level

$$(\text{CNR})_{\text{o uplink rain}} = (\text{CNR})_{\text{o clear air}} \text{ dB Regenerative transponder or AGC} \tag{4.48}$$

The above equation will hold only if the received signal is above threshold and the BER of the recovered digital signal in the transponder is small. If the signal falls below threshold, the uplink will contribute significantly to the BER of the digital signal at the receiving earth station. In a link that employs a regenerative transponder, the BERs in the transponder and the earth station add. This results in a lower overall BER than with a linear transponder, because the uplink and downlink CNRs are always higher than the overall CNR. Typically, with a large uplink earth station the downlink CNR will be significantly lower than the uplink CNR, so most of the bit errors on the link will occur in the earth station receiver.

4.7.3.2 Downlink Attenuation and $(\text{CNR})_{\text{dn}}$

The earth station receiver noise temperature can change very significantly when rain is present in the downlink path from the satellite. The sky noise temperature can increase to a value close to the physical temperature of the individual raindrops, particularly in very heavy rain. A reasonable temperature to assume for temperate latitudes in a variety of rainfall rates is 270 K, although values above 290 K have been observed in the tropics. An increase in sky noise temperature to 270 K will increase the receiving antenna temperature markedly above its clear air value. See Section 4.5 for an illustration of this effect. The result is that the received power level, C, is reduced and the noise power, N, in the receiver increases. The result for downlink CNR is given by Eq. (4.35), repeated here

$$(\text{CNR})_{\text{dn rain}} = (\text{CNR})_{\text{dn ca}} - A_{\text{rain}} - \Delta N_{\text{rain}} \text{ dB} \tag{4.49}$$

The overall CNR is then given by Eq. (4.42), using linear values for CNRs, not decibels

$$(\text{CNR})_{\text{o}} = \frac{1}{1/(\text{CNR})_{\text{up}} + 1/(\text{CNR})_{\text{dn}}} \tag{4.50}$$

As noted earlier, unless we are making a loop-back test, we will assume that the value of $(\text{CNR})_{\text{up}}$ is for clear air, and remains constant regardless of the attenuation on the downlink.

4.8 System Design for Specific Performance

A typical two-way satellite communication link consists of four separate paths: an outbound uplink path from one terminal to the satellite and an outbound downlink to the second terminal, and an inbound uplink from the second terminal to the satellite and an inbound downlink to the first terminal. The links in the two directions are independent and can be designed separately, unless they share the same transponder using FDMA. A broadcast link, like the DBS-TV system described earlier in this chapter, is a one-way system, with just one uplink and one downlink.

4.8.1 Satellite Communication Link Design Procedure

The design procedure for a one-way satellite communication link can be summarized by the following 10 steps. The return link design follows the same procedure.

1. Determine the frequency band in which the system must operate. Comparative designs may be required to help make the selection.
2. Determine the communications parameters of the satellite. Estimate any values that are not known.
3. Determine the parameters of the transmitting and receiving earth stations.
4. Start at the transmitting earth station. Establish an uplink power budget and a transponder noise power budget to find $(CNR)_{up}$ in the transponder.
5. Find the output power of the transponder based on transponder gain or output backoff.
6. Establish a downlink power and noise budget for the receiving earth station. Calculate $(CNR)_{dn}$ and $(CNR)_o$ for a station at the edge of the coverage zone (worst case).
7. Calculate SNR or BER in the baseband channel. Find the link margins.
8. Evaluate the result and compare with the specification requirements. Change parameters of the system as required to obtain acceptable $(CNR)_o$ or SNR or BER values. This may require several trial designs.
9. Determine the propagation conditions under which the link must operate. Calculate outage times for the uplinks and downlinks.
10. Redesign the system by changing some parameters if the link margins are inadequate. Check that all parameters are reasonable, and that the design can be implemented within the expected budget.

The following sample system designs demonstrate how the ideas developed in this chapter can be applied to the design of satellite communication systems.

Example 4.11 System Design

This example examines the design of a satellite communication link using a Ku-band geostationary satellite with bent-pipe transponders to distribute digital TV signals from an uplink earth station to many receiving stations throughout the United States. The design requires a clear sky overall CNR of 17.0 dB at the receiving earth station, which will result in no bit errors in the received digital video signal when FEC encoding is applied to the transmitted signal. The minimum allowed overall CNR is 8.0 dB to ensure that the BER in the earth station receiver does not exceed 10^{-6}. The transmission method is QPSK with 6 dB FEC coding gain, using the DVB-S standard. (See Chapter 5 for details of BER in digital links.) The uplink transmitter power and the receiving antenna gain and diameter to meet the specification are determined. The available link margins for each of the systems are found and the performance of the systems is analyzed when rain attenuation occurs in the satellite-earth paths. The advantages and disadvantages of implementing UPC are considered. Since the uplink station will be distributing TV signals to hundreds of cable TV head ends, the probability of an outage on the uplink must be made very small. Figure 4.14 shows an illustration of the satellite video distribution system.

In this example, the satellite is located at 73°W and has 28 Ku-band transponders. However, for international registration of this satellite location, the location would be

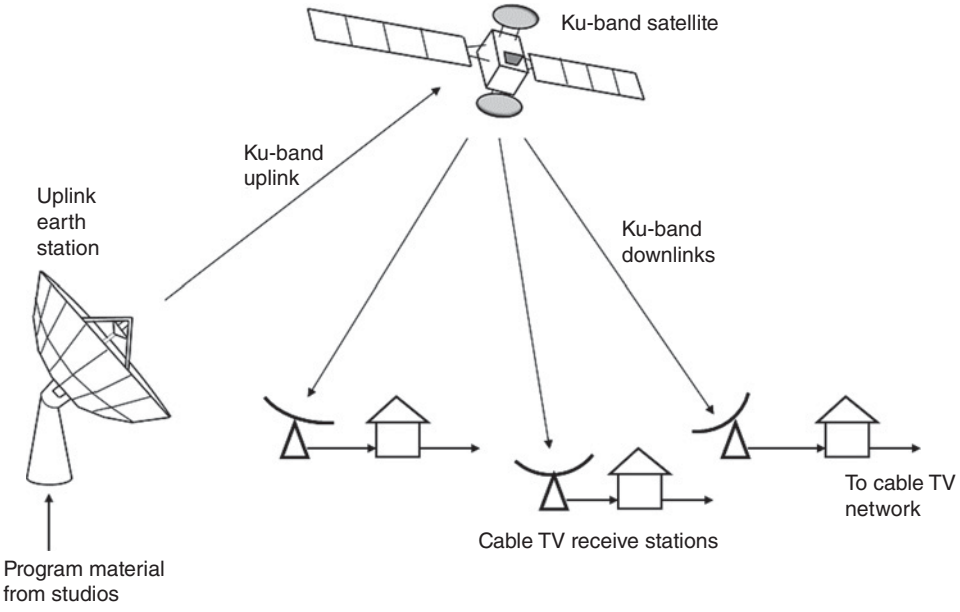


Figure 4.14 Illustration of video distribution system supplying cable TV signals via a GEO satellite. Satellite distribution of TV programming to cable TV systems is widely employed because a single uplink earth station and GEO satellite can send hundreds of TV channels to every cable TV system in an entire continent.

denoted as 287°E. The link budgets developed in the examples below use decibel notation throughout. The satellite and earth stations are specified in Table 4.8a and the propagation conditions in Table 4.8b.

4.8.1.1 Ku-Band Uplink Design

We must first calculate uplink antenna gain and path loss. The uplink antenna has a diameter of 5.0 m and an aperture efficiency of 68%. At 14.0 GHz the wavelength is 2.120 cm = 0.0212 m. The antenna gain is

$$G_t = 10 \log_{10} [0.68 \times (\pi D/\lambda)^2] = 55.7 \text{ dB}$$

Assuming a distance to the satellite of 38 500 km, the free space path loss is

$$L_p = 10 \log_{10} [(4\pi R/\lambda)^2] = 207.2 \text{ dB}$$

We will next calculate the noise power in the transponder for 43.2 MHz bandwidth, and then add the uplink CNR value of 27.0 dB to find the transponder input power level.

$$N_{xp} = -228.6 + 27.0 + 76.4 = -125.2 \text{ dBW}$$

Table 4.8c sets out these calculations as an uplink carrier and noise power budget.

The received power P_r at the transponder input must be 27.0 dB greater than the noise power.

$$P_r = -125.2 + 27.0 = -98.2 \text{ dBW}$$

Table 4.8a Satellite and earth station specification

Ku-band satellite		
Antenna gain, on axis, (transmit and receive)	G_t, G_r	31 dB
Transponder system noise temperature	$T_{s\,sat}$	500 K
Transponder saturated output power	$P_{t\,sat}$	80 W
Transponder bandwidth:	B_{transp}	54 MHz
Signal		
Compressed digital video signals: symbol rate	R_s	43.2 Msps
Minimum permitted overall CNR in receiver	$(CNR)_o$	8.0 dB
Uplink earth station		
Antenna diameter	D	5.0 m
Transmitting antenna aperture efficiency	η_A	68%
Uplink frequency	f_{up}	14.15 GHz
Required CNR in Ku-band transponder	$(CNR)_{xp}$	27 dB
Transponder HPA output backoff	$B_{o\,xp}$	1.0 dB
Miscellaneous uplink losses	$L_{misc\,up}$	1.0 dB
Location: −2 dB contour of satellite uplink antenna		
Downlink earth station		
Receiving antenna aperture efficiency	η_A	65%
Downlink frequency	f_{down}	11.45 GHz
Receiver IF noise bandwidth	B_n	43.2 MHz
Antenna noise temperature	T_a	30 K
LNA noise temperature	T_{LNA}	75 K
Required overall CNR in clear air	$(CNR)_o$	17 dB
Miscellaneous downlink losses	$L_{misc\,dn}$	0.8 dB
Location: −3 dB contour of satellite transmitting antenna		

Table 4.8b Rain attenuation and propagation factors

Ku-band clear air attenuation (worst case)	
Uplink 14.25 GHz	0.7 dB
Downlink 11.45 GHz	0.5 dB
Rain attenuation (worst case)	
Uplink 0.01% of year	6.0 dB
Downlink 0.01% of year	5.0 dB

Hence the required uplink transmitter power P_t is given by

$$P_t - 124.2\ \text{dB} = -98.2\ \text{dBW}$$

$$P_t = 26.0\ \text{dBW or 400 watts}$$

This is a relatively high transmit power so we could increase the transmitting antenna diameter to increase its gain, allowing a reduction in transmit power. For example, a 7 m antenna has a gain 2.9 dB greater than a 5 m antenna, which would allow a reduction in transmitter power to 204 W.

4.8.1.2 Ku-Band Downlink Design

The first step is to calculate the downlink $(CNR)_{dn}$ that will provide $(CNR)_o = 17$ dB when $(CNR)_{up} = 27$ dB. Rearranging Eq. (4.50)

$$1/(CNR)_{dn} = 1/(CNR)_o - 1/(CNR)_{up} \text{ (not in dB)}$$

Thus $1/(CNR)_{dn} = 1/50 - 1/500 = 0.018$

$$(CNR)_{dn} = 55.5 \Rightarrow 17.4 \text{ dB}$$

We must find the required receiver input power to give $(CNR)_{dn} = 17.4$ dB and then find the receiving antenna gain, G_r. The downlink clear air attenuation of 0.5 dB gives a sky temperature of 6.6 x 5 = 33 K.

The LNA noise temperature is 75 K giving a system noise temperature in clear sky conditions of 108 K or 20.3 dBK. With a noise bandwidth of 43.2 MHz or 76.4 dBHz, the receiver noise power is −131.9 dBW.

The power level at the earth station receiver input must be 17.4 dB greater than the noise power in clear air. Hence

$$P_r = -131.9 \text{ dBW} + 17.4 \text{ dB} = -114.5 \text{ dBW}$$

We need to calculate the path loss at 11.45 GHz. At 14.15 GHz path loss was 207.2 dB. At 11.45 GHz path loss is

$$L_p = 207.2 - 20 \log_{10} (14.15/11.45) = 205.4 \text{ dB}$$

The transponder is operated with 1 dB output back off, so the output power is 1 dB below 80 W (80 W ⇒ 19.0 dBW)

$$P_t = 19 \text{ dBW} - 1 \text{ dB} = 18 \text{ dBW}$$

Table 4.8c Uplink carrier power and noise power link budget

Earth station transmitter power	P_t	P_t dBW
Earth station antenna gain	G_t	55.7 dB
Satellite receiving antenna gain	G_r	31.0 dB
Free space path loss	L_p	−207.2 dB
Earth station on −2 dB contour	L_{ant}	−2.0 dB
Atmospheric path loss	L_{up}	−0.7 dB
Miscellaneous losses	L_{misc}	−1.0 dB
Received power at transponder	P_r	P_t − 124.2 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz
Transponder noise temperature 500 K	T_{xp}	27.0 dBK
Receiver noise bandwidth 43.2 MHz	B_n	76.4 dBHz
Transponder noise power	N_{xp}	−125.2 dBW

Table 4.8d Downlink CNR link budget

Satellite transponder output power	P_t	18.0 dBW
Satellite antenna gain	G_t	31.0 dB
Earth station antenna gain	G_r	G_r dB
Free space path loss	L_p	−205.4 dB
Earth station on −3 dB contour of satellite antenna	L_{ra}	−3.0 dB
Clear sky atmospheric loss	L_{dn}	−0.5 dB
Miscellaneous losses	L_{misc}	−0.8 dB
Received power at earth station	P_r	G_r − 160.7 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz
Receiving system noise temp = 33 + 75 K = 108 K	T_s	20.3 dBK
Receiver noise bandwidth 43.2 MHz	B_n	76.4 dBHz
Receiver noise power	N_r	−131.9 dBW

Table 4.8d shows the link budget for downlink from the satellite to the gateway earth station.

The required power into the earth station receiver to meet the $(CNR)_{dn}$ objective is

$$P_r = N_{xp} + (CNR)_{dn} = -131.9 + 17.4 = -114.5 \text{ dBW}$$

Hence the receiving antenna must have a gain G_r where

$$G_r - 160.7 \text{ dB} = -114.5 \text{ dBW}$$

$$G_r = 46.2 \text{ dB or } 41{,}690 \text{ as a ratio}$$

The earth station antenna diameter, D, is calculated from the formula for antenna gain, G, with a circular aperture and an aperture efficiency of 0.65 (65%)

$$G_r = 0.65 \times (\pi D/\lambda)^2 = 41{,}690$$

At 11.45 GHz, the wavelength is 2.62 cm = 0.0262 m. Evaluating the above equation to find D gives the required receiving antenna diameter as $D = 2.11$ m. A 2.2 m antenna could be used in practice, or a larger antenna could be used to increase the downlink rain attenuation margin.

4.8.1.3 Rain Effects at Ku-Band: Uplink

Under conditions of heavy rain, the Ku-band path to the satellite station suffers an attenuation of 6 dB for 0.01% of the year. We must find the uplink attenuation margin and decide whether UPC would improve system performance at Ku-band.

The uplink CNR was 27 dB in clear air. With 6 dB uplink path attenuation, the CNR in the transponder falls to 21 dB, and assuming a linear transponder characteristic and no UPC, the transponder output power falls to 18 − 6 = 12 dBW. The downlink $(CNR)_{dn}$ falls by 6 dB from 17.4 to 11.4 dB, and the overall $(CNR)_o$ falls by 6 dB to 11 dB. The required minimum overall $(CNR)_o$ is 8.0 dB, so the link margin available on the uplink is 9.4 dB without UPC. This is an adequate uplink rain attenuation margin for many

parts of the United States, and would typically lead to rain outages of less than one hour total time per year.

UPC could be implemented so that the earth station transmitter output power is increased when the uplink attenuation is estimated to have reached 3 dB, as illustrated in Figure 4.13. This would hold the value of overall $(CNR)_o$ in the receiver at 14.4 dB. If the UPC system has a dynamic range of 6 dB, the uplink rain attenuation margin is increased to 15.4 dB and the maximum Ku-band transmitter power is increased to 32.0 dBW (1580 W). Rain attenuation exceeds 15.4 dB at 14 GHz for only a few minutes at a time in very heavy thunderstorms, but there would only be a handful of such occurrences in an average year. UPC definitely improves the ability of the uplink to resist rain attenuation, but at the expense of a considerably more powerful, and expensive, uplink transmitter. The extra expense can be justified in a video distribution system with many receiving stations. There is also an increased risk that the additional power radiated by the uplink station when UPC is active will cause interference at an unacceptable level into other satellite links using the same frequencies. It would be advisable to increase the earth station antenna diameter to increase its gain and narrow its beamwidth, and thus reduce the maximum transmit power required and also reduce interference with adjacent satellites. With a 7 m uplink earth station antenna, the maximum transmitter power is reduced to 29.1 dBW (813 W). More than one uplink earth stations is required for an operational video distribution system, to provide alternate capacity if one earth station is down for maintenance, or fails, or suffers excessive rain attenuation. The earth stations are typically located in different geographical areas to reduce the risk that heavy rain affects two stations at the same time.

4.8.1.4 Downlink Attenuation and Sky Noise Increase

The 11.45 GHz path between the satellite and the receive station suffers rain attenuation exceeding 5 dB for 0.01% of the year. Assuming 100% coupling of sky noise into antenna noise, and 0.5 dB clear air gaseous attenuation, calculate the overall CNR under these conditions. Assume that the uplink station is operating in clear air. We must calculate the available downlink fade margin.

We need to find the sky noise temperature that results from a total excess path attenuation of 5.5 dB (clear air attenuation plus rain attenuation); this is the new antenna temperature in rain, because we assumed 100% coupling between sky noise temperature and antenna temperature. We must evaluate the change in received power and increase in system noise temperature in order to calculate the change in CNR for the downlink.

In clear air, the atmospheric attenuation on the downlink is 0.5 dB. The corresponding sky noise temperature is 33 K.When the rain causes 5 dB attenuation, the total path attenuation from the atmosphere and the rain is 5.5 dB.

The corresponding sky noise temperature is given by

$$T_{\text{sky rain}} = T_o(1 - G)\ \text{K}$$

where

$$G = 10^{-\frac{A}{10}} = 0.282$$
$$T_{\text{sky rain}} = 270\ (1 - 0.282) = 194\ \text{K}$$

Thus the antenna temperature has increased from 33 K in clear air to 194 K in rain. The system noise temperature in rain, $T_{s\ rain}$, is increased from the clear air value of 108 K (33 K sky noise temperature plus 75 K LNA temperature) to

$$T_{s\ rain} = 194 + 75 = 269\ \text{K}$$

The increase in noise power is

$$\Delta N = 10\ \log_{10}(269/108) = 4.0\ \text{dB}$$

The signal is attenuated by 5 dB in the rain, so the total reduction in downlink CNR is 9.0 dB, which yields a new value

$$(\text{CNR})_{\text{dn rain}} = 17.4 - 9.0 = 8.4\ \text{dB}$$

The overall CNR is then found by combining the clear air uplink $(\text{CNR})_{\text{up}}$ of 27 dB with the rain faded downlink $(\text{CNR})_{\text{dn rain}}$ of 8.4 dB, giving

$$(\text{CNR})_{\text{o rain}} = 8.3\ \text{dB}$$

The overall $(\text{CNR})_{\text{o}}$ is just above the minimum acceptable value of 8.0 dB. The downlink link margin is

$$\text{Downlink Fade Margin} = (\text{CNR})_{\text{dn}} - (\text{CNR})_{\text{min}} = 17.4 - 8.4 = 8.0\ \text{dB}$$

We have met all the requirements of the system specification, and can be confident that our video distribution system will provide the required availability.

4.8.1.5 Summary of Ku-Band Link Performance

The Ku-band link with a 2.1 m earth station antenna will suffer rain outages because attenuation exceeding 8.3 dB could occur occasionally on the downlinks, affecting individual customers. Outages will rarely occur on the uplink. With UPC and a more powerful transmitter, uplink outages can be restricted to a few minutes per year. A 2.2 or 3.0 m receiving antenna is needed to ensure that the Ku-band downlink will be out for no more than 0.01% of an average year with the given attenuation statistics. The threshold value for overall CNR was set at 8.0 dB because we can use QPSK and half rate error correction coding to obtain an equivalent (CNR) ratio of about 14.0 dB without coding. Allowing a 0.5 dB *implementation margin* (see Chapter 5), the BER on the downlink will remain below 10^{-7} except when very heavy rain affects the downlink. In clear sky conditions there will be no errors on the link. The 43.2 Msps QPSK signal with half rate FEC can deliver a data rate of 43.2 Mbps, which can support up to 10 live MPEG 2 video channels and 20 pre-recorded video channels.

The video distribution system described here is designed to deliver multiple video channels to cable TV stations with low risk of outages. DBS-TV delivers video signals directly to the customer's location using a much smaller 0.5 m receiving antenna. The smaller antenna can be used because the DBS-TV satellites transmit at a higher power level (typically 160 W), the symbol rate is lower (20–27 Mbps) and availability of the signal at the receiving antenna is guaranteed for only 99.7% of the year.

The gateway station can transmit multiple carriers on different frequencies to separate transponders on the same satellite. For example, by transmitting to five transponders, 50 full motion MPEG-2 compressed digital video channels and 100 pre-recorded video

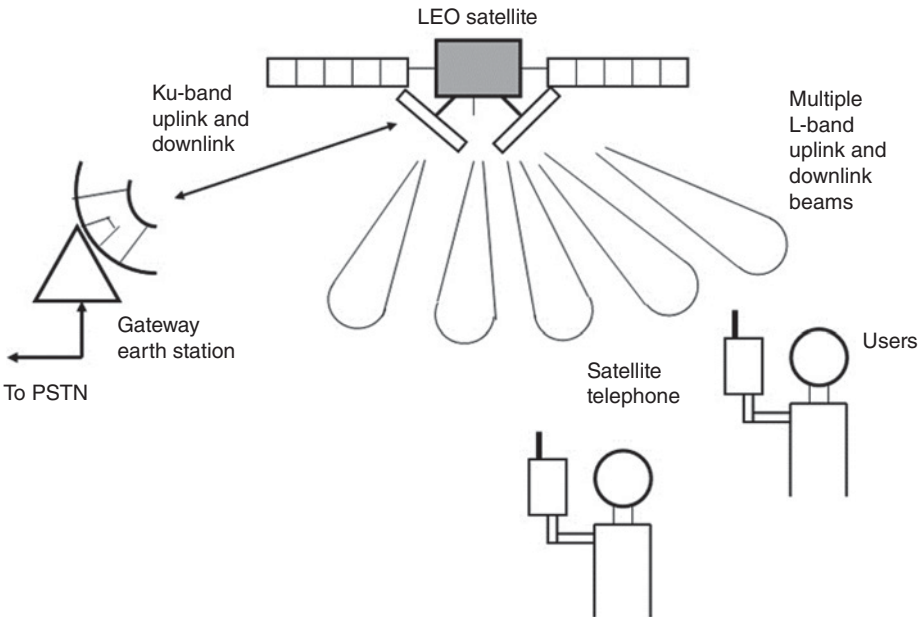


Figure 4.15 Illustration of a satellite telephone system using a low earth orbit satellite with multiple uplink and downlink beams. PSTN: Public switched telephone network.

channels can be sent to every cable TV head end for distribution over hundreds or thousands of local cable TV systems. This requires five separate high power transmitters and an RF combining network at each uplink earth station.

Example 4.12 Personal Communication System Using Low Earth Orbit Satellites

LEO satellite systems can be designed to provide *personal communication service* similar to a cellular telephone, but over a much wider area. LEO satellite systems can cover sparsely populated regions of a country, or the world, where there are no terrestrial cellular telephone systems. The shorter path length to and from a LEO satellite incurs a much shorter delay than with a GEO satellite. The user has a *handset* similar to a cellular telephone handset that provides two-way voice communications through a gateway station, usually to a conventional telephone in a home or office connected to the public switched telephone network (PSTN). Satellite telephones can equally well connect to another satellite handset, or to a terrestrial cellular telephone. Figure 4.15 illustrates a satellite telephone system.

LEO satellite communication systems use a large number of satellites in multiple orbital planes at altitudes between 350 and 1400 km. The satellites can communicate with a limited portion of the earth's surface because of the low orbital altitude, and appear to an observer on the earth to fly across the sky in a few minutes. Communication is maintained between the user and a gateway station by switching the channels from one satellite to the next as one satellite goes below the horizon and another comes into view. The

satellites are all identical, so the link design is based on a single link between one user and the gateway station via one satellite. Examples of such LEO satellite systems are Iridium and Globalstar (see Chapter 9). In 2018 there were many proposals for LEO satellite systems employing hundreds or thousands of satellites to provide worldwide internet access.

Early LEO satellite systems operated in L-band, in the 1500 and 1600 MHz bands, and in the lower part of S-band around 2460 MHz, frequency bands that are allocated for mobile satellite communications. Proposed internet access systems use Ka-band and higher frequencies to access wider bandwidths and increase capacity. Some LEO systems use ISLs so a user can connect to any point in the world without an intermediate return to earth. However, the signals invariably pass through a gateway station at each end of the link to facilitate control of the call and to ensure that users can be charged for using the system. Connections between the gateway earth stations and the satellites use S-band, C-band, Ku-band, or Ka-band frequencies, depending on the system requirements. Only a small portion of the radio spectrum at L-band is allocated to LEO and MEO satellite systems, so L-band frequencies are reserved for the critical links between the user and the satellite. Rain attenuation at L-band is very small and can be ignored, but with a mobile terminal blockage by buildings and trees is significant and can cause outages (Allnutt 2011, pp. 471–486). See Chapter 7 for a discussion of mobile satellite service propagation effects.

A hand-off process is required for LEO satellites similar to that used in cellular telephone networks, and the hand-off between satellites should not be apparent to the user. Most LEO satellites have multiple beam antennas, and the beam pattern moves across the earth's surface at the speed of the satellite – typically about 7.7 km/s or 17 200 mph. A single beam is typically 250–500 km in diameter for L-band satellites, much smaller when Ka- or V-band is used, so an individual user is in any one beam for less than a minute. The system provides automatic switching from beam to beam within the same satellite antenna footprint, much like a cellular telephone system switches users from cell to cell, which may require a change in the link frequencies.

The example below analyzes the links between a user and a gateway station. LEO satellite systems employ digital transmission so that advantage can be taken of FEC coding and speech compression techniques. The bit rate of digital voice in a LEO satellite link is typically 4800 bps, requiring powerful compression algorithms. The low bit rate allows more signals to be sent in the available transponder bandwidth and also helps maintain the CNR in the receivers. When FEC in the form of turbo or Low density parity code (LPDC) coding is applied to a digital bit stream, CNRs down to 2 dB can be used. The low bit rate, and operation of the receivers at low CNRs are essential to make personal communication via a LEO satellite possible. Higher bit rates are needed for internet access.

The link between the gateway station and the mobile terminal is defined as the *Outbound Link*, and the link from the mobile terminal to the gateway is the *Inbound Link*. Note that there are four satellite paths, just as in all other two-way satellite communication systems: outbound uplink, outbound downlink, inbound uplink, inbound downlink. Each has its own unique frequency, and in most LEO satellite systems, one of the links will be weaker than the other three links and will thus limit the system performance. One objective in the example that follows is to identify the weakest path and to

then attempt to improve that part of the system. Figure 4.15 illustrates the two-way link between the gateway station and the handset. Note that separate transponders are used for the inbound and outbound paths.

In this example, the mobile terminals transmit to a transponder on the satellite using FDMA and single channel per carrier (SCPC) techniques. FDMA and SCPC are discussed in Chapter 6. However, the principle is simple: each transmitter is allocated its own frequency, just like broadcast stations. The available frequencies are shared among active users on demand, so a call begins with a start-up sequence that establishes communication between the mobile terminal and the local gateway station via the nearest LEO satellite. The gateway station then allocates frequencies for the call. At the end of the call, the frequencies are released and become available for another user. This is called demand assignment (DA), and the multiple access technique is identified by the acronym SCPC-FDMA-DA, or, alternatively, SCPC-FDMA-DAMA where DAMA stands for *demand assignment multiple access*. A common set of control channels at pre-assigned frequencies enables call setup and tear down. Other multiple access techniques can be used, including time division multiple access (TDMA) and time division duplexing (TDD) – see Chapter 6 for details.

The link from the gateway station via the satellite to the mobile terminal uses time division multiplexing (TDM). The TDM signal consists of a sequence of packets with addresses that repeats every 20 ms. The addresses identify which terminal should receive each packet. The TDM bit stream rate must exceed the total bit rate of all active terminals in a two-way telephone system so that there is sufficient capacity available for each terminal within the TDM bit stream. All links require that signals pass through a gateway station at both the transmitting and receiving ends of the link, resulting in four signal paths through space. The longest time delay for a round trip from one mobile terminal to another is approximately 50 ms, considerably shorter than the 500 ms delay with a GEO satellite. However, packets are delivered to user terminals at 20 ms intervals, adding further delay to the link.

In this example we begin by assuming that 100 active users share the common TDM channel on the outbound link from the gateway station to the mobile terminal. We will also assume that the gateway earth station operates at Ku-band to and from the satellite, and that the satellite employs a linear transponder (bent pipe) rather than having onboard processing. The parameters of the satellite transponder, the mobile terminal, and the gateway station are given in Tables 4.9a–4.9d. The tables give the maximum path length for any satellite-earth link.

The transmit and receive antennas on the satellite in this example create multiple beams that have individual gain of 22 dB and a 3 dB beamwidth of 14 degrees. The LEO satellite needs approximately 50 beams to cover the visible portion of the earth. Minimum elevation angle is set at 10° to avoid excessive blocking by trees and buildings. A separate transponder with an output power of 10 W is connected to each of the multiple downlink beams, giving a total RF power at L-band of 500 W. The uplink antennas on the satellite have low gain because a broad beam is needed to maintain communication with a gateway station as the satellite moves across the sky.

The user's transmitter and receiver is called a *mobile terminal* in this example. It could be a handheld device like a cellular telephone, sometimes called a *satellite telephone* or *satphone,* or the terminal could be mounted on a vehicle. The capacity of any mobile satellite communication system is limited by the low gain of the antenna on the mobile terminal, which is typically omnidirectional on a handheld device. The performance of

Table 4.9a LEO satellite personal communication system parameters

Satellite parameters		
Saturated output power per transponder	$P_{\text{t sat}}$	10 W
Transponder bandwidth	B_{transp}	1 MHz
Uplink frequency from mobile terminal		1650 MHz
Downlink frequency to mobile terminal		1550 MHz
Satellite uplink antenna gain 1650 MHz (one beam)	$G_{\text{r sat L}}$	22 dB
Satellite downlink antenna gain 1550 MHz (one beam)	$G_{\text{t sat L}}$	22 dB
Number of transmit and received beams		50
Satellite uplink antenna gain 14 GHz	$G_{\text{t es Ku}}$	6 dB
Satellite station downlink antenna gain 11.5 GHz	$G_{\text{r es Ku}}$	6 dB
Satellite receiver system noise temperature	$T_{\text{s sat}}$	300 K
Satellite altitude		865 km
Maximum range to edge of coverage zone	R_{max}	2500 km
Gateway station parameters		
Transmitter output power (maximum per transponder)	$P_{\text{t es}}$	10 W
Antenna gain (transmit, 14.0 GHz)	$G_{\text{t es}}$	55 dB
Antenna gain (receive, 11.5 GHz)	$G_{\text{r es}}$	53.5 dB
Receive system noise temperature (clear sky)	T_{s}	110 K
Transmit bit rate (before FEC encoder)	R_{b}	500 kbps
Mobile terminal parameters		
Transmitter output power	$P_{\text{t mob}}$	0.5 W
Minimum antenna gain (transmit and receive)	$G_{\text{t mob}}$, $G_{\text{r mob}}$	0 dB
Receiver system noise temperature	$T_{\text{s mob}}$	300 K
Transmit bit rate	$R_{\text{b tx}}$	4800 bps
Receive bit rate	$R_{\text{b rx}}$	500 kbps
Required maximum bit error rate	BER_{max}	10^{-4}

a handheld terminal improves as the RF frequency is lowered because of the reduced path loss – hence the use of L-band for voice connections and the very high frequency (VHF) and ultra high frequency (UHF) bands for LEO data transfer satellite systems.

Consider the case of a handheld terminal with an omnidirectional antenna receiving a signal from a LEO satellite at a frequency f Hz and wavelength λ m transmitting an EIRP $P_{\text{t}}\, G_{\text{t}}$ watts. The satellite has a defined footprint that spreads its transmitted energy over an area of A m^2 on the earth's surface. The average flux density is F watts/m^2 where

$$F = P_{\text{t}} G_{\text{t}} / (4\pi R^2) \text{W/m}^2 \tag{4.51}$$

An antenna with a gain of 0 dB (ratio 1) has a gain G and an effective area A_{e} where

$$G = 4 \times \pi \times A_{\text{e}} / \lambda^2 = 1 \tag{4.52}$$

Hence the effective area of the omnidirectional receiving antenna is

$$A_e = \lambda^2/4\pi m^2$$

and the power received is P_r where

$$P_r = F \times A_e = \left[P_t G_t/(4\pi R)^2\right] \times \left(\lambda^2/4\pi\right) = \frac{P_t G_t \lambda^2}{(4\pi R)^2} \text{ watts} \tag{4.53}$$

Thus received power increases as the square of the wavelength, and therefore decreases as the square of the frequency, which favors lower RF frequencies for any link of this type where omnidirectional antennas are used. As an example, a link operating at 150 MHz will create a received power that is 20 dB higher than the same link operated at 1500 MHz when each has the same satellite beam footprint.

The satellite has multiple L-band beams serving different parts of its instantaneous coverage zone because a single beam from a LEO satellite has both low gain and limited capacity. For an antenna with a gain of 22 dB, $G = 160$ and the beamwidth is approximately θ_{3dB} where

$$\theta_{3dB} = (30{,}000/160)^{1/2} = 13.7 \text{ degrees}$$

The use of a multiple beam antenna on the satellite increases the antenna gain toward the mobile terminal, which increases the CNR of the signals in the mobile terminal gateway station receivers. The LEO satellite is at an altitude of 865 km and 50 beams are needed to cover the footprint of the LEO satellite. The uplink CNR on the link between the gateway station and the satellite is high through the use of a large antenna and a high transmitter power at the gateway station, allowing the use of small Ku-band antennas on the satellite.

The antenna gain at the mobile terminal is low, with a value of 0 dB used for calculation, because the antenna coverage of the terminal must be very broad. If the terminal is a satellite telephone, an omnidirectional antenna allows the user to move around freely. If the mobile terminal antenna gain were to be increased, its beam would be correspondingly narrower, and the user would have to point the handset at the satellite. In a LEO satellite system, the user does not know which satellite is being used nor where it is in the sky, so requiring the user to point the handset antenna at the satellite is not a feasible option. When the mobile terminal is mounted in a vehicle with the antenna on the roof, pointing the antenna at the satellite is not possible unless an automatic tracking phased array antenna is used.

In this example, we will begin by assuming that there are 100 users sharing a single transponder on the satellite, and that one transponder serves one of the L-band beams within the LEO satellite coverage, operating within a given set of frequencies. A large number of users can share a LEO satellite through the provision of many transponders, each of which is connected to one of the individual beams in the multiple L-band antenna coverage of the satellite. The signal received by a mobile terminal from the gateway is a TDM sequence of packets carrying 100 digital voice channels, each at 4800 bps. The bit rate of the TDM signal would be 480 kbps if it carried only the voice signals, but will be higher in practice because additional overhead bits must be sent with each packet; a TDM bit rate of 500 kbps is used in this example. Individual mobile terminals pull off their assigned packets from within the TDM stream and ignore the

rest. All signals are transmitted with QPSK modulation and half rate FEC. The symbol rates on the links are therefore equal to the data bit rates.

All digital links are designed with ideal SRRC filters, which have noise bandwidth, B_n Hz, numerically equal to the symbol rate of the digital signal in symbols per second. The maximum permitted BER of the digital signal of 10^{-4} leads to a theoretical SNR in the speech channel of 34 dB. (SNR = $1/4\,P_e$, where P_e is the BER – see Chapter 5.) However, with compressed speech a *mean opinion score* (MOS) should be used instead of SNR.

4.8.2 Inbound Link: Mobile Terminal to Gateway Station

Each terminal transmits a QPSK signal at 4800 sps at an allocated frequency. The satellite transponder shifts all received L-band signals in frequency before retransmission at Ku-band to the gateway station, and also amplifies the signals with a linear transponder. At the gateway station, the antenna and RF receiver are connected to many identical IF receivers tuned to the individual frequencies (translated by the transponders) of the handheld transmitters. Each IF receiver has a noise bandwidth of 4800 Hz, set by a SRRC filter with $\alpha = 0.25$, giving an occupied channel bandwidth of 6.0 kHz (see Chapter 5 for details of the design of digital links).

At the receiving end of the link, the CNR at the input to the QPSK demodulator must be high enough to provide an acceptable BER. Here, we require a maximum BER of 10^{-4}, which provides a theoretical minimum SNR of 34 dB in the speech channel. We will assume that the required CNR to achieve a BER of 10^{-4} with QPSK modulation is 3.0 dB when half rate FEC and turbo coding is employed, which gives BER performance similar to the DVB-S2 curve for QPSK and half rate FEC coding in Figure 4.11. We need a minimum overall CNR of 3.0 dB to meet the BER and SNR specifications. We can now design the satellite link to achieve this minimum CNR.

4.8.3 Mobile Terminal to Satellite Link

We will establish power and noise link budgets for each of the four paths, beginning with the uplink from the mobile terminal to the satellite.

The received power at the output of the uplink antenna on the satellite from Eq. (4.11) is

$$P_r = \text{EIRP} + G_r - L_p - L_{misc}\ \text{dBW}$$

where EIRP is the product of transmitter output power and transmitting antenna gain, $P_t\,G_t$ in dBW, G_r is the satellite receive antenna gain, L_p is the path loss of the link, and L_{misc} accounts for all other losses. The noise power, N_{xp}, at the input to the satellite receiving system from Eq. (4.13) is

$$N_{xp} = kT_sB_n\ \text{watts}$$

or using decibel units

$$N_{xp} = k + T_s + B_n\ \text{dBW}$$

Path loss L_p is found from Eq. (4.12)

$$L_p = 20\ \log_{10}\,(4\,\pi R/\lambda)\ \text{dB}$$

where R is the distance in meters between the transmitting and receiving antennas in the link and λ is the wavelength in meters.

The uplink frequency is 1650 MHz, giving $\lambda = 0.1818$ m. The maximum range is 2500 km so maximum path loss is

$$L_p = 20 \log_{10} (4\pi \times 2.5 \times 10^6 / 0.1818) = 164.8 \text{ dB}$$

We will assume that there are miscellaneous losses in the uplink of 0.5 dB, caused by polarization misalignments, gaseous absorption in the atmosphere, and so on. The calculation of the CNR is made for the worst case of an earth station located on the −3 dB contour of a satellite antenna beam, so a 3 dB reduction in satellite antenna gain is applied, making the value of $L_{misc} = -3.5$ dB. We can now set out the link power and noise budgets for clear line of sight conditions, when there is no attenuation caused by obstructions in the path. Table 4.9b is the link budget for the uplink from the mobile terminal to the satellite, and Table 4.9c is the link budget for the downlink from the satellite to the gateway earth station.

Table 4.9b Uplink inbound CNR budget. Mobile terminal to satellite

EIRP of handheld unit	$P_t G_t$	−3 dBW
Gain of satellite receiving antenna	G_r	22 dB
Path loss at 1650 MHz	L_p	−164.8 dB
Miscellaneous losses	L_{misc}	−3.5 dB
Received power at satellite	P_r	−149.3 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz
System noise temperature 300 K	T_s	24.8 dBK
Noise bandwidth 4800 Hz	B_n	36.8 dBHz
Noise power	N_{xp}	−167.0 dBW
$(CNR)_{up} = P_r$ dBW − N dBW		17.7 dB

The inbound uplink CNR in the transponder is 17.7 dB. This is the lowest CNR that should occur in the transponder in clear sky conditions, since the calculation was made for a mobile terminal at the longest range from the satellite and at the edge of a satellite antenna beam. The mobile terminal antenna gain has also been set to its minimum value of 0 dB. If the satellite were directly overhead the range would be 865 km instead of 2500 km, making the path loss lower by 9.2 dB, and the miscellaneous losses would be 3 dB lower at the center of the satellite antenna beam, making the power received at the transponder 12.2 dB greater, and then $(CNR)_{up} = 29.9$ dB. However, we cannot use this figure for the system design, otherwise there would be only one user who could make calls, and then only for a brief moment as the satellite passes directly overhead. We must ensure that all users within the satellite's coverage zone have adequate CNRs in their links for successful communication and must therefore start by calculating CNR values for the worst case. An ideal antenna for communication with LEO satellites should have maximum gain at 10° elevation angle falling to a value 9 dB lower at zenith.

4.8.4 Satellite to Gateway Station Down Link

The next step in calculating the CNR for the inbound link is to calculate $(CNR)_{dn}$ in the gateway receiver. We are operating the transponder in FDMA, so the individual mobile terminal signals must share the output power of the transponder. We will assume

that 100 active terminal signals share the 1 MHz transponder bandwidth and that 3 dB backoff is used at the transponder output to obtain quasi-linear operation of the transponder HPA (remembering that we have assumed linear transponder operation in this example). The transponder output power is therefore 10 dBW − 3 dB = 7 dBW (5 W). The 5 W transponder output power must be shared equally between the 100 signals in the transponder, giving 0.05 W = −13 dBW per signal at the transponder output for the downlink to the gateway station. At the edge of the satellite downlink beam the satellite antenna gain is 3 dB. We will use the same worst case conditions as for the uplink – maximum path length and minimum satellite antenna gain, with miscellaneous losses of 0.5 dB, giving 3.5 dB loss on the downlink.

We can now establish a link budget for a single channel downlink from the satellite to the gateway station. The link budget is shown in Table 4.9c.

Table 4.9c Inbound downlink CNR budget. Satellite to gateway station

EIRP per channel	$P_t\ G_t$	−13.0 dBW
Gain of receiving antenna	G_r	53.5 dB
Path loss at 11.5 GHz	L_p	−181.6 dB
Downlink losses	L_{misc}	−3.5 dB
Received power at gateway station	P_r	−144.6 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz
System noise temperature 110 K	T_s	20.4 dBK
Noise bandwidth 4800 Hz	B_n	36.8 dBHz
Noise power	N_{es}	−171.4 dBW
$(CNR)_{dn} = P_r$ dBW − N dBW		26.8 dB

The CNR in the 4.8 kHz noise bandwidth of a gateway station IF receiver is 26.8 dB. $(CNR)_{dn}$ for the inbound downlink is higher than $(CNR)_{up}$ for the inbound uplink because of the high gain of the gateway station antenna. The gain of the antenna, 53.5 dB, corresponds to an antenna diameter of 5 m and an aperture efficiency of 60%, so its beamwidth is narrow, about 0.4°, and the gateway station must track the satellite as it crosses the sky. Several gateway station antennas are needed to maintain continuous communication. As one satellite goes below the minimum operating elevation angle of 10° a different gateway antenna must be waiting to connect to another satellite that has appeared above the horizon in a different part of the sky.

The overall $(CNR)_o$ at the gateway is calculated by combining the uplink CNR and downlink CNR values from Tables 4.8b and 4.8c using Eq. (4.43), since both the transponder and the gateway station receiver add noise to the signal. The values used in the formula are ratios, that is, CNR values are not in decibels.

$$1/(CNR)_o = 1/(CNR)_{up} + 1/(CNR)_{dn}$$

For the inbound uplink, $(CNR)_{up}$ = 17.7 dB ⇒ 58.9 as a ratio. For the inbound downlink, $(CNR)_{dn}$ = 26.8 dB ⇒ 478 as a ratio. Hence the overall CNR is in the gateway station receiver is

$$(CNR)_o = 1/(1/58.9 + 1/478) = 52.44 \text{ or } 17.2 \text{ dB}$$

The overall CNR of 17.2 dB at the gateway station receiver guarantees that with the selected modulation and FEC there will be no bit errors under clear air conditions and the SNR of the speech channel will be set by the performance of the speech compression system. The maximum permitted BER is 10^{-4}, which occurs with $(\text{CNR})_o = 3.0$ dB. We therefore have an inbound link margin of $(\text{CNR})_o = (17.1 - 3.0) = 14.1$ dB. However, we must calculate the individual link margins for the uplink and downlink in order to be able to use the margins for fading analysis. This will be done at the end of the example.

4.8.5 Outbound Link

The outbound link from the gateway station to the mobile terminal sends a continuous 500 ksps TDM bit stream using QPSK modulation and half rate FEC coding and a separate transponder with 1 MHz bandwidth. The bit stream is a series of packets addressed to all 100 active terminals. The noise bandwidth of the terminal receiver is 500 kHz, assuming ideal root raised cosine (SRRC) filters. The outbound uplink and downlink CNR values are calculated in exactly the same way as for the inbound link, and the power and noise budgets are combined to give CNRs directly from a single table. The link budget for the outbound uplink and downlink are shown in Tables 4.9d and 4.9e.

Table 4.9d Outbound uplink CNR budget. Gateway station to satellite

Gateway station EIRP	$P_t\ G_t$	65.0 dBW
Gain of receiving antenna	G_r	6.0 dB
Path loss at 14.0 GHz	L_p	−183.3 dB
Miscellaneous and edge of beam losses	L_{misc}	−4.0 dB
Received power at satellite	P_r	−116.3 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz
System noise temperature 500 K	T_s	27.0 dBK
Noise bandwidth 500 kHz	B_n	57.0 dBHz
Noise power	N_{xp}	−144.6 dBW
$(\text{CNR})_{up} = P_r$ dBW $- N$ dBW		28.3 dB

Table 4.9e Outbound downlink CNR budget. Satellite to mobile terminal

EIRP of satellite	$P_t\ G_t$	31.0 dBW
Gain of receiving antenna	G_r	0 dB
Path loss at 1550 MHz	L_p	−164.2 dB
Miscellaneous and edge of beam losses	L_{misc}	−3.5 dB
Received power at mobile	P_r	−136.7 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz
System noise temperature 300 K	T_s	24.8 dBK
Noise bandwidth 500 kHz	B_n	57.0 dBHz
Noise power	N_{es}	−146.8 dBW
Downlink CNR	$(\text{CNR})_{dn}$	10.1 dB

At the uplink frequency of 14 GHz, clear air atmospheric attenuation of 1.0 dB is included in the miscellaneous losses, together with the usual 3 dB loss for the user at the edge of the satellite antenna beam. The Ku-band receive system noise temperature at the satellite is 500 K.

The satellite transponder carrying the single 500 ksps TDM outbound signal can be operated close to saturation because there is only one signal in the transponder, thus eliminating intermodulation problems. We will allow 1.0 dB backoff at the transponder output to avoid saturating the transponder, giving a transmitted power $P_t = 9.0$ dBW. Downlink satellite antenna gain is 22 dB on axis, giving an EIRP of 31.0 dB. Miscellaneous losses on the downlink are 0.5 dB atmospheric loss and 3 dB for the edge of the antenna beam.

Combining the CNR values for the uplink and downlink gives the overall $(CNR)_o$ ratio at the mobile terminal receiver. Converting the CNR values from decibels gives

$$(CNR)_{up} = 28.3 \text{ dB} = 676 \text{ as a ratio}$$

$$(CNR)_{dn} = 10.1 \text{ dB} = 10.23 \text{ as a ratio}$$

Hence, the overall $(CNR)_o$ for the outbound link is

$$\begin{aligned}(CNR)_o &= 1/[1/(CNR)_{up} + 1/(CNR)_{dn}] = 1/[0.00148 + 0.0978] \\ &= 10.07 \text{ or } 10.0 \text{ dB}\end{aligned}$$

Note that the downlink CNR is so much lower than the uplink CNR that the overall CNR is almost equal to the downlink CNR.

The clear sky $(CNR)_o$ value is 7.0 dB above the minimum allowed for BER = 10^{-4} on the outbound link, leaving a 7 dB margin for blockage by buildings, the user's head, multipath effects, the ionosphere, or vegetative shadowing on the downlink. The link margins for the outbound link are lower than for the inbound link, and it is therefore the weakest part of the system. Attenuation exceeding 7 dB in the downlink from the satellite to the mobile terminal will cause the BER to exceed 10^{-4} and the link will fail. Because of the very steep characteristics of the BER vs CNR curve for turbo coding, the speech channel will be unusable if downlink attenuation exceeds 7 dB.

The link margins are quite small for a mobile system in which the line of sight between the satellite and the user can easily be blocked by building, trees, or by the user's body. It is the link between the satellite and the mobile terminal that sets the overall CNR value for both the inbound and the outbound links, but there is little room to change the system parameters to yield higher margins.

When the mobile terminal is a satellite telephone handset, the transmitter power is limited by FCC regulations to ensure that there is no short term biological hazard to the user when the handset is transmitting. (See Chapters 7 and 9 for further details on radiation limits for portable equipment.) The power from the satellite is limited by the transponder HPA output power and the low gain of the handset antenna. However, a higher gain antenna would have a narrower beam and would have to track the satellite automatically – a smart antenna could be built to do this, but the small size of most mobile telephone handsets limits the available improvement to no more than 3–4 dB.

4.8.6 System Performance

The preceding calculations show that the LEO satellite system can support two-way digital speech with 100 active users per transponder, and provides link margins of approximately 16 dB in the inbound link and 7 dB in the outbound link. However, these values were calculated for the worst case of a mobile terminal at the edge of the satellite footprint. Half of the footprint area of the satellite has a path loss at least 3 dB lower than the worst case, so on average, link margins are at least 3 dB better than the worst case for half of the mobile terminals.

With 50 spot beams in the satellite footprint, there can be a maximum of 50 000 simultaneous mobile terminal connections to the terrestrial telephone system, or 25 000 mobile to mobile connections. The satellite footprint is 4400 km wide, which covers the entire United States, and system capacity compares poorly to the number of simultaneous users of cellular phones. Consequently, charges per minute for a satellite phone connection will have to be much higher than for cellular telephones, restricting the use of satellite phones to terminals that cannot reach the cellular network. The low bit rate in this system (4800 bps) makes data connections very slow compared to a cellular phone, and suitable only for short text messages and email.

The RF bandwidth used by the inbound and outbound links is found from the symbol rates and the α values of the SRRC filters (see Chapter 5). For the outbound link, the symbol rate is 500 ksps, giving

$$B_{\text{outbound}} = 500 \times (1 + \alpha) = 625 \text{ kHz}$$

For the inbound link, the symbol rate for one speech channel is 4800 baud (4.8 ksps), giving

$$\text{B}_{\text{inbound}} = 4.8 \times (1 + \alpha) = 6.0 \text{ kHz}$$

The inbound channels access the satellite transponder using SCPC-FDMA, so the RF signals are distributed across the transponder bandwidth. We must space the channels more than 6 kHz apart in the transponder so that the narrow bandpass filters in the gateway station receiver can extract each speech channel without interference from the adjacent channels. If we use a 10 kHz channel spacing, there will be a frequency gap, called a *guard band* of 4 kHz between each channel, which will ensure minimal interference from adjacent channels. With 100 channels sharing one transponder, the total bandwidth occupied in the inbound link transponder will be 1 MHz.

4.8.7 Rain Attenuation at Ku-Band

Rain causes attenuation at Ku-band, as discussed in Chapter 9. We must calculate the rain attenuation margins for the inbound downlink and the outbound uplink and determine the probability of an outage. The link margin is the number of decibels by which the CNR on an uplink or a downlink can be reduced before the overall $(\text{CNR})_{\text{o}}$ for that link falls to the threshold value. We will use 3.5 dB as the threshold value for overall CNR in each case, assuming that half rate FEC with turbo coding is used. We will also assume that clear sky conditions prevail on the uplink when extreme attenuation occurs on the downlink, and vice versa.

For the inbound Ku-band downlink, using half rate FEC, the clear air CNR is 26.8 dB (ratio 785) and the L-band clear air uplink CNR is 14.7 dB (ratio 29.5). With a threshold at 3.5 dB (ratio 2.24), the minimum downlink CNR will be given by (using ratios, not dB)

$$\begin{aligned} 1/(\mathrm{CNR})_{\mathrm{dn\ min}} &= 1/(\mathrm{CNR})_{\mathrm{o}} - 1/(\mathrm{CNR})_{\mathrm{up}} \\ &= 1/2.24 - 1/478 = 0.444 \end{aligned}$$

corresponding to a $(\mathrm{CNR})_{\mathrm{dn\ min}}$ of 3.5 dB. The downlink margin is 26.8 − 3.5 = 23.3 dB. Rain attenuation at 11.5 GHz very rarely exceeds this value in the United States, so for a US system, the Ku-band downlink margin is adequate.

The Ku-band uplink has a clear sky $(\mathrm{CNR})_{\mathrm{up}}$ ratio of 24.9 dB, but attenuation of the uplink signal causes a reduction in received power at the transponder input. If the satellite transponders are linear (bent pipe), the output power will fall when the input power is reduced by uplink rain attenuation. Because the transponder is operated close to saturation, there will not be a one to one correspondence in the changes in power level at the input and output, but exact analysis is beyond the scope of this example; a linear relationship will be assumed here. The transponder non-linearity actually increases the uplink rain attenuation margin, because the output signal from the satellite will fall less than the input signal to the satellite, so the results that follow represent a pessimistic estimate of the margin available. A regenerative repeater always transmits at constant output power and is very desirable in a digital system. It avoids the difficulty of attenuation on the uplink causing a reduction in transponder output power.

Applying the same analysis as used for the Ku-band downlink, with $(\mathrm{CNR})_{\mathrm{dn}} = 10.1$ dB (ratio 10.2) in clear sky conditions and $(\mathrm{CNR})_{\mathrm{o\ min}} = 3.5$ dB (ratio 2.24)

$$1/(\mathrm{CNR})_{\mathrm{dn\ min}} = 1/(\mathrm{CNR})_{\mathrm{o}} - 1/(\mathrm{CNR})_{\mathrm{dn}} = 1/2.24 - 1/10.2 = 0.348$$

Thus the minimum $(\mathrm{CNR})_{\mathrm{up}}$ ratio is $10 \log_{10} (1/0.348) = 4.6$ dB, ignoring the effects of non-linear coupling between input and output power in the transponder. When the latter effect is considered with a linear transponder characteristic, the limit is set by the $(\mathrm{CNR})_{\mathrm{o}}$ ratio falling to 3.5 dB. This will occur with 10.1 − 4.6 = 5.5 dB uplink attenuation, which is the limiting value. The uplink will not be reliable with a 5.5 dB margin, so UPC will be needed to prevent the input power level of the transponder from falling when rain affects the uplink. It would be straightforward to use UPC in this case. Attenuation on the downlink at 11.5 GHz is measured using the satellite beacon, scaled to 14.0 GHz, and used to set the gateway station transmit power level.

Let's use a UPC system with a dynamic range of 7 dB, and rain attenuation on the uplink is allowed to reach 2 dB at 14.0 GHz before the UPC comes in. The downlink CNR will fall from 10.1 to 8.1 dB and then remain constant until uplink rain attenuation is 9.0 dB. The minimum uplink CNR is 4.6 dB for an overall CNR in the mobile receiver of 3.5 dB, so the maximum rain attenuation on the uplink is 12.5 dB with 7 dB of UPC. This gives the 14.0 GHz uplink a rain attenuation margin of 12.5 dB, which would maintain the link for better than 99.99% of a year throughout most of the United States. Typically, uplink earth stations are located where heavy rain occurs infrequently to further lessen the chance of outages. The open-loop UPC system discussed here would probably have a margin of error of at least ±1 dB in estimating uplink attenuation under low fading conditions. Uncertainties in identifying the propagation mechanism that is causing the fading and the difficulty of accurately setting the clear sky baseline for the signal make greater accuracy unlikely. If two mobile terminals are located within the same satellite beam coverage, and are therefore operating through the same gateway earth station,

the assumption of non-simultaneous outage of the two links would not be valid. Such situations are expected to be rare occurrences.

The gateway station would typically be sited in a dry region, such as Wyoming or Idaho in the United States, to minimize the number and severity of rain attenuation events. Thus rain attenuation at Ku-band can be overcome by a large link margin for the downlink and implementation of UPC in the uplink, and by intelligent siting of the gateway station. All 100 channels can be guaranteed to be unaffected by rain in the outbound uplink.

4.8.8 Path Blockage at L-Band

Trees, buildings, and people are the most likely causes of blockage that affect the performance of the mobile terminal at L-band. Blockage by buildings is too severe to allow the L-band link to operate, and most LEO satellite telephones will not work indoors. Some systems like Iridium incorporate a cellular telephone into the handset. The cellular telephone is used in preference to the satellite phone to reduce loading on the LEO satellite system, and also whenever the satellite signal is unavailable, such as indoors. Paging options have been designed into some mobile satellite systems, which permit users to be alerted that there is an incoming call. The user still has to run outdoors to be able to receive the call, and this has evidently been a factor deterring the use of satellite telephone by business people.

The link margins for the L-band links are calculated in the same way as the Ku-band margins. Repeating the calculations with a minimum overall CNR of 3.5 dB and no rain attenuation in the Ku-band links gives

L-band uplink margin = 13.7 dB

L-band downlink margin = 5.5 dB

The downlink from the satellite to the mobile terminal is therefore the most vulnerable of the links, and cannot be made robust without reducing the number of users per transponder. However, the value of 5.5 dB for the downlink margin is a worst case value and most of the users will have a margin several dB higher. A margin of 5.5 dB can be exceeded by attenuation through a stand of trees. For example, if the user is in a vehicle traveling along a road cut through a forest, and the satellite has a low elevation angle, the 5.5 dB attenuation margin may be exceeded from time to time, causing repeated breakup of the downlink signal. Transmission protocols and signal buffering can be designed to reduce the impact of this type of intermittent loss of signal. Multipath effects when the satellite is at a low elevation angle can also cause variations in signal level leading to lower performance and occasional outages in extreme cases, such as an over-water path.

4.8.9 Summary of L-Band Mobile PCS System Performance

The personal communication system in this example uses a network of LEO satellites to link a user anywhere in the system's coverage zone to a gateway station, and then to the PSTN or another mobile terminal. The user's terminal operates in L-band and is similar to a cellular telephone, with a low gain, omnidirectional antenna. The transmissions are digital, and use speech compression to achieve a bit rate of 4.8 kbps per speech channel. There are a maximum of 100 users in each of the satellite's 50 L-band beams, giving

a nominal satellite capacity of 50 000 mobile terminals. With this maximum number of simultaneous users, the satellite would need to generate 2 kW of RF power, but the typical demand is much lower because all the channels are unlikely to be occupied at the same time. A factor called the *contention ratio* is applied in systems that share common resources, so the satellite might have a maximum RF power available of 1 kW, allowing a contention ratio of two. What this means in practice is that not all 50 beams have 100 active users at the same time; some beams covering more densely populated areas could reach this level, but other beams over sparsely populated regions or the oceans will have very few users.

The inbound link from the user to the gateway station has a margin of 13.7 dB for tree shadowing on the uplink to the satellite. The downlink from the satellite has a blockage margin of 5.5 dB. The Ku-band links between the satellite and the gateway station have large margins, and UPC is used to prevent uplink rain attenuation at 14 GHz from adversely affecting the downlinks to the mobile terminals. TDMA is used on the outbound link with half rate FEC coding and a bit rate of 500 kbps. SCPC-FDMA-DAMA is used on the inbound links, with a channel bit rate of 9.6 kbps after FEC coding is applied.

There are several reasons why this system could not be implemented in practice. The L-band mobile frequencies of 1650 and 1550 MHz are fully utilized by existing systems, notably Inmarsat and Iridium. The broad antenna beam of the mobile terminal would cause interference with these systems, and would also receive interfering signals from them. One advantage of GEO satellites is that narrow beam antennas can be used and any given frequency band can be reused by satellites spaced as close as two degrees. The mobile communication system requires 50 satellites to provide continuous coverage, which must be constructed and launched at a cost that is likely to exceed $1 B. The per minute charges to users to recover the system costs make the satellite system non-competitive with terrestrial cellular service except where cellular service is not available. However, LEO satellite systems have application for internet access in many parts of the world that lack a good terrestrial communication network. Chapter 11 examines this topic.

Texting with cellular telephones is popular. Text transmissions use the short messaging system (SMS) with a bit rate of 19.6 kbps. The bit rate for voice channels in the satellite system is 4.8 kbps, making text transmission much slower than with a cell phone. The bit rate can be increased to 19.6 kps by adopting a higher order modulation, but this requires higher CNR. There is sufficient margin available on the inbound link to implement SMS text transmission at 19.6 kbps, but a lower bit rate would have to be tolerated on the outbound link because of the smaller margin. Dynamic allocation of bandwidth is needed in a SCPC-FDMA-DAMA system, making TDMA more attractive. Compatibility between cell phone transmission techniques and satellite telephones is highly desirable so that the same high density integrated circuits can be used in both applications.

4.9 Summary

This chapter has set out the procedures for calculation of received power from a satellite and noise power in a receiver. Together, these figures give the CNR values for the receiving systems. The specification of a system will always require a minimum CNR in each receiver, below which the link is considered inoperable. The design of a link to achieve that minimum CNR requires repeated calculation of the link and noise

powers to give CNR values for clear air conditions with acceptable bandwidth and antenna dimensions. When a linear (bent pipe) transponder is used, the clear air value of CNR for the uplink and downlink must be combined to give the overall $(CNR)_o$ ratio in the earth station receiver. Once clear air performance has been calculated, the effect of rain on the slant paths must be determined and the propagation path statistics need to be studied to determine how much margin is required to meet worst-case conditions. Examples are presented throughout Chapter 4 showing how link power and noise budgets are used to find the overall $(CNR)_o$ ratio for different systems.

Fading of both uplink and downlink simultaneously is unlikely for 6/4 and 14/11 GHz systems and can safely be ignored when computing link statistics. At 30/20 GHz the possibility cannot be ignored and the joint effect has to be calculated. No attempt has been made to derive an optimization procedure for the design of the "best" system within a frequency band and CNR specification. There are too many variables in the system, including the cost of antennas, receivers, and other components to produce a single optimization procedure. Iterative techniques must be used to find a set of parameters for the earth stations and satellite that provide the performance required from the satellite communication system. The designer of a satellite communication system may have to go through many trial design procedures and compare the resulting systems to determine which one best suits the particular application.

Exercises

4.1 Calculate the path loss for a satellite to earth down link with a distance of 38 500 km at a frequency of 4.0 GHz. What is the uplink path loss if the earth station transmits to the same satellite at a frequency of 6.0 GHz?

4.2 A LEO satellite at an altitude of 400 km has a distance to a receiving terminal of 500 km. The satellite radiates a power of 1.0 W from a spot beam antenna with a gain of 35 dB in the direction of the terminal. Find the flux density at the receiving terminal, and the power received with a tracking antenna with a gain of 30 dB.

4.3 Vacuum tubes are inherently noise devices because they rely on the random emission of electrons from a heated filament to create a current between the cathode and anode. The noise figure of a typical vacuum tube is 16 dB. Calculate its noise temperature.

4.4 A Ku-band DBS-TV receiving terminal has an antenna with a gain of 34 dB at 12.5 GHz. The noise temperature of the receiver under clear air conditions is 80 K.

a. Calculate the G/T ratio of the terminal in dB.

b. If the same reflector is used for reception of a Ka-band DBS-TV signal at 21 GHz, and the terminal has a noise temperature of 200 K at this frequency, what is the terminal G/T ratio, in dB.

4.5 Ku-band satellite DBS-TV with regional beams.
The link budget in Table 4.6a details the performance of the downlink from the satellite to a customer's receiving system using DVB-S format signals. With a

later design of receiver, the DVB-S2 standard can employed with 8-PSK modulation and 2/3 rate FEC, which delivers two bits for each hertz of noise bandwidth. QPSK and half rate FEC delivers one bit per hertz of noise bandwidth.

a. If the signal format is changed from DVB-S to DVB-S2 what is the new bit rate, assuming all other parameters stay the same. What is the new link margin?

b. A receiving terminal is located in Mexico on the −6 dB contour of the satellite's regional beam. What receiving antenna gain is required to achieve the same performance as a station in the United States located on the −3 dB contour of the satellite beam, assuming all other parameters are unchanged? Estimate the diameter of a reflector antenna with a circular aperture and this gain.

c. The Mexican receiver is re-equipped with a new LNA with a noise temperature of 64 K. What diameter antenna is required now to achieve the same performance as a station in the United States located on the −3 dB contour of the satellite beam, assuming all other parameters are unchanged?

4.6 DBS-TV receiving location on −7 dB contour of satellite footprint

A DBS-TV receiver is located on the −7 dB contour of a DBS-TV satellite. The system parameters are the same as those in Table 4.6a except that the receiving antenna diameter is increased to 0.9 m and the minimum permitted CNR on the downlink is 8.0 dB.

a. Calculate the gain of the receiving antenna at 12.5 GHz with an aperture efficiency of 70%.

b. Calculate the received power and downlink CNR for this station in clear sky conditions.

c. Calculate the downlink CNR with 2 dB rain attenuation on the path.

d. Using the iterative procedure in Section 4.5.2 find the rain attenuation margin.

4.7 An earth station equipped with UPC transmits at 17.0 GHz to a DBS-TV satellite with a beacon transmitter at 12.7 GHz.

a. Calculate the ratio of uplink attenuation to down link attenuation for these frequencies, using a value $a = 2.4$ in Eq. (4.4).

b. In clear air, the transmit power is 300 W. The saturated output power of the transmitter is 3 kW and 3 dB backoff is always applied to the transmitter output. What is the dynamic range of the UPC?

4.8 A Ku-band satellite carries a bent pipe transponder with 36 MHz bandwidth, which is operated in its linear region. In clear air conditions, the uplink CNR is 24.0 dB and the downlink CNR at a receiving earth station is 14.0 dB.

a. Calculate the overall $(CNR)_o$ in clear sky conditions.

b. Rain affects the uplink causing 4 dB of attenuation. Calculate the overall $(CNR)_o$ at the receiving earth station.

c. Rain affects the downlink causing 4 dB of attenuation. Calculate the overall $(CNR)_o$ at the receiving earth station.

d. Interference from a Ku-band line of sight link occasionally affects the receiving earth station. If the interference creates a C/I ratio in the receiver of 24 dB, recalculate your answers for parts a, b, and c above.

e. One of the transponders on the satellite has AGC. The AGC holds the output power from the transponder constant over a 10 dB range of input power.

Recalculate your answers to parts a, b, and c above with the AGC equipped transponder.

4.9 Video distribution with a Ku-band satellite

The parameters in Table 4.8a define the video distribution system. Make the following changes. Uplink earth station diameter 7.0 m with aperture efficiency 63%. Receiving earth station diameter 3.0 m with 60% aperture efficiency. Earth station HPA output power 100 W. Transponder bandwidth 100 MHz, transponder saturated output power 100 W. The uplink and downlink earth stations are at a distance of 38 000 km from the satellite. Give all final answers to the nearest 0.1 dB.

- **a.** Calculate the gain of the uplink earth station antenna at a frequency of 14.0 GHz, and the gain of the downlink earth station antenna at a frequency of 11.0 GHz. Calculate path loss for the uplink and downlink.
- **b.** The signal transmitted by the uplink earth station has a bandwidth of 50 MHz and its power level at the input to the transponder is set such that the output from the transponder is half of its operating output power. The transponder is operated with 3.0 dB backoff. Find the output power of the transponder for this signal.
- **c.** Set out the uplink and downlink power and noise budgets similar to Tables 4.8c and 4.8d and find the uplink CNR and downlink CNR in clear air conditions. The receiver noise bandwidth is 50 MHz. Then calculate the overall CNR for the receiving earth station.
- **d.** Rain attenuation of 5 dB affects the uplink. What is the overall $(CNR)_o$ at the receiving earth station, assuming linear operation of the transponder?
- **e.** Rain attenuation of 3 dB affects the downlink. What is the overall $(CNR)_o$ at the receiving earth station, assuming linear operation of the transponder?
- **f.** What is the downlink rain margin for a threshold (minimum required) $(CNR)_o$ of 8.0 dB?

4.10 Modified L-band mobile communication system

The mobile satellite communication system described in Section 4.8 is modified in the following ways. Maximum path length to the mobile device is reduced to 2000 km, mobile device antenna gain is increased to 6 dB in the elevation angle range 10°–30° for both receive and transmit, and the number of satellite L-band beams is increased to 200 with same footprint as before.

- **a.** Calculate the new path loss values at 1550, 1650 MHz, 11.5 GHz and 14.0 GHz.
- **b.** Find the new L-band satellite antenna gains and beamwidths, and the antenna diameter at each L-band frequency assuming a circular aperture.
- **c.** Calculate the new $(CNR)_{up}$ for the L-band uplink and the overall $(CNR)_o$ at the gateway earth station receiver. What is the increase in overall CNR for the inbound link with the new system parameters?
- **d.** Repeat the calculations in part (d) above for the outbound link.
- **e.** The improvement in overall CNR is traded for an increase in bit rate to 19.6 kbps so that the mobile terminal can send text messages that are compatible with the SMS service of cell phones. The available channel bandwidth using SCPC-FDMA-DAMA is 10 kHz. FEC encoding is applied to the 19.6 kbps message signal making its bit rate 39.2 kbps after FEC coding is applied. The SRRC filters in the link have $\alpha = 0.25$. How many bits per symbol

are required to reduce the transmission bandwidth below 10 kHz? What is the available guard band with this number of bits per symbol? Is a reduction in the number of channels needed to meet this requirement?

f. The modulation on the link is changed to 32-APSK (amplitude phase shift keying), which requires CNR = 22.5 dB for BER = 10^{-4}. What is the overall CNR margin on the inbound and outbound links with the receiver threshold at 3.0 dB? (See Section 6.1 for a discussion of higher order PSK modulations.)

4.11 Calculate the power received by the earth station when the satellite in Example 4.1 has a regional beam with a gain of 26 dB.

i) What is the diameter of a circular aperture antenna with a gain of 26 dB at 11 GHz?

ii) What is the effective receiving area of the earth station antenna with a gain of 52.3 dB?

iii) What is the diameter of the earth station antenna if its aperture efficiency is 65%?

iv) If we increase the earth station antenna diameter by 20%, what is its new gain and the new received power in dBW and dBm?

4.12 Spread sheet for link budgets

Once the process of calculating link budgets is thoroughly understood, a spread sheet is a useful way to calculate link budgets, but needs to be tested thoroughly to ensure that it produces answers that are always correct. Generate a link budget spread sheet using Excel® or MatLab® and enter the data in Tables 4.6a and 4.6b for the DBS-TV downlink. Print out intermediate results for antenna gain, path loss, received power, noise power, and CNR. Check that the results from the spread sheet are the same as those in Tables 4.6a and 4.6b. Repeat the check using Tables 4.8c and 4.8d for the video distribution system uplink and downlink.

4.13 Moon to earth link

a. The average distance from the moon to earth is 384 400 km. A spacecraft with an S-band transmitter is located on the moon and operates at 2295 MHz, a frequency assigned to space to earth links for space research. The transmitter output power is 10 W and a steerable reflector antenna with a diameter of 1.0 m and aperture efficiency 60% on the spacecraft points toward earth whenever the receiving terminal on earth is visible. A receiving antenna on earth with a system noise temperature of 25 K is used to receive the spacecraft transmissions. Turbo coding of the data transmitted from the spacecraft allows the threshold CNR to be 3.0 dB.

Create a table of receiving antenna diameters with an aperture efficiency of 60% and maximum data rate for the link. Which combination would you recommend for this project?

b. Repeat the exercise for a spacecraft on the surface of Mars, using a distance of 100 million kilometers. The distance between earth and Mars varies between its closest approach of 57 M km and its most distant point at 100 M km.

References

Allnutt, J.E. (2011). *Satellite-to-Ground Radiowave Propagation*, 2e, Chapter 6. London, UK: IET.

Dicks, J. and Brown, M. (1975). INTELSAT IV-A transmission system design. *Comsat Technical Review* 5: 73–103.

FCC Online Table of Frequency Allocations (2017). http://www.transition.fcc.gov/oet/spectrum/table/fcctable.pdf (accessed 24 January 2018).

Glover, I.A. and Grant, P.M. (1998). *Digital Communications*. Hemel Hempstead, UK: Prentice Hall Europe.

ITU Radio Regulations (2001). https://www.itu.int/en/ITU-R/software/Pages/ant-pattern.aspx (accessed 3 February 2018).

ITU Radio Regulations (2017). http://www.itu.int/pub/R-REG-RR (accessed 23 January 2018).

Kanipe, D.B. (2014). *Estimating-the-Cost-of-Space-Systems*. Department of Aerospace Engineering, Texas A & M https://engineering.tamu.edu/media/1832106/Estimating-the-Cost-of-Space-Systems_2014.pdf (accessed 3 February 2018).

Krauss, H.L., Bostian, C.W., and Raab, F.H. (1980). *Solid State Radio Engineering*. New York, NY: Wiley.

Maral, G. and Bousquet, M. (2002). *Satellite Communication Systems*, 4e. Chichester, UK: Wiley.

NASA earth observations (2013). https://neo.sci.gsfc.nasa.gov/view.php?datasetId=MOD11C1_M_LSTDA&year=2013 (accessed 6 February 2018).

Rappaport, T.S. (2002). *Wireless Communications, second edition*. Upper Saddle River, NJ: Prentice Hall.

Recommendation ITU-R S 465.5 (1993). Reference Earth Station Antenna Patterns 3–30 GHz.

Recommendation ITU-R S.1328-3 (2001). Satellite system characteristics to be considered in frequency sharing analyses between geostationary-satellite orbit (GSO) and non-GSO satellite systems in the fixed-satellite service (FSS) including feeder links for the mobile-satellite service (MSS).

Recommendation ITU-R S.1591 (2002). Sharing of inter-satellite link bands around 23, 32.5 and 64.5 GHz between non-geostationary/geostationary inter-satellite links and geostationary/geostationary inter-satellite links.

Recommendation ITU-R S.524-6 (2000). *Maximum Permissible Levels of Off-Axis EIRP Density from Earth Stations in GSO Networks Operating in the Fixed-Satellite Service Transmitting in the 6 GHz, 14 GHz and 30 GHz Frequency Bands*. Geneva, Switzerland: ITU.

de Selding, P.B. (2016). http://spacenews.com/spacexs-new-price-chart-illustrates-performance-cost-of-reusability (accessed 4 February 2018).

Silver, S. (ed.) (1949). *Microwave Antenna Theory and Design*, vol. 12. MIT Radiation Lab Series. (Republished by Peter Perigrinus, Stevenage, Herts, UK, 1984).

Stutzman, W.L. and Thiele, G.A. (2013). *Antenna Theory and Design*, 3e. Hoboken, NJ: Wiley.

5

Digital Transmission and Error Control

Communications satellites are used to carry telephone, video, and data signals, and can employ both analog and digital modulation techniques. When geostationary earth orbit (GEO) satellites were first used for communications in the 1960s and 1970s, the signals were mainly analog. The advent of satellite communications made possible the transmission of wide bandwidth signals between continents. For the first time, video signals could be sent between North America, Europe, and Asia. Thousands of telephone channels could be multiplexed through one transponder and sent across the United States or across the Atlantic or Pacific Oceans. The modulation and multiplexing techniques that were used at that time were analog, adapted from the technology developed for microwave links in the previous two decades. Frequency modulation (FM) was the modulation of choice and *frequency division multiplexing* (FDM) was used to combine hundreds or thousands of telephone channels onto a single microwave carrier. Regional domestic and international satellite systems were developed to exploit the high capacity and bandwidth that satellites offered.

In the 1980s, optical fibers came into widespread use, and GEO satellites were no longer used for telephony within the United States. The long round trip delay of 500 ms in a typical GEO satellite voice link could cause echoes that were unsettling to many telephone users, so GEO satellite telephone links were restricted to routes that cannot use optical fibers over the majority of the route. Echoes occur because of mismatches between different pieces of equipment along the communication path, especially at the receiving end of the call. On a terrestrial telephone link, echoes typically occur with a delay of a few milliseconds and are not noticeable. On a GEO satellite voice link, the echo is heard as the speaker's voice delayed by half a second. Two techniques for reducing the nuisance caused by echoes were developed. Echo cancellers were installed by US telephone companies and proved effective, so callers to the United States were not troubled by echoes. European telecommunication authorities installed echo suppressors which were less effective and resulted in noticeable echoes on calls made from the United States to Europe.

Long distance telephone links using optical fibers are digital, so all telephone signals sent via optical fiber have to be in digital form. At the same time as the migration to optical fibers occurred, telephone exchanges became digital computers instead of large banks of mechanical switches. The change to digital voice signals made it easier for long distance communication carriers to mix digital data and telephone traffic and send it through the same optical fibers and telephone exchanges. This forced telephone signals to be converted to digital form at the telephone exchange, and rendered all the analog

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

multiplexing methods obsolete. FDM has disappeared as a way to combine analog telephone signals, replaced by time division multiplexing (TDM) of digital voice signals. In the first edition of *Satellite Communications*, the frequency modulation – frequency division multiplexing (FM-FDM) techniques were covered in detail and in the second edition in Appendix B (Pratt and Bostian 1986, pp. 156–176; Pratt et al. 2003, Appendix B). These techniques are now obsolete and are omitted in this edition, but there may be parts of the world where the older analog technology is still in use. An internet search for FDM/FM/FDMA will provide a multitude of sources that describe analog communication technology, and many texts on communication systems cover this technology (Couch 2007; Haykin 2001).

The distribution of television program material in North America and much of the rest of the world is by satellite. Satellites are particularly useful for distributing the same signal to hundreds or thousands of receivers (*point to multipoint*, broadcasting) and many of the world's GEO satellite transponders are used for video distribution to cable television systems. Direct broadcast satellite television (DBS-TV) developed rapidly in the 1990s after the introduction of digital video transmission and had over 33 million subscribers in the United States in 2017, with many millions more in other countries. DBS-TV is an example of what GEO satellites do best: broadcasting. Several GEO satellites can provide hundreds of entertainment channels to the whole of North America, where there are more than 100 million households, or to any other continent. Satellite TV is especially valuable to people who live in rural areas where there is no access to cable TV. Chapter 10 describes DBS-TV in detail.

Analog multiplexing in the form of FDM has virtually disappeared, but frequency division multiple access, FDMA, remains one of the major ways in which transponder capacity is shared among users. FDMA divides up the frequency band in the transponder into channels which are allocated to different signals on a fixed or on-demand basis. Multiple access techniques are discussed in Chapter 6: this chapter concentrates on the digital transmission techniques that are used for voice, video, and data signals. Since most signals are now transmitted digitally, analog signals must first be converted to digital form and that process is also described. Once an analog signal is in digital form, it can be transmitted over any digital communication link, multiplexed with other digital signals, and sent very long distances without degradation. One major advantage of digital transmission systems over analog is that *error free transmission* is possible.

Digital signals can be compressed to reduce the bandwidth required to transmit the signal, an essential feature in mobile telephones and in the transmission of video signals. Digital signals can easily be encrypted to maintain security of the message content, while effective encryption of analog signals is very difficult. In a digital telephone system, error free transmission means that no noise is injected into the baseband channel, regardless of the transmission distance, so a telephone call over a distance of 10 000 km has the same quality as a call over a distance of 10 km. This is not the case when analog transmission is used.

The design of digital communication links requires a different approach from the design of analog links, although achieving optimum performance in a digital communication system requires a link with linear characteristics – an analog circuit rather than a digital circuit. The techniques for the design of digital communication systems are covered in some detail because a good understanding of these methods is essential in setting carrier to noise ratios (CNRs) correctly for a digital satellite link, and for estimating the bit error rate that can be expected. The topics of bit error rates and the

use of error detecting and error correcting codes for error control are covered in this chapter. There is also a review of the compression techniques used for digital voice and video signals. For a more extensive treatment, the reader is referred to any of several excellent texts on communication theory and digital communication systems (Couch 2007; Haykin 2001; Lathi and Ding 2009; Haykin and Moher 2009).

5.1 Digital Transmission

Digital modulation is the obvious choice for satellite transmission of signals that originate in digital form and that are to be used by digital devices. Virtually all signals sent via satellites are now digital. Familiar examples are data transmissions to and from internet hubs, communications between remote terminals and computers, digital telephony, and TV signals in digital form, such as high definition television (HDTV) and DBS-TV. Digital transmission lends itself naturally to TDM and *time division multiple access* (TDMA). Analog signals that are transmitted digitally can share channels with digital data, since all digital signals are handled in the same way, and their content is immaterial. Thus a digital satellite link can carry a mix of telephone, video, and data signals that varies with traffic demand.

This section contains a review of digital transmission techniques. All digital links are designed in much the same way, using a specific *symbol rate*, and specific filters and waveforms that minimize *intersymbol interference* (ISI). Signals are generated at baseband as pulses of current or voltage. A digital baseband voice signal, for example, could consist of voltage pulses of 0 and +3.3 V, typical voltage levels used by logic circuits, or pulses of light on an optical fiber representing the logical state 1 and no pulse representing the logical state 0. For transmission over a wire link, baseband voltage pulses need to be symmetrical in the form $+V$ and $-V$ volts to avoid the generation of a DC level. A $+V$ or $-V$ voltage pulse is a *symbol* in a digital link, representing a single bit from a bit stream. Symbols can have multiple voltage levels so that one symbol can represent more than one bit. For example, two bits can be represented by a voltage pulse that has four levels: +3 V, +V, −V, and −3 V. This is described as two bits per symbol transmission, and has the advantage of a higher bit rate than a binary link using only two voltage levels. Systems with 10 or more bits per symbol are in widespread use. As we shall see later, the more bits per symbol there are, the higher the carrier to noise ratio (CNR) required at the receiver.

In a digital radio link, a symbol is almost always a phase state or a phase and amplitude state. For historical reasons related to the telegraph, digital transmission methods are referred to as keying. *Binary phase shift keying* (BPSK) and *quadrature phase shift keying* (QPSK) send one bit and two bits per symbol as phase states of a radio wave, called a *carrier*. Higher levels carrying more bits per symbol are also used, for example, 8-PSK has eight phase states and carries three bits per symbol. A combination of multiple voltage levels and phase states can be employed to carry a large number of bits per symbol using quadrature amplitude modulation (QAM) or amplitude phase shift keying (APSK). For example, 4-QAM uses four phase states and one carrier amplitude (V) in a rectangular constellation and is therefore QPSK, which transmits two bits per symbol. 16-APSK uses eight phase states and two carrier amplitudes (V and 3 V) in a circular constellation to transmit four bits per symbol. The topic of modulation is explored in more depth in Chapter 6.

It is important to distinguish between symbols and bits. They are easily confused because with binary modulations, such as BPSK, the symbol rate and bit rate are the same. Symbol rates are given in *baud* (largely obsolete) or in *symbols per second*, abbreviated to *sps*.

We will use symbols per second in the analysis that follows because the difference between symbols per second and bits per second is then more obvious, with the comment that the baud as the unit of transmission rate is equal to symbols per second and is still in use. The name of the unit, the baud, is derived from the name Baudot, who was an early French pioneer of the telegraph, born in 1845 (Baudot 2018).

5.1.1 Baseband Digital Signals

We will represent baseband digital signals as serially transmitted logical ones and zeroes. While in computer circuitry a logical zero may be represented by a low voltage (nominally zero) and a logical one may be represented by a high voltage (e.g., 3.3 V), this arrangement is inconvenient for transmission over any significant distance and is not used. To understand why, imagine a transmission line carrying a bit stream encoded this way and containing approximately equal numbers of ones and zeroes. About half the time the line voltage will be 3.3 V and about half the time it will be 0 V; hence the line signal will have a 1.65 V DC component. All circuits that carry this signal must have a frequency response that extends to DC, and this is difficult to achieve since many communication circuits contain transformers or are not DC coupled. To avoid this problem, digital modulators usually accept their input in a polar non-return-to-zero (NRZ) format: logical ones and zeroes are transmitted as plus or minus a stated voltage value. Thus a logical one might be transmitted as +1 V and a logical zero might be transmitted as −1 V. Zero volts is not transmitted except as a transient value. Throughout this text we will assume a polar NRZ format for data signals unless we explicitly state otherwise. Since a receiver has only to decide whether a voltage pulse is positive or negative by using a *threshold* of zero volts, the actual value of V is unimportant in a binary system.

5.1.2 Baseband Transmission of Digital Data

Satellite links always consist of RF signals, which requires that data be modulated onto a radio frequency carrier for transmission. However, to provide a better understanding of the way in which digital transmission systems are designed, we will begin by examining the case of a baseband data link. In a baseband link, the frequency response of the link is assumed to extend from DC to an upper limit $f_{\max}$, where $f_{\max}$ is equal to the bandwidth of the link, B Hz. Data is transmitted in the form of polar pulses; in a binary system, the pulses have amplitudes $+V$ and $-V$ volts, where V can take any value. As mentioned earlier, the average number of $+V$ and $-V$ pulses is made equal so that the average DC voltage on the transmission line is zero. Because the link has a finite bandwidth, we cannot transmit rectangular voltage or current pulses, since those have infinite bandwidth – the Fourier transform of a rectangular pulse yields a spectrum that extends from 0 Hz to infinity. We must therefore transmit a pulse that ideally has all its energy confined to a finite bandwidth of B Hz. Readers unfamiliar with the relationship between waveforms and their frequency spectra should refer to a text on communication theory such as (Couch 2007; Haykin 2001).

We begin the analysis of baseband links by determining the conditions under which ISI can be minimized. ISI occurs when part of one pulse carries over to the time interval of the next pulse. The numerical results for bandwidth and symbol rate in this section do not correspond to the requirements for a satellite link; they refer only to a baseband link using a wire transmission line.

The random sequence of rectangular binary pulses shown in Figure 5.1 has a power spectral density

$$G(f) = T_s \left[\frac{\sin(\pi f T_s)}{\pi f T_s} \right] \tag{5.1}$$

where T_s *is* the duration of the pulse (Couch 2007). This spectrum is illustrated in Figure 5.1d.

The familiar sin x/x spectrum (and the identical spectrum of sinc $x = \sin \pi x/\pi x$) shows that energy exists at all frequencies; to retain the rectangular pulse shape would require an infinite transmission bandwidth. Practical communication systems are always bandwidth limited. Not only is infinite bandwidth not available, interference considerations in radio links dictate that a communication system should use the smallest possible bandwidth, and this is usually one of the design criteria of a radio communication system.

In any digital communication system, symbols are sent over the link in the form of voltage or current pulses at baseband, or changes in phase angle of a carrier, for example, in a phase shift keying (PSK) system. Our discussion of digital transmission systems will be based on symbols, rather than bits, because we often want to send more than one bit per symbol in an RF system to conserve bandwidth. *Nyquist's criterion* for zero ISI, which forms the design basis for every digital transmission system, is based on the use of *square root raised cosine* (SRRC) filters and corresponding waveforms, and a specified symbol rate on the link (Shanmugam 1979, p. 195). If the transmission is binary, the symbol is a bit, and the symbol has two states. When two bits are sent per symbol, the symbols have four possible states and the system is denoted as quaternary, hence one reason for the Q in QPSK. (QPSK is better known as *quadrature phase shift keying*.) If a symbol represents more than one bit, the system is known generically as *m-ary*, with one symbol having *m* states.

If we take the random pulse train shown in Figure 5.1a and bandlimit it by passing the pulses through a low-pass filter, the pulse shape will be altered. As an example, consider the effect of passing the rectangular pulse train through a single *resistor-capacitor* (RC) section, representing a very simple low pass filter. The resulting waveform, shown in Figure 5.1b, has been delayed and pulses are smeared in time – the decaying pulse from one transition extends into the next pulse interval. Where the pulse pattern is 10 or 01, the amplitude of the second pulse at the sampling instant shown in Figure 5.1b has been reduced by the presence of a delayed portion of the preceding pulse. This is called inter-symbol interference (ISI) and is likely to occur whenever a digital signal is passed through a band-limiting filter. When noise is added to the waveform, ISI increases the likelihood that the receiver will detect a bit incorrectly, causing a bit error. In a baseband system, ISI can be avoided by an appropriate choice of low-pass filter or waveform. In 1928, Nyquist was looking for ways to improve telegraph transmission over telephone lines and proposed a technique that can theoretically produce zero ISI, now known as

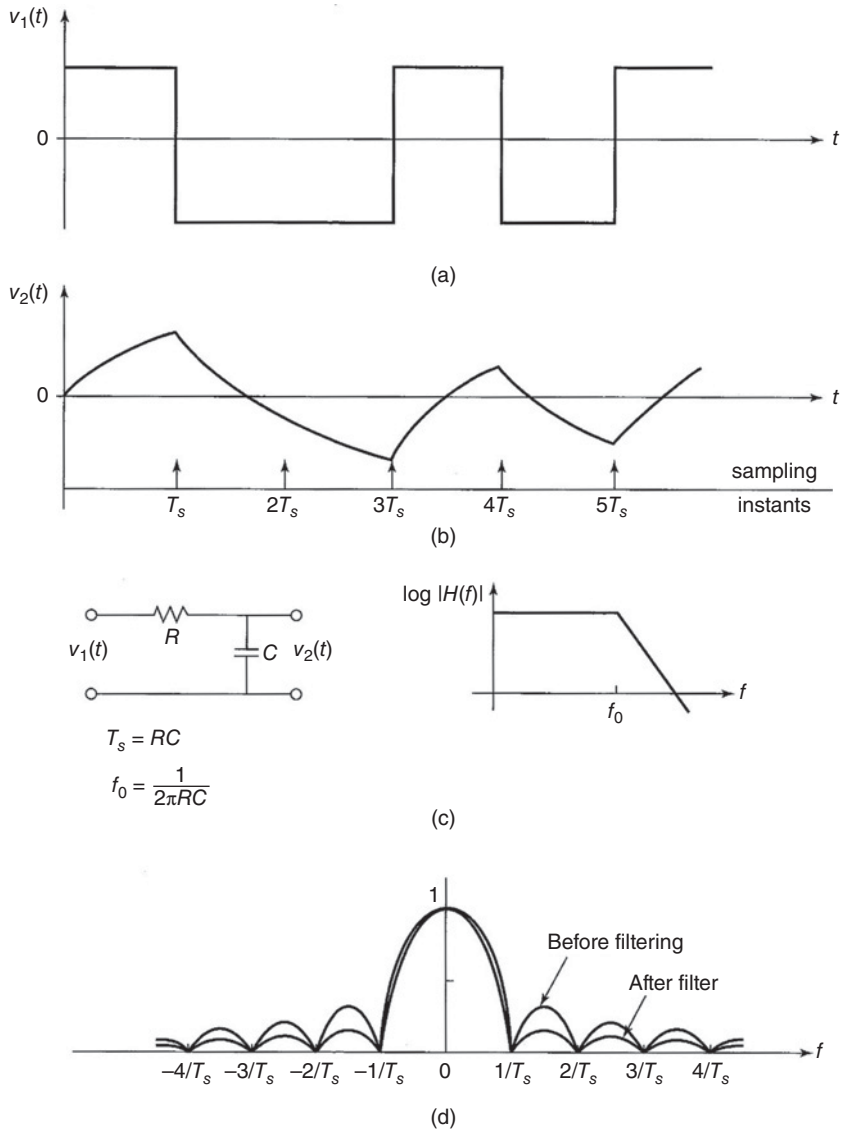


Figure 5.1 Illustration of the effect of low pass filtering on a NRZ pulse train. (a) Random NRZ pulse train. (b) Waveform output from an RC filter with $T_s = RC$ and cut off frequency $f_o = 1/(2\pi RC)$. (c) RC filter with transfer function $H(f)$. (d) Spectra of pulse trains before and after the RC filter.

the *Nyquist criterion* (Nyquist 1928). The objective is to create in the receiver a pulse that resembles the sin x/x shape, crossing the axis at intervals of T_s, where T_s, is the symbol period. The receiver samples the incoming wave at intervals of T_s, as shown in Figure 5.2, so that at the instant one pulse is sampled, the tails from all preceding pulses have zero value. Thus previous pulses cause zero intersymbol interference (zero ISI) at each sampling instant.

Filters that produce the required zero ISI waveform in the receiver can be realized in several ways. The baseband transfer function proposed by Nyquist was the *raised*

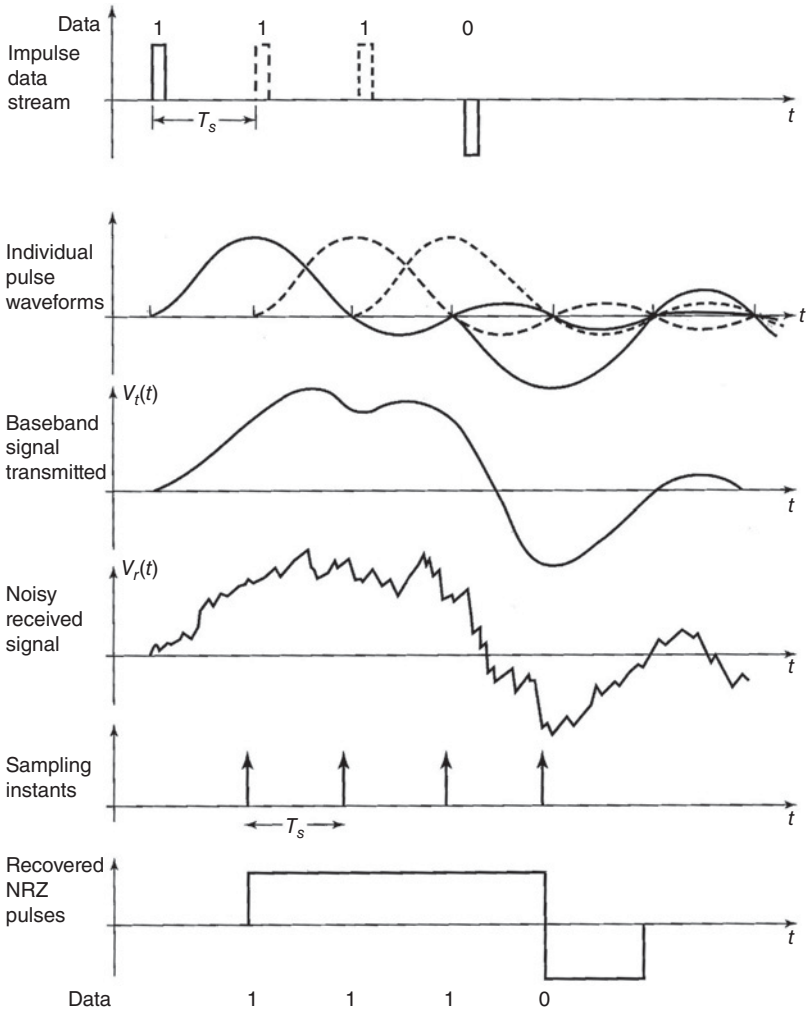


Figure 5.2 Transmission and reception of baseband zero ISI pulses. The data stream is the sequence 1 1 1 0 with a pulse period T_s. The transmitted pulses have square root raised cosine shape with zero crossings at every increment of T_s. A data 1 is sent as a positive waveform and a data zero as a negative waveform. The received waveform is noisy but sampling at the correct instants recovers the data stream correctly. The time sidelobes that precede the main lobe of the zero ISI pulses are not shown.

cosine function, $V_{\mathrm{NQ}}(f)$, which has a normalized two sided frequency characteristic given by

$$V_{\mathrm{NQ}}(f) = 1 \quad \text{for } |f| < \frac{R_s}{2}(1-\alpha)$$

$$V_{\mathrm{NQ}}(f) = cos^2\left\{\frac{\pi}{2\alpha R_s}\left[|f| - \frac{R_s}{2}(1-\alpha)\right]\right\}$$

$$\text{for } \frac{R_s}{2}(1-\alpha) \leq |f| \leq \frac{R_s}{2}(1+\alpha)$$

$$V_{\mathrm{NQ}}(f) = 0 \quad \text{for } |f| > \frac{R_s}{2}(1+\alpha) \tag{5.2}$$

where $0 < \alpha < 1$ and $R_s = 1/T_s$ is the symbol rate in symbols per second. The parameter α is called the *roll-off* factor. The name *raised cosine* comes from the squared cosine term in Eq. (5.2). (Many texts and papers use the symbol *r, β, or ρ* instead of *α,* each with a different definition, and present the equations in a different form that does not include a squared term, obscuring the reason for the name *raised cosine*.) The entire communication link must have this transfer function to ensure zero ISI. The pulse shape generated at the output of the link is $v_{NQ}(t)$, the required zero ISI waveform, when the filter input is driven by an impulse, $\delta(t)$. The waveform $v_{NQ}(t)$ is obtained as the inverse Fourier transform (the *impulse response*) of the output from the Nyquist raised cosine transfer function, which is simply the spectrum of the input pulse multiplied by the frequency response of the system.

$$v_{NQ}(t) = F^{-1}[V_{NQ}(f) \times S(f)] \tag{5.3}$$

where $F^{-1}[]$ indicates the inverse Fourier transform and $S(f)$ is the spectrum of the input pulse. If we use an impulse $\delta(t)$ as the input signal, the input signal spectrum is $S(f) = 1$, which is referred to as a *flat spectrum,* and then

$$v_{NQ}(t) = F^{-1}[V_{NQ}(f)] \tag{5.4}$$

The raised cosine transfer function with $V_{NQ}(f)$ given by Eq. (5.2) has an impulse response $v_{NQ}(t)$ given by (Couch 2007, p. 183)

$$v_{NQ}(t) = \left[\frac{\cos(\pi\alpha R_s t)}{1 - (2\alpha R_s t)^2}\right] \times \left[\frac{\sin(\pi\alpha R_s t)}{\pi\alpha R_s t}\right] \tag{5.5}$$

Figure 5.3 shows the shape of several raised cosine function characteristics for values of α between 0 and 1, and the corresponding waveforms generated by the impulse response of these filters. Note that the raised cosine function shown in Figure 5.3 is a voltage transfer function, and that all functions have a value $V_{NQ}(f) = 0.5$ at a baseband frequency $f = R_s/2$ Hz. The case of $\alpha = 0$ in Eq. (5.5) yields a rectangular transfer function with a bandwidth of $R_s/2$ Hz. This is the minimum bandwidth through which a baseband signal with a symbol rate R_s can be transmitted while still satisfying the zero ISI condition. Such a function is not realizable in practice, since we cannot have an infinitely rapid attenuation slope at one frequency. (Texts on communication theory call this *non-causal*.) In fact, none of the Nyquist raised cosine functions can be created in practice. The requirement in Eq. (5.2) that there be zero transmission above a frequency $(R_s/2) \times (1 + \alpha)$ Hz cannot be met with any real circuit. Consequently, all digital transmission systems can only approximate the ideal Nyquist transfer function, and will therefore always generate some ISI. However, the basis for the design of all digital links is the ideal zero ISI Nyquist criterion. Real filters that give the link a transfer function that approximates the ideal Nyquist zero ISI transfer function will minimize ISI and maximize symbol rate. The Nyquist raised cosine function is not the only transfer function that can create a zero ISI waveform at the receiving end of a link. Any transfer function that possesses the same symmetry as the Nyquist raised cosine function has this property. Specifically, the normalized baseband transfer function must have a magnitude 0.5 at a frequency $f_{max}/2$, and reverse symmetry about this point. The reader is reminded again that these results apply to a baseband link, not to a satellite link.

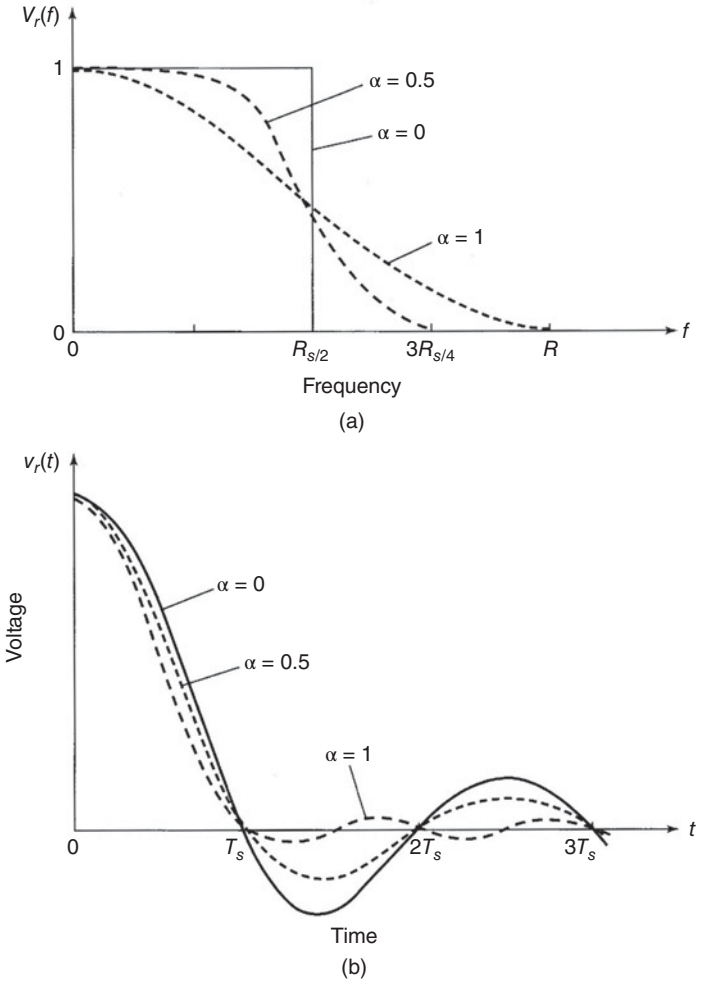


Figure 5.3 Baseband raised cosine transfer function (frequency response) and impulse response for a signal with period $T_s = 1/R_s$. (a) Frequency response for α values of 0, 0.5, 1.0. (b) Raised cosine waveform for α values of 0, 0.5, and 1.0. Note zero crossings occur at intervals of T_s. This is the ideal zero ISI waveform in the baseband receiver at the output of the SRRC filter.

Implementation of a link with a Nyquist transfer function requires specific parts of the link to have different transfer functions. There are three separate parts to any communication system: the transmitter, the transmission link, and the receiver. These three parts are in series, so the overall system transfer function is the product of the three individual transfer functions. We will denote their transfer functions as $H_t(f)$ for the transmitter, $L(f)$ for the transmission link, and $H_r(f)$ for the receiver. We want the output of the receiver to be a zero ISI waveform, which we achieve by creating a zero ISI spectrum $V_{NQ}(f)$ at the receiver output.

The spectrum of the waveform at the output of the receiver is given by

$$V_r(f) = S(f) \times H_t(f) \times L(f) \times H_r(f) \tag{5.6}$$

where $S(f)$ is the spectrum of the signal at the input of the transmitter. We will specify that $S(f) = 1$, corresponding to an input consisting of delta functions $+\delta(t)$ or $-\delta(t)$ representing data ones and zeroes. We will also specify that the transfer function of the link must be flat, such that $|L(f)| = 1$ and that the phase response of the link is linear with frequency. With these conditions in place, we want the end-to-end transfer function of the link to be a Nyquist zero ISI raised cosine transfer function.

Hence

$$V_r(f) = S(f) \times H_{\mathrm{t}}(f) \times L(f) \times H_{\mathrm{r}}(f) = 1 \times H_{\mathrm{t}}(f) \times 1 \times H_{\mathrm{r}}(f) = V_{NQ}(f) \tag{5.7}$$

or

$$V_{\mathrm{NQ}}(f) = H_{\mathrm{t}}(f) \times H_r(f) \tag{5.8}$$

Equation (5.8) is an important result in the design of digital communication links. It states that the transfer functions of the transmitter and receiver when multiplied together must equal the Nyquist raised cosine transfer function. One obvious way to achieve this result is to make the transfer functions of the transmitter and receiver identical, so that

$$H_t(f) = H_{\mathrm{r}}(f) = \sqrt{V_{\mathrm{NQ}}(f)} \tag{5.9}$$

A filter with a transfer function equal to the square root of a raised cosine function is called a *square root raised cosine filter* (a SRRC filter) or often just a *root raised cosine* (RRC) filter. SRRC filters are used as the basis for the design of most digital communication links, even though no such filters actually exist. Real filters, both in analog and digital form (such as *Butterworth, Chebychev, Elliptic function,* etc.) can only approximate the SRRC filter's transfer function.

Many textbooks and papers assume that the reader has a good knowledge of the foregoing analysis, and make statements such as "assuming ideal SRRC filtering" when discussing digital communication links. Ideal SRRC filtering can never be achieved and the performance of all digital links is never quite as good as predicted on that theoretical basis.

Using identical filters in the transmitter and receiver satisfies another desired criterion in communication systems: the *matched filter* criterion (Couch 2007, pp. 448–449; Haykin and Moher 2009, pp. 509–511). A communication link achieves the best possible SNR or bit error rate at the receiver when the spectrum of the signal at the receiver input is a replica of the transfer function of the receiver. The easiest way to achieve this condition is with identical (matched) filters in the transmitter and receiver. The resulting structure for a baseband digital communication link is shown in Figure 5.4, together with the corresponding waveforms and spectra. Because the SRRC filters in the transmitter and receiver have zero transmission above the frequency $f_{\max} = R_{\mathrm{s}}/2 \times (1 + \alpha)$ Hz, the requirement that the link transfer function have a flat magnitude and linear phase extends only up to $f_{\max}$. Similarly, the delta functions at the input can be narrow rectangular pulses with a sin x/x spectrum that has a first null well above f_{max}. A useful rule of thumb is that a pulse of duration $T_{\mathrm{s}} = 1/10\, f_{\max}$ is sufficiently short that its spectrum approximates to a delta function.

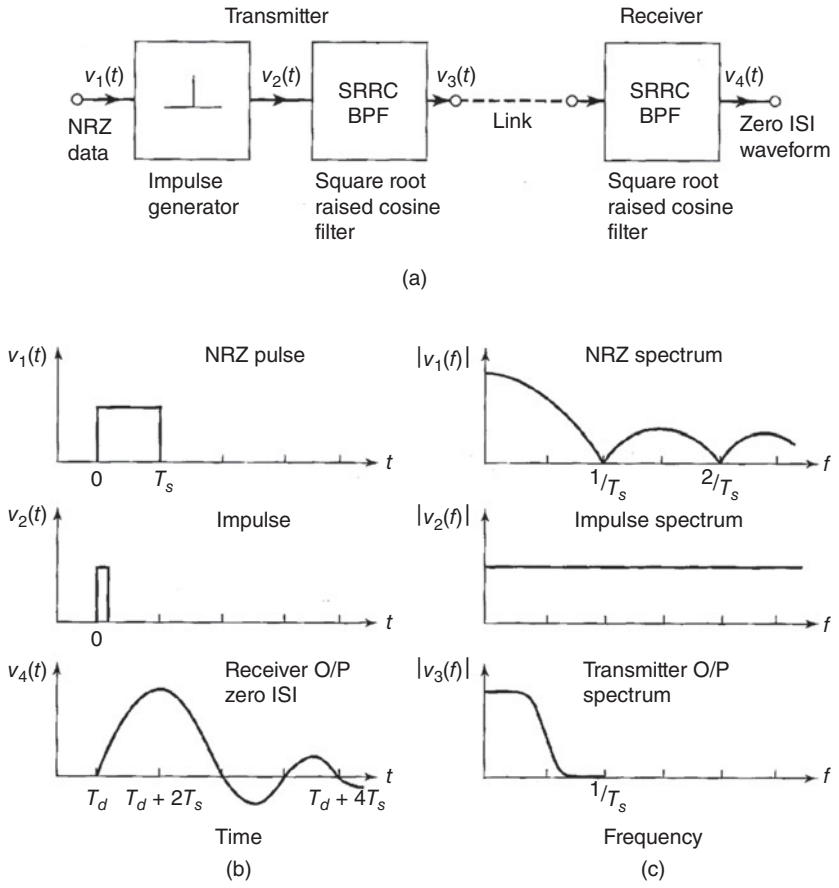


Figure 5.4 Waveforms and spectra in a baseband digital transmission system using SRRC filters. (a) Block diagram of system. (b) The NRZ input pulse $v_1(t)$ is converted to a very short pulse $v_2(t)$ that approximates an impulse. The waveform $v_4(t)$ at the output of the SRRC filter in the receiver is a raised cosine waveform. (c) Spectra of the NRZ input signal $V_1(f)$, the impulse spectrum $V_2(f)$ and the transmitted spectrum $V_3(f)$. The time domain sidelobes before time T_d in (b) are omitted.

5.1.3 Radio Frequency Transmission of Digital Data

In a radio frequency communication system that transmits digital data, a parameter of the RF wave must be varied, or modulated, to carry the baseband information. The most popular choice of modulation for a digital satellite communication system is phase shift keying (PSK), as described in the following section. Bandpass (or radio frequency) transmission of digital data differs from baseband transmission only because modulation of an RF wave is required: the receiver demodulates the modulated RF wave to recover the baseband data stream. Thus ISI will occur at the receiver due to band limiting of the modulated waveform unless filters that satisfy the Nyquist criterion are used.

An additional constraint usually exists with radio communication systems. The bandwidth occupied by any radio transmission is specified to avoid interference with other transmissions at adjacent frequencies. The output of a transmitter must have a carefully controlled spectrum that reduces out-of-band signals to a low level. Figure 5.5 shows

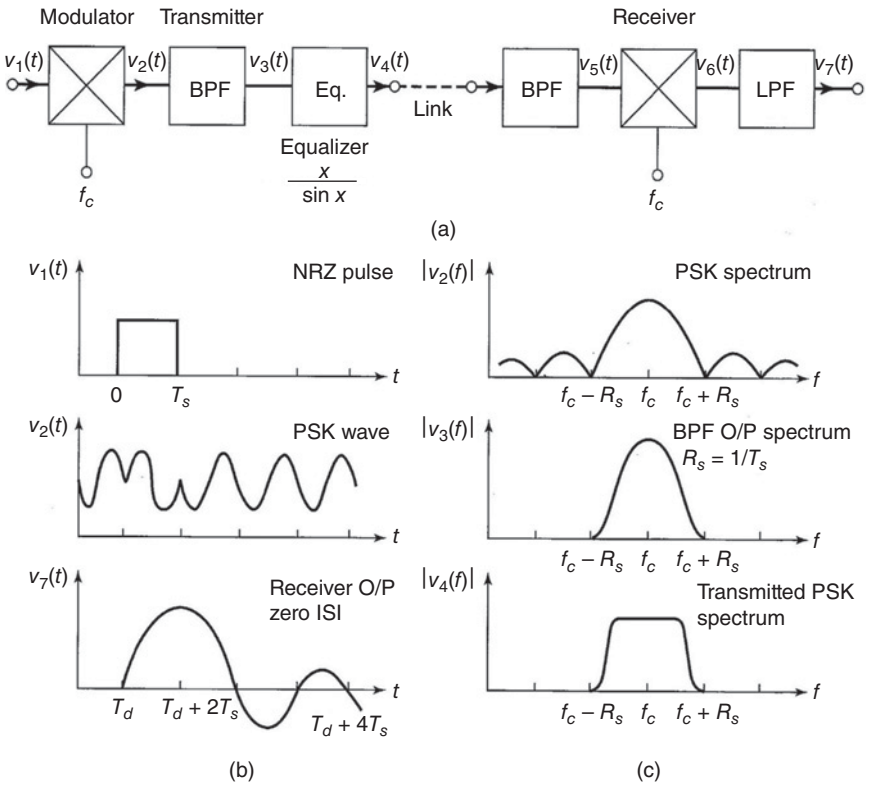


Figure 5.5 Waveforms and spectra in a BPSK radio link with SRRC filtering and NRZ pulse equalization. (a) Link block diagram. The band pass filters (BPFs) are SRRC filters implemented at the frequency f_c. The $x/\sin x$ equalizer can be before, after, or combined with the transmitter SRRC filter. (b) Waveforms of the NRZ input pulse $v_1(t)$, the BPSK waveform $v_2(t)$ before filtering and the receiver zero ISI waveform $v_7(t)$. (c) Spectra of the BPSK waveform $V_2(f)$ before SRRC filtering and after SRRC filtering $V_3(f)$, and the transmitted BPSK spectrum $V_4(f)$ after $x/\sin x$ equalization. The low pass filter after the demodulator removes any unwanted frequency components. T_s is the period of the input waveform and the bit rate is $R_s = 1/T_s$ bps.

the spectrum of a binary PSK (BPSK) signal generated from a random train of binary digits. The slow decay of the spectrum beyond $f_c \pm R_s$ Hz results from the sudden phase reversals of the PSK waveform.

The Nyquist filters used in a radio link must be bandpass filters, centered at the carrier frequency of the RF signal. The single sided transfer function of a bandpass SRRC filter is identical to the two sided baseband frequency response of the equivalent baseband filter, with its center frequency shifted from 0 Hz to the carrier frequency f_c Hz. Thus an important difference between baseband SRRC filters and bandpass (RF) SRRC filters is that the RF version has a bandwidth twice that of the baseband filter. In radio transmitters and receivers, SRRC bandpass filters are usually implemented in the *intermediate frequency* (IF) section of the transmitter or receiver, rather than the much higher radio frequency section. The reason for using IF filters is that the percentage bandwidth of an IF filter is much larger than that of an RF filter, making it much easier to build the filter. For example, a 10 MHz filter at an IF frequency of 100 MHz represents a 10% bandwidth.

The same filter at an RF frequency of 10 GHz has a 0.1% bandwidth and is much more difficult to construct.

At the transmitter, the SRRC filter limits the bandwidth of the transmitted baseband signal to $B_{occ} = R_s\ (1 + \alpha)$ Hz, where B_{occ} is called the *occupied* bandwidth of the transmitted signal, and R_s is the symbol rate. Some authors call this the *absolute bandwidth* of the signal (Couch 2007, p. 103). At the receiver, the SRRC filter limits the noise that can reach the receiver output to a noise bandwidth of $B_n = R_s$. Note the very important distinction here: the signal occupies a bandwidth B_{occ} Hz, but the noise bandwidth of every bandpass SRRC filter is R_s Hz. It is common in electronic circuits to think in terms of 3 dB bandwidth, but the analysis of radio communication systems requires knowledge of two different bandwidth – the total bandwidth required to carry the radio signal, and the noise bandwidth of the receiver. The total bandwidth, which is the occupied bandwidth, B_{occ} when Nyquist filters are used in the transmitter, defines the RF spectrum required to transmit all of the signal energy. It is often referred to as *channel bandwidth*, because we can think in terms of establishing a radio channel with bandwidth B_{occ}.

Hence for the case of a satellite communication system with a symbol rate R_s sps, the required SRRC filters are identical bandpass filters, which have the following bandwidths:

At the transmitter, the SRRC filter creates a signal with an occupied bandwidth

$$B_{occ} = R_s(1 + \alpha)\text{Hz} \tag{5.10}$$

At the receiver, the noise bandwidth of the SRRC filter is

$$B_n = R_s\text{Hz} \tag{5.11}$$

Every SRRC bandpass filter, regardless of the value of the roll-off factor α has a noise bandwidth equal to the symbol rate of the link. This is the most important design features of digital radio links, and must be observed whenever noise power is calculated in a digital radio receiver. Every SRRC bandpass filter in a digital radio transmitter generates a radio signal with a bandwidth $R_s(1 + \alpha)$ Hz, which is always greater than the numerical value of the symbol rate since $\alpha > 0$ in any real link. Satellite communication links typically use SRRC filters with α values between 0.2 and 0.35.

As stated above, the bandpass filters in the transmitter and receiver are usually implemented at an intermediate frequency f_{IF}. The frequency response of the bandpass SRRC filters therefore extends from a lower limit $f_{IF} - ½R_s\ (1 + \alpha)$ Hz to an upper limit $f_{IF} + ½R_s(1 + \alpha)$ Hz. The corresponding frequency range of the RF signal is the occupied bandwidth B_{occ}, centered on the RF carrier frequency f_c

$$B_{occ} = \left[f_c + \frac{R_s}{2}(1 + \alpha)\right] - \left[f_c - \frac{R_s}{2}(1 + \alpha)\right] \text{Hz} \tag{5.12}$$

Throughout the radio link, every RF and IF component must have a flat frequency response (and linear phase characteristics) over the bandwidth occupied by the signal. If any part of the link does not meet the frequency response requirement, an equalizer must be added in series with that part to force the required frequency response, otherwise ISI will be generated.

Example 5.1

A satellite link has an RF channel with a bandwidth 1.0 MHz. The transmitter and receiver have SRRC filters with $\alpha = 0.35$. What is correct symbol rate for this link?

The relationship between symbol rate and bandwidth is given by Eq. (5.10).

$$B_{\text{occ}} = R_{\text{s}}(1 + \alpha)\text{Hz}$$

$$10^6 = R_{\text{s}}(1 + 0.35) = 1.35R_{\text{s}}$$

$$R_{\text{s}} = \frac{10^6}{1.35} = 740.7\text{ksps}$$

Example 5.2

A Ku-band satellite uplink has a carrier frequency of 14.125 MHz and carries a symbol stream at $R_{\text{s}} = 16$ Msps. The transmitter and receiver have SRRC filters with $\alpha = 0.25$.

What is bandwidth occupied by the RF signal, and what is the frequency range of the transmitted RF signal?

From Eq. (5.10)

$$B_{\text{occ}} = R_{\text{s}}(1 + 0.25) = 1.25R_{\text{s}}\text{Hz}$$

$$= 1.25 \times 16 \times 10^6 = 20\,\text{MHz}$$

The RF signal occupies the frequency range given by Eq. (5.12)

$$f_{\text{c}} - \frac{R_{\text{s}}}{2}(1 + \alpha) \leq f_{\text{RF}} \leq f_{\text{c}} + \frac{R_{\text{s}}}{2}(1 + \alpha)$$

Hence, the frequency band occupied by the uplink signal extends from

$$(14.125 - 0.01) \text{ to } (14.125 + 0.01) = 14.115 \text{ GHz to } 14.135 \text{ GHz}$$

In many data transmission systems the baseband waveform at the transmitter input has an NRZ format. A link with a Nyquist transfer function will produce a zero ISI waveform at the receiver output only when driven by an impulse, as shown by Eqs. 5.6 and 5.7. If the transmitter is driven by a NRZ waveform with a symbol rate R_{s}, period $T_{\text{s}} = 1/R_{\text{s}}$, the spectrum of the driving pulse has a shape

$$S(f) = \frac{\sin(\pi T_{\text{s}} f)}{\pi T_{\text{s}} f} \tag{5.13}$$

and the spectrum of the output of the SRRC filter will be

$$V_{\text{t}}(f) = \sqrt{V_{\text{NQ}}(f)} \left[\frac{\sin(\pi T_{\text{s}} f)}{\pi T_{\text{s}} f}\right] \tag{5.14}$$

To obtain zero ISI at the receiver, we must force the spectrum of the signal from the transmitter to be $\sqrt{V_{\text{NQ}}(f)}$, which can be achieved by using an equalizer in the transmitter with a transfer function given by

$$E_{\text{t}}(f) = \frac{\pi T_{\text{s}} f}{\sin(\pi T_{\text{s}} f)}$$

Then the output spectrum from the transmitter is given by

$$V_t(f) = S(f) \times E_t(f) \times H_t(f)$$

$$= \left[\frac{\sin(\pi T_s f)}{\pi T_s f}\right] \times \left[\frac{\pi T_s f}{\sin(\pi T_s f)}\right] \times \sqrt{V_{nq}(f)} = \sqrt{V_{nq}(f)} \tag{5.15}$$

The arrangement is illustrated in Figure 5.5a. The SRRC filters have zero transmission beyond frequencies defined by $f = f_c \pm f_{max}$ Hz, where $f_{max} = \frac{R_s}{2} \times (1 + \alpha)$Hz. The $x/\sin x$ equalizer operates only within the central lobe of the sin x/x function. At $f = 1/T_s$ Hz the $x/\sin x$ function goes to infinity, so the SRRC filter parameter α must be less than 1 for this system to work. In practice, RF filters with SRRC shaping use α values between 0.1 and 0.5. The maximum gain of the equalizer for $\alpha = 0.5$ is given by

$$G = 20\log_{10}\left(\frac{0.75\pi}{\sin(0.75\pi)}\right) = 10.5\text{dB}$$

at the edge of the equalizer band. Transfer functions are given in voltage terms, hence we must use 20 log to obtain the gain of the equalizer in decibels. Reminder: Power is proportional to (voltage^2) so 10 $\log_{10}$ power = 10 $\log_{10}$ (voltage)^2 = 20 $\log_{10}$ (voltage.)

The bandwidth of the link must at least equal the bandwidth occupied by the signal, otherwise the spectrum of the received signal will be altered by transmission over the link and we will not have a zero ISI raised cosine waveform at the output of the receiver. Thus the transfer function of the link, $L(f)$, must be such that $|L(f)| = 1$ and $\Phi(f) = kf$, where $\Phi(f)$ is the phase characteristic of the link with frequency, and k is a constant, over the bandwidth occupied by the signal, R_s $(1 + \alpha)$ Hz. If this condition is not met, either in magnitude or phase, we can insert an equalizer in series with the link to force the required condition. The equalizer can be at the transmitter or the receiver. In systems where the characteristics of the link vary with time, as in a mobile link that suffers from multipath interference, for example, the link equalizer can be adaptive. Multipath is less of a problem for mobile satellite links than for cellular telephone systems, where adaptive equalizers are commonly used, because the satellites are used only when their elevation angle is at least five degrees.

Adaptive equalizers are used in cellular radio links because the frequency characteristics of the radio channel change as the transmitter or receiver moves around. The equalizer is implemented as a *transversal equalizer*, which operates in the time domain rather than the frequency domain (Rappaport 2002, p. 359). A transversal equalizer works to improve the pulse shape at the output of the receiver so that the pulses more closely resemble the ideal zero ISI shape. The received pulses are sampled repeatedly within the pulse period, and then weighted and delayed samples are added to the original pulse to improve its shape. A *training sequence* is required in the signal to allow the transversal equalizer to adjust to the received signal. In a mobile radio system the equalizer must continuously adapt to the changing propagation path. Transversal equalizers are not generally used in fixed satellite communication links because the path between the satellite and earth is stable. They may be used in mobile links, however, because multipath and blocking conditions can exist, creating a less stable propagation path.

For more detail on transversal equalizers, the reader should refer to a text on communications system theory (Ziemer and Tranter 2015).

The discussion of filter characteristics and signal spectra thus far has ignored the phase response of the filters and the resulting phase spectra of the waveforms. It can readily be shown that the phase response of all filters and equalizers must be linear with frequency for the zero ISI condition to be met (Couch 2007, pp. 83–84) Achieving a linear phase response throughout a communication system can be difficult in practice, and may require the use of *phase equalizers*.

5.1.4 Transmission of QPSK Signals Through a Band-Limited Channel

Figure 5.5 shows a block diagram of a typical BPSK link. Modulation of the digital signal is achieved by multiplying an IF carrier by the NRZ data stream. When the modulating signal is a logical 1 the carrier wave is multiplied by +1 and remains unchanged with a reference phase angle of zero degrees. When the modulating signal is a zero, the carrier is multiplied by −1 causing a 180° phase shift. The modulated wave is a *phase reversal keyed* (PRK) waveform, more generally known as BPSK. The phase reversals in the BPSK waveform generate a wide spectrum, which is limited to the channel bandwidth by the SRRC filter that follows the modulator.

The same transmitter and receiver system, and the same satellite link, can be used to send QPSK signals, since QPSK is nothing more than two BPSK signals generated from carriers that are in phase quadrature, which can share a common link bandwidth. (Any pair of signals that is orthogonal can be separated by suitable signal processing in a radio receiver. Signals in frequency bands that do not overlap are orthogonal, and signals in phase quadrature are orthogonal. Code division multiple access (CDMA) signals using ideal codes are also orthogonal.) The input equalizer with a $x/sin\ x$ transfer function $E_i(f)$ is placed after the PSK modulator because typically PSK modulators are binary state devices that need to be driven by NRZ digital signals with voltage levels $\pm V$ volts. The input to the PSK modulator must therefore be a $\pm V_{signal}$, with $+V =$ data 1 and $-V =$ data *0* (or vice-versa), created from the logic levels of the NRZ waveform that carries data bits (usually +5 V and 0 V, or +3.3 V and 0 V).

The waveform at the output of the receiver, $v_r(t)$ must be a zero ISI waveform. This can be achieved only if we can make the transfer function of the entire link, from input terminal to output terminal, equal to $V_{NQ}(f)$, the Nyquist raised cosine transfer function. The condition for zero ISI in the link is therefore

$$V_i(f) \times H_t(f) \times E_i(f) \times L(f) \times H_r(f) \times G_{LPF}(f) = V_{NQ}(f) \tag{5.16}$$

where

$V_i(f)$ = sin x/x, the spectrum of the NRZ data pulses at the system input
$H_t(f)$ = $H_r(f)$, the transfer function of the bandpass SRRC filters in the transmitter and receiver
$E_i(f)$ = $x/\sin x$, the equalizer for the input NRZ waveform
$L(f)$ = the transfer function of the link, equalized if necessary to make it linear
$G_{LPF}(f)$ = the transfer function of the baseband low pass filter following the demodulator.

In order to satisfy the conditions for Eq. (5.16) to be true within the limits of the bandwidth occupied by the signals in the link, $R_s\ (1 + \alpha)$, we must ensure that

$$V_i(f) \times E_i(f) = 1$$
$$L(f) = 1$$
$$G_{LPF}(f) = 1$$

then

$$H_t(f) \times H_r(f) = V_{NQ}(f) \tag{5.17}$$

QPSK is a popular choice of modulation technique for use in satellite communication links carrying digital data. Modulation methods for satellite links are described in more detail in Section 6.1 of Chapter 6. In QPSK modulation a digital data stream is taken two bits at a time and used to generate one of four possible phase states of the transmitted carrier. This requires an in-phase (I) channel and a Q (quadrature) channel. The four QPSK states are generated by adding the signals from the I and Q channels according to a *mapping function*. (See Section 6.1 of Chapter 6 for details of PSK modulations.) If the data rate is R_b, bits/second, the symbol rate for the QPSK carrier is $R_s = R_b/2$ symbols/second.

In order to recover the symbol stream with zero ISI, we must shape the transmitted spectrum such that after demodulation a single symbol creates a zero ISI waveform at the output of the demodulator. Then sampling of the symbol stream can be achieved with zero ISI. In practice, a QPSK receiver has two demodulators, one for the I channel and one for the Q channel. The outputs of the demodulators are added to give one of four signal states. We shall consider only one channel in looking at ISI.

Figure 5.6 shows a typical arrangement for one half of a QPSK transmit-receive link. The other half is identical except that the carrier used for modulation and demodulation is shifted in phase by 90°. Since the carriers in the two channels have a 90° phase difference, the channels are identified as *in-phase* (I) and *quadrature* (Q). The data presented to the QPSK modulators is in NRZ format and causes a jump in carrier phase at each symbol transition, as seen in Figure 5.5b. The input data rate to the demodulator is $R_s = R_b/2$ sps, giving the QPSK spectrum shown in Figure 5.6b.

The central lobe of the unfiltered QPSK spectrum extends from $(f_c - R_s)$ to $(f_c + R_s)$, giving a band occupancy of $2R_s$. The spectrum must be narrowed for transmission via a radio channel, and this is achieved by use of a bandpass filter meeting the zero ISI criterion, for example, a SRRC filter.

The bandpass raised cosine function is a transformation of the low-pass raised cosine function and has a response $|H(f)| = 1/2$ at frequencies $(f_c - \frac{R_s}{2})$Hz and $(f_c + \frac{R_s}{2})$Hz. The frequencies at which $H(f)$ falls to zero are determined by the roll-off factor α in Eq. (5.10). Matched filter operation of the link requires that the raised cosine function response be split between the transmit end and the receive end of the link. Thus a square root raised cosine response filter is required after the modulator in the transmitter and before the demodulator in the receiver. Finally, because we are using NRZ pulses rather than impulses, we need an $x/\sin x$ equalizer to equalize the spectrum from the modulator.

Example 5.3

A satellite transponder has a bandwidth of 36 MHz. Earth stations use ideal SRRC filters with $\alpha = 0.4$. What is the maximum bit rate that can be sent through this transponder with

i) BPSK
ii) QPSK?

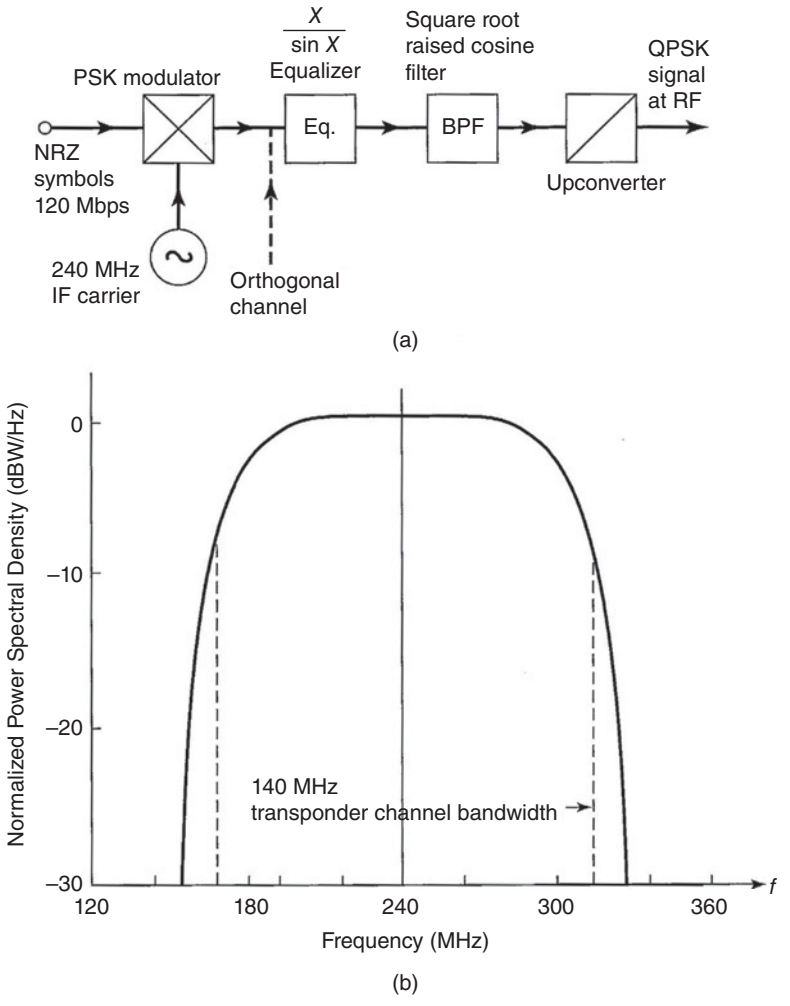


Figure 5.6 QPSK transmission for a bit stream at 240 Mbps and symbol rate 120 Msps. Only the in-phase portion of the transmitter is shown. An identical modulator driven by a quadrature carrier supplies the orthogonal channel signal. IF frequency is 240 MHz. (a) The 240 Mbps data stream is split into alternate bits to create two bit streams at 120 Mbps. Both signals are applied to the same SRRC filter and *x*/sin*x* equalizer, then upconverted to the transmitted radio frequency. (b) Spectrum of the filtered and equalized IF signal. Transmission through a transponder with a bandwidth of 140 MHz will cause some ISI in the receiver.

The maximum symbol rate for an RF link is given by Eq. (5.10), reordered as

$$R_s = B/(1+\alpha) = 36\,\text{M}/1.4 = 25.7\,\text{Msps}$$

The corresponding bit rates for BPSK and QPSK are

i) BPSK $R_b = R_s = 25.7$ Mbps
ii) QPSK $R_b = 2 \times R_s = 51.4$ Mbps

Example 5.4

A data stream at 240 Mbps is to be sent via a satellite using QPSK. The receiver IF frequency is 240 MHz. Find the RF bandwidth needed to transmit the QPSK signal when ideal SRRC filters with $\alpha = 0.35$ are used.

The symbol rate for the link is $R_b/2 = 120$ Msps.

Using Eq. (5.10)

$$B = R_s(1 + \alpha) = 162\,\text{MHz}$$

The 240 Mbps signal is divided into two 120 Mbps symbol streams and applied to I and Q channel modulators fed by two IF carriers, with a 90° phase difference. The resulting spectrum from each modulator has a width of 240 MHz between zeros of the central lobe of the PSK spectrum. The I and Q signals are added and applied to an x/sin x equalizer with $x = [f_c - \pi f R_s]$ extending to ±87 MHz from the carrier. The maximum gain of the function x/sin x at ±87 MHz from the carrier is 9.5 dB. Figure 5.6a shows a block diagram of one half of the transmit end of the QPSK link.

The equalized spectrum is applied to the square root raised cosine filter. The response of this filter is 3 dB down at $f_c \pm R_s/2$ Hz, that is, at $f_c \pm 60$ MHz. In practice, a single filter combining the square root raised cosine and x /sin x responses is used. A first order Chebychev filter is a good approximation to the combination of an SRRC filter and an x/sin x equalizer. The ideal combined response of this single filter is shown in Figure 5.7c. Thus in the IF amplifier of the receiver, the signal spectrum is 6 dB down at 180 and 300 MHz. The bandpass filter will theoretically have zero response for $f < f_c - ½R_s(1 + \alpha)$ Hz and $f > f_c + ½R_s(1 + \alpha)$ Hz, that is, below 153 MHz and above 327 MHz. Figure 5.7b shows the transmitted QPSK spectrum centered on the IF carrier.

If we examine the spectrum of the QPSK signal at the receiver, we find that the 3 dB bandwidth is 120 MHz and the total frequency band containing all of the signal energy is 174 MHz. A typical satellite transponder for such a signal would have a 3 dB bandwidth of 160 MHz. Beyond 160 MHz the spectrum of the QPSK signal would be attenuated by the transponder filter, leading to some spectral distortion of the receiver signal and consequent ISI in the demodulated waveform. However, the energy contained in the QPSK spectrum beyond ±70 MHz from the carrier is small, and the ISI caused by the transponder filter is minimal.

Practical filters invariably cause some ISI because it is impossible to realize the square root raised cosine characteristic exactly. Typically, with a high speed digital link operating at 120 Msps, up to 2 dB of additional carrier power may have to be provided to achieve a 10^{-8} BER, compared to the theoretical power needed for this error rate in a carefully filtered QPSK link. The extra power is called an *implementation margin*. Implementation margin accounts for the non-ideal nature of real communication systems. Timing jitter leading to incorrect sampling instants, phase instability in the receiver local oscillator, real filters rather than Nyquist SRRC filters, and phase and amplitude distortions all contribute to ISI levels that are higher than theory predicts. An implementation margin covers all of these effects. In links operating at lower symbol rates and with higher CNRs, implementation margins of 0.2–0.5 dB are common. Links that have low CNR and high bit rates, as in DBS-TV, implementation margin may be 1.5–2.0 dB. Chapter 10 discusses these aspects of DBS-TV in more detail.

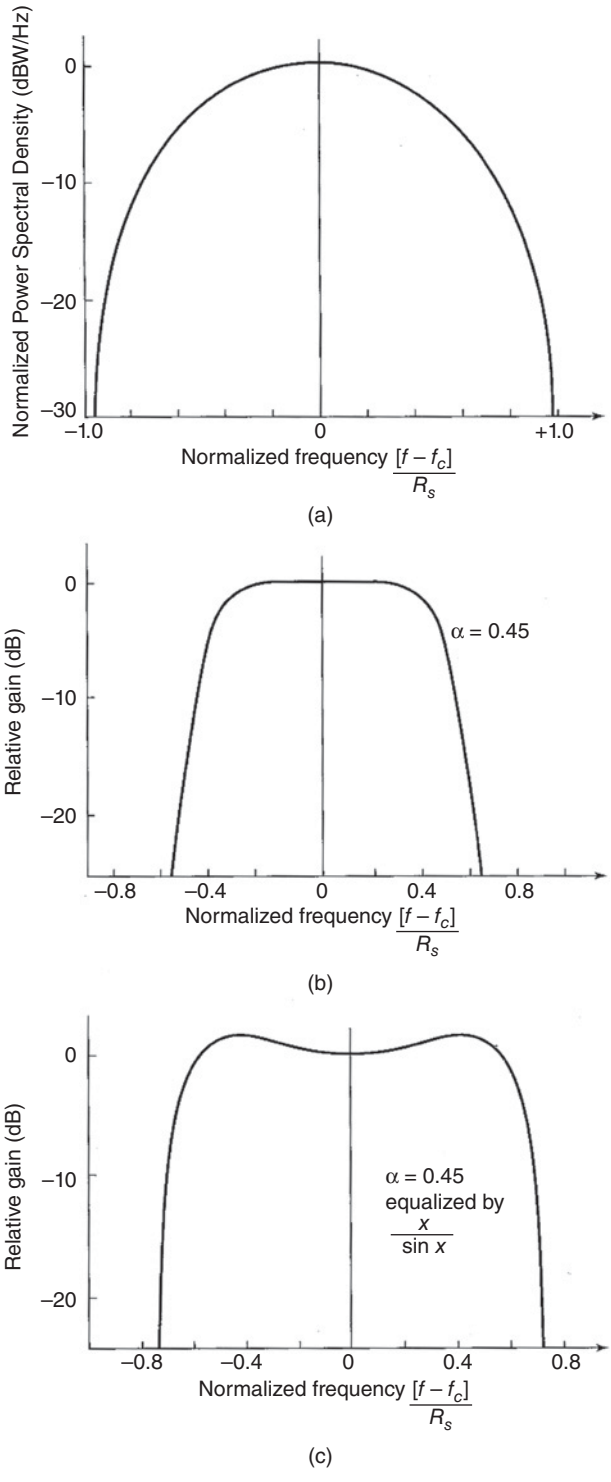


Figure 5.7 (a) Frequency spectrum of an unfiltered QPSK signal with carrier frequency f_c and symbol rate R_s. Only the central lobe of the spectrum shown. (b) Transfer function of a SRRC filter with $\alpha = 0.45$. (c) Spectrum of the QPSK signal after equalization by an $x/\sin x$ equalizer. Note that the frequency scale has been normalized to be centered at the actual carrier frequency f_c Hz.

5.2 Implementing Zero ISI Transmission in the Time Domain

The inclusion of microprocessors and application specific integrated circuits (ASICs) into radio communication devices, namely transmitters and receivers, revolutionized the way in which they are implemented, but also introduced a high level of complexity. In Section 5.1 the classical approach to zero ISI transmission is described in the frequency domain, where signals are represented by their spectra and are processed by filters with closely specified transfer functions. The ideal square root raised cosine filter is the controlling element in both the transmitter and the receiver. The transmitter generates a SRRC waveform at an intermediate frequency using a filter that approximates the SRRC transfer function as closely as possible. In the receiver an identical IF filter results in the baseband signal having a good approximation to a raised cosine waveform with zero ISI. The ideal SRRC filter cannot be created in practice because it cannot have zero transmission above its cut off frequency, so there is always some ISI present in the receiver output waveform when real filters are used, contributing to the implementation margin of the system.

Filters tend to be bulky devices, made up of inductors and capacitors. Inductors, in particular are much larger than most other components in a transmitter or receiver and cannot easily be included in integrated circuits. This is an important consideration in small devices such as cellular phones, where the size of the handheld device is an important consideration. As cellular telephones developed their size steadily reduced, from the early days of *bag phones* the size of a house brick to a thin package that can fit into a pocket. The reduction in size required the elimination of virtually all the inductors that were used in the earlier generations of cellphones, and has been achieved by the use of digital filters. A digital filter is a device operating in the time domain that has a transfer function that matches the classical hardware inductor and capacitor filter. Digital filters are implemented in the time domain using infinite impulse response (IIR) or finite impulse response (FIR) designs. IIR filters use feedback loops to generate the required impulse response, while FIR filters use only feed forward loops. Replacing hardware filters and other operations such as modulation and demodulation by digital time domain equivalents allows much of a radio to be implemented in a microprocessor or specialized digital signal processing (DSP) integrated circuits or field programmable gate arrays (FPGAs).

5.2.1 Generation and Transmission of Square Root Raised Cosine Waveforms

An alternative to the use of SRRC filters in the transmitter of a digital link is to generate the waveform corresponding to the output of the SRRC filter – a SRRC waveform. This can be done at baseband, before the modulator in a digital radio link. Since the waveform at the output of an ideal SRRC filter driven by an impulse extends from $-\infty$ to $+\infty$ in time, we must approximate the waveform by truncating it to a finite time interval.

As an example, consider the case of a system designed with ideal SRRC filters with $\alpha = 0$, allowing the use of the minimum bandwidth of R_s Hz in the radio link. The corresponding time waveform at the receiver has a sin x/x shape (a sinc pulse) as seen in Figure 5.3b. If we truncate the transmitted waveform between $-3T_s$ and $+3T_s$ we will eliminate all the energy outside this time interval. The power associated with the missing part of the transmitted waveform can be calculated by integrating the SRRC waveform

Table 5.1 Energy contained within a truncated SRRC waveform for various values of alpha and waveform length

	Percent energy contained within sequence length			
Alpha	**2T**	**4T**	**6T**	**8T**
0	92.23	96.79	98.38	99.19
0.1	93.05	97.47	98.89	99.52
0.2	94.64	98.62	99.61	99.90
0.3	96.04	99.38	99.91	99.98
0.4	97.15	99.75	99.96	99.98
0.5	98.02	99.90	99.97	99.99
1.0	99.79	99.98	99.99	100.00

from $-3T_s$ and $+3T_s$ and comparing the result to the integral from $-\infty$ to $+\infty$. The sinc function has no analytical integral, so the integration must be performed numerically. The result for the six symbol period shows 1.62% of the pulse energy is omitted when the truncated pulse is used in place of the infinitely long pulse. Some intersymbol interference will result because part of the energy in the SRRC waveform has been lost, but the impact on link performance is unlikely to be any greater than when non-ideal SRRC filters are used. However, truncating the waveform will result in a widening of its spectrum, and all of the transmitted energy will no longer lie within the specified bandwidth. Additional filtering is required in the transmitter to keep the spectrum within the allocated channel bandwidth, and additional ISI is generated. With an SRRC waveform corresponding to an SRRC filter with a higher value of α, the missing power is reduced and less ISI will result.

The SRRC waveform is given by Haykin and Moher in slightly different form (Haykin and Moher 2005, p. 121)

$$p(t) = \frac{\sin\left[\pi \frac{t}{T}(1-\alpha)\right] + 4\alpha \frac{t}{T}\cos\left[\pi \frac{t}{T}(1+\alpha)\right]}{\pi \frac{t}{T}\left[1 - \left(4\alpha\pi \frac{t}{T}\right)^2\right]} \tag{5.18}$$

where T is the symbol period and α is the roll off factor of the SRRC filter. Table 5.1 shows the percentage of energy included in a truncated SRRC waveform for various values of α and truncation length, calculated by numerical integration of Eq. (5.18). When α is small, a longer waveform is needed to reduce the lost energy to less than 1%. With $\alpha = 0$ this requires a waveform of eight periods.

With $\alpha = 0.2$, an SRRC waveform of six periods reduces the energy loss to 0.39%. Close inspection of the truncated waveforms shows that the zero crossings are no longer at regular intervals of T. However, when the SRRC waveform is passed through the SRRC filter in the receiver, the zeroes of the wave appear at regular intervals of T avoiding any additional ISI in the receiver. Determining the impact of ISI caused by the use of truncated SRRC waveforms requires the use of a communication simulator such as Simulink® (Simulink® 2019).

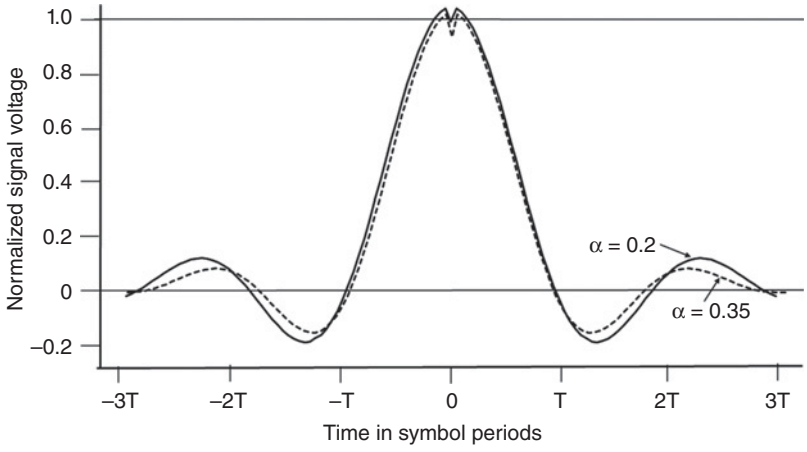


Figure 5.8 SRRC waveforms with α = 0.2 and 0.35 generated with six period truncation. *T* is the bit period. Alpha is the roll off factor for RRC waveforms.

Figure 5.8 shows examples of truncated SSRC waveforms for $\alpha = 0.2$ and 0.35 truncated to six periods. The waveforms were generated in Microsoft Excel® using 20 samples per bit period, corresponding to the way the waveform is generated digitally in a software transmitter.

The transmitter SRRC waveforms for three bit sequences with $\alpha = 0.2$ and a waveform length of six periods are shown in Figure 5.9a–c. The six period truncated waveform with $\alpha = 0.2$ shown in Figure 5.8 was used to create a 14 bit sequence by adding positive and negative SRRC waveforms for data inputs of 1 and 0 offset by successive bit periods of T seconds. Only the central eight bits of the waveform are shown in Figure 5.9a–c to eliminate starting and ending transients.

In Figure 5.9a, the digital bit sequence is 101010101010 resulting in a waveform with values ±1 V at the sampling intervals. In Figure 5.9b, the bit sequence is 1010101101010 resulting in sample values that are not exactly 1.0 V, and some transients are evident in the waveform. In Figure 5.9c the bit sequence is 01010111010101. The transients are the result of truncation of the SRRC waveform to six periods and the use of 20 samples per period to generate the waveform digitally. The SRRC waveform generated this way

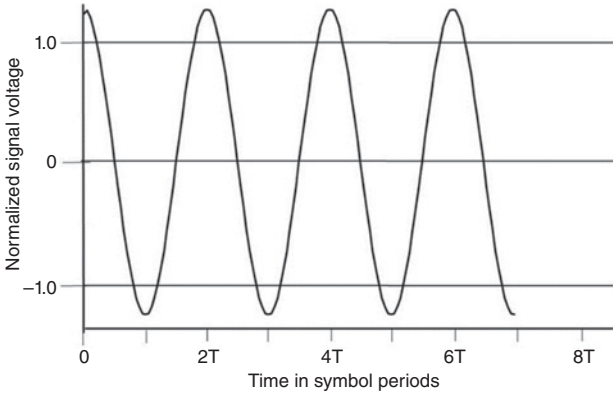


Figure 5.9a SRRC waveform for continuous 10101010 bit pattern.

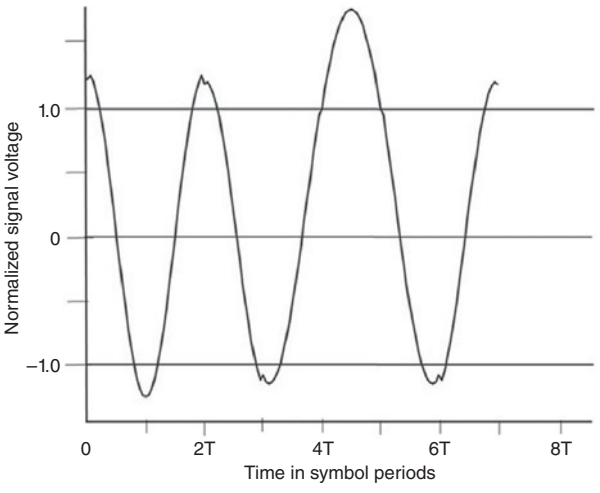


Figure 5.9b SRRC waveform for 101011101 bit pattern.

requires filtering to control the spectral spreading caused by these transients. Amplitude modulation can be seen in these waveforms; they no longer have constant magnitude. This causes ISI when the signal is passed through a non-linear satellite transponder because the peaks of the waveform will be compressed.

Linearization of the transponder, discussed in Chapter 10, is desirable to reduce the ISI.

Any waveform can be generated digitally using a read only memory (ROM). Digital values representing the amplitude of the waveform are stored in the ROM. In the example case above for $\alpha = 0.2$, sample values of the positive and negative SRRC waveforms over a six symbol period are stored, with typically 10 or 20 samples per symbol period. When a digital 1 is to be sent, the numerical values for a positive waveform are read out of the ROM at a rate 10 or 20 times higher than R_s and sent to an adder. When a digital 0 is to be sent, the numerical values for a negative waveform are read out of the ROM and sent to the adder. Successive pules are generated with offsets of T_s, $2T_s$, $3T_s$ … so that the output of the adder is the sum of each six symbol length waveform. The output of the

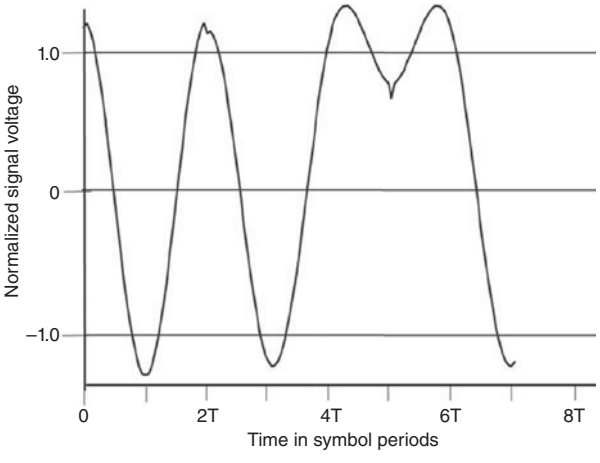


Figure 5.9c SRRC waveform for 10101110 bit pattern.

adder is sent to a digital to analog converter (DAC) followed by a smoothing filter. In a hardware radio transmitter the analog SRRC waveform is applied to the input of a linear modulator, typically a phase modulator, and a near-zero ISI PSK transmitted waveform results. Note that with a six period truncated SRRC waveform there is a three symbol delay between the decision to send a logical 1 or 0 and the SRRC waveform reaching its maximum value. In a software radio a linear PSK modulator can be implemented by using the SRRC waveform voltage to select a sample of a digitally generated IF waveform that is advanced or retarded in time in proportion to the SRRC waveform voltage. The resulting steps in the PSK waveform are smoothed out by a bandpass filter following the modulator.

5.2.2 Software Radios and Digital Filtering

In a digital radio, filters are implemented in the time domain using IIR or FIR designs. For links that require good stability and linear phase response, the FIR filter is preferred. In a digital radio, FIR filters can be implemented in ASICs and FPGAs (Viasat 2017). In a software radio, a microprocessor programmed to execute each operational block in the radio is implemented in the time domain as a series of instructions. The advantage of a software radio is that the programming can be changed, via a satellite if necessary, giving a degree of flexibility not available in hardware designs. Development of digital and software radios has been spurred by the requirement of pocket-sized cellular telephones, leading to the availability of many of the needed component blocks of transmitters and receivers in ASIC form, and very powerful microprocessors that can be used to create software radios. Figure 5.10 illustrates the structure of a digital radio transmitter for a single data stream with *adaptive coding and modulation* (ACM).

Figure 5.11a shows the structure of a digital receiver that uses sampling of the first IF signal. The RF signal (f_{rf}) is down converted to a suitable intermediate frequency (f_{if}) and sampled by two analog to digital converters (ADCs) that are driven in phase quadrature at a sampling rate f_s, creating I and Q channels that preserve the phase information in the signal. The sampled signals are applied to SRRC filters implemented as digital FIR filters and then digitally demodulated. The receiver can adapt to the modulation of the signal

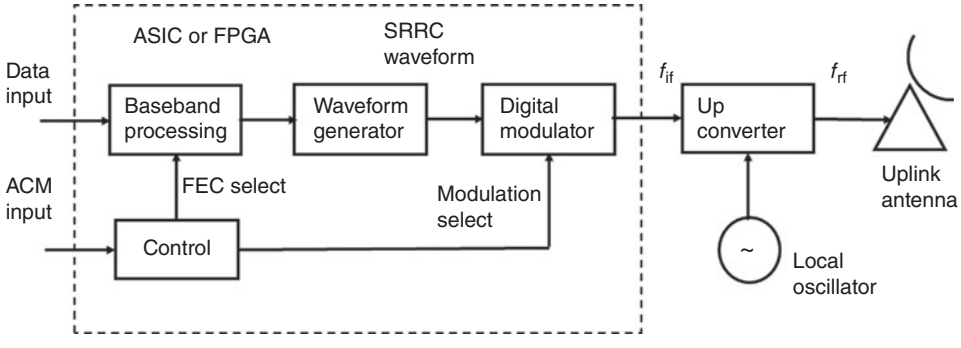


Figure 5.10 Digital transmitter structure for single data stream with adaptive FEC coding and modulation. The blocks within the dotted line are implemented in an application specific integrated circuit (ASIC) or a field programmable gate array (FPGA). The waveform generator creates SRRC waveforms at the IF frequency f_{if}. The upconverter translates the signals to the RF frequency f_{rf}. A microprocessor provides the ACM control. ACM: Adaptive coding and modulation.

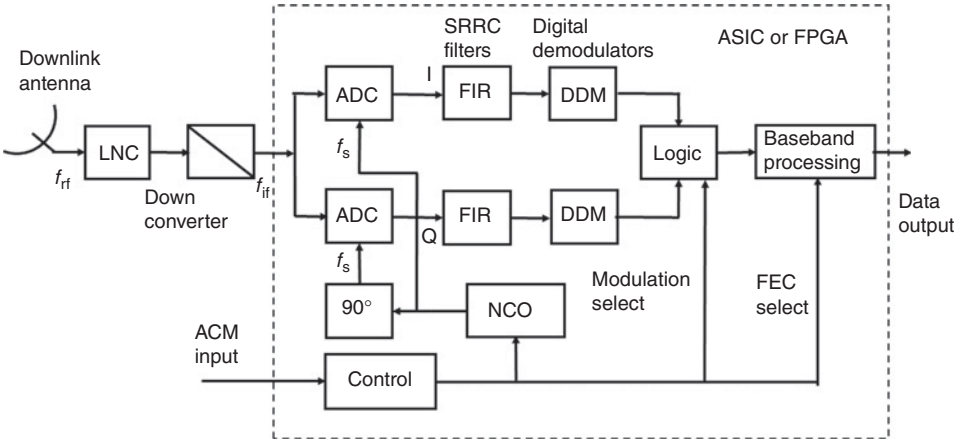


Figure 5.11a Structure of a digital receiver that uses sampling of a second IF signal. The First IF signal from the LNC is down converted to a suitable intermediate frequency (f_{rf}) and sampled by two analog to digital converters (ADCs) driven in phase quadrature at a sampling rate f_s, creating I and Q channels. SRRC filters are implemented as digital FIR filters at the IF frequency. The I and Q signals are demodulated digitally. By changing the logic that combines the baseband I and Q signals after the demodulators, the receiver can adapt to different forms of phase shift keying. The receiver can also adapt to changes in the FEC coding. LNC: Low noise block down converter, in this case just a low noise amplifier (LNA). ADC: analog to digital converter. FIR: Finite impulse response digital filter. DDM: Digital demodulator. NCO: numerically controlled oscillator. ACM: Adaptive coding and modulation.

by changing the logic that combines the output of the demodulators, and also adapt to changes in the forward error correction (FEC) coding. A microprocessor provides the ACM control.

Figure 5.11b shows the structure of a digital receiver that uses direct conversion of the RF signal to baseband. The RF signal (f_{rf}) is sampled by two ADCs that are driven in phase quadrature at a sampling rate f_{rf}, creating baseband I and Q channels. The SRRC

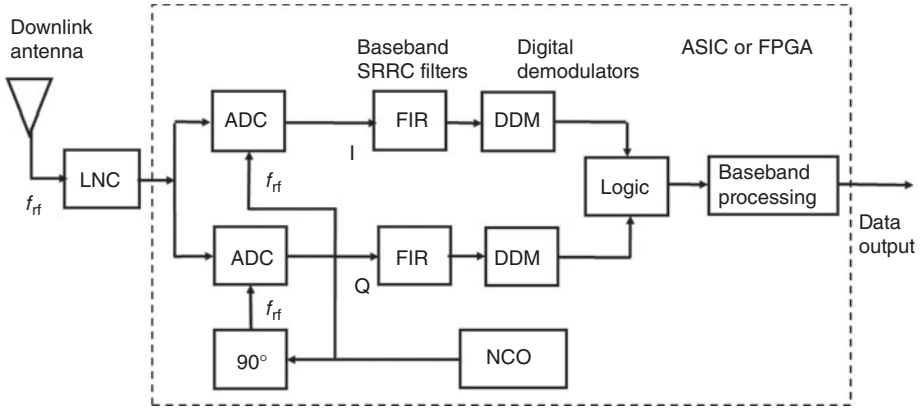


Figure 5.11b Structure of a digital receiver using direct conversion of the RF signal to baseband. The SRRC filters are implemented as low pass digital FIR filters. ADC: analog to digital converter. FIR: Finite impulse response digital filter. DDM: Digital demodulator. NCO: numerically controlled oscillator.

filters are implemented as baseband digital FIR filters and the signal is demodulated digitally.

The process of convolving an analog waveform within an FIR filter requires that the waveform be digitized at a rate much higher than its carrier frequency, so software radios for satellite communications typically use an RF front end that down converts the RF signal to an IF frequency suitable for digitizing with a fast ADC. DSP devices are available (2017) that can accept inputs at frequencies up to 2 GHz.

FIR filters have a limited number of stages that correspond to the truncation of the SRRC waveform, so the transfer function does not precisely match that of an ideal SRRC filter needed in the transmitter and receiver. Some ISI is therefore generated whenever digital filters are employed, just as occurs with hardware filters. Implementing all the functions of a transmitter and receiver digitally is especially important in handheld devices where size is of great importance. Modern cell phones and global positioning system (GPS) receivers, for example, cannot contain large components like analog inductors and rely entirely on digital implementation.

The output of the digital filter must be converted back to an analog waveform with a DAC when an analog waveform is required. The process of digitizing an analog waveform with an ADC and then recovering the filtered waveform with a DAC introduces the same problems of quantization error and aliasing that exist with digital voice links. FIR filters and the ADC–DAC combination introduce a delay (*latency*) into the communication link, which can have undesirable effects. The FIR filter design, its operating frequency, and the number of bits in the ADC and DAC must be chosen as a compromise between accuracy and latency.

The details of FIR and IIR filtering and the design of software radios are beyond the scope of this text. There are many texts devoted to the design and implementation of digital filters, and also on the topic of software radios. (See, for example, Taylor 2011; Antoniou 2018; Reed 2002.)

5.3 Probability of Error in Digital Transmission

A received radio signal will always be accompanied by noise and may also be subjected to interference from other radio sources. Unless a very high CNR can be guaranteed in the receiver, a digital radio will occasionally suffer bit errors at its output. Almost all digital radio links implement some form of error control. Satellite links are typically affected by rain in the earth-space path, resulting in attenuation of the signal and lowered CNR. At higher radio frequencies, 10 GHz and above where the majority of satellite links operate, rain attenuation events are inevitable and will cause significant bit errors at the output of the receiver and occasional outages. Error control measures aim to minimize the effect of these errors on the end user. We will first examine error rates, more accurately described as the probability of bit errors, for different modulation techniques as a function of the receiver CNR. Then we will consider how to mitigate the errors, either by FEC, which reduces the rate at which errors reach the end user, or by strategies that allow the user to decide what to do when errors are detected.

The noise power in a radio receiver is calculated using the techniques discussed in Chapter 4, which provide a value of CNR at the input to the demodulator. The CNR expresses noise as a power ratio relative to the received carrier. The noise in the receiver is assumed to be *additive white Gaussian noise* (AWGN), leading to a description of the

satellite link as an *AWGN channel*. The assumption that the noise is additive, white, and has a Gaussian voltage distribution is necessary to simplify the analysis of error probability. In satellite links, the assumption is usually valid provided the noise is thermal in origin (e.g., from the receiver front end and satellite transponder). If the noise is actually interference from another communication link, the assumption may not be valid, but often AWGN conditions are assumed for want of a better way to analyze interference. A communication link simulator that emulates the time waveforms of the actual link can be used to analyze the effect of an interfering signal, for example, Simulink (Simulink® 2019).

5.3.1 Probability of Symbol Error

For the analysis of error probabilities, we need to work with voltages. The noise voltage at the output of the demodulator is given by $n_o(t)$. At the sample instant, we will assume a noise voltage n_o volts at the demodulator output. The decision circuit will make an error if noise changes the sign of the received signal $v(t)$ at the sample instant. This is illustrated in Figure 5.12, where the received signal has sample voltages of $+V$ and $-V$ volts. The threshold for the decision circuit is at zero volts. If the sample is less than zero volts (i.e., a negative voltage) the decision circuit will output a data zero. If the sample voltage is more than zero volts (i.e., a positive voltage) the decision circuit will output a data one.

At the sampling instant, the output of the demodulator v_o will be the sum of the signal sample $\pm V$ and the noise sample n_o.

$$v_o = \pm V + n_o \text{ volts} \tag{5.19}$$

There are two possible ways that an error can occur, depending whether a $+V$ or a $-V$ signal was transmitted. If a signal $+V$ was sent and $n_o < -V$ (i.e., the magnitude of the noise sample is negative and larger than V) the sum of the signal and noise will be less than zero, giving a symbol error. If the transmitted signal was $-V$ and noise at the sampling instant was greater than $+V$, an error will occur because the sum is greater than zero volts. We will assume that, on average, the transmitter always sends an equal number of logical ones and zeroes.

We can calculate the probability that an error will happen, and thus the symbol error rate P_e, by the following argument. At the correct decision time, the amplitude of the signal will be $\pm V$, where V is the peak magnitude of the waveform at the output of the demodulator. The Gaussian distribution is symmetrical about zero volts, so the probability of an error occurring is

$$\begin{aligned} P_e &= \tfrac{1}{2} P(\text{output is} > 0 \text{ when} - V \text{ was sent}) + \tfrac{1}{2} P(\text{output is} < 0 \text{ when} + V \text{ was sent}) \\ &= P(\text{output is} > 0 \text{ when} - V \text{ was sent}) \\ &= P(n_o > +V) \end{aligned} \tag{5.20}$$

Thus the probability of an error occurring in the transmission of symbols reduces to the simple condition that, at the instant the receiver output waveform is sampled, the noise voltage at the receiver output is larger in the wrong direction than the sample value of the signal. Since the noise is defined to be Gaussian, we can find the probability that the noise exceeds a given value $+V$ volts. The probability that the sampled value of

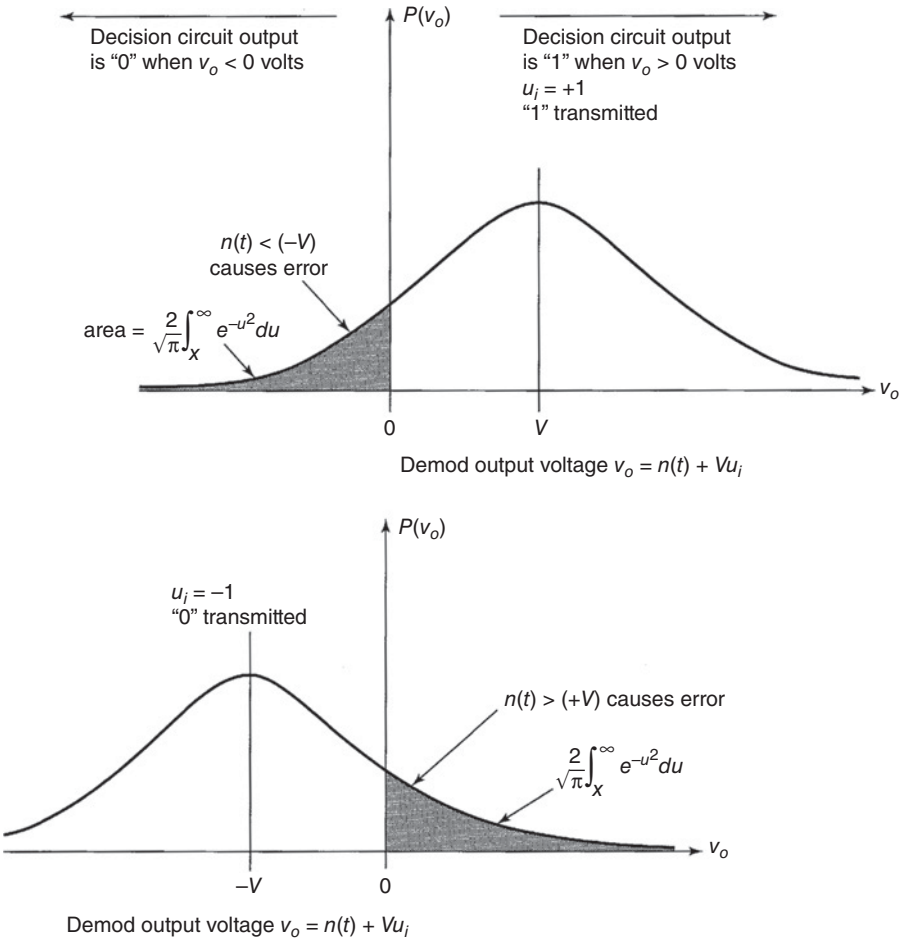


Figure 5.12 Illustration of errors in a binary decision circuit. Received pulses have amplitude $+V$ representing a data 1 and $-V$ representing a data 0. AWGN noise $n(t)$ volts with a Gaussian probability distribution is added to the signal at the sampling instant resulting in a received voltage with pdf $P(v_o)$. Errors occur in the shaded areas when the instantaneous noise voltage exceeds the signal voltage in the opposite polarity and the resulting sampled voltage $n(t) + Vu_i$ is less than 0 V when the transmitted signal was a data 1, or more than 0 V when the transmitted signal was a data 0.

the AWG noise voltage σ exceeds a value $+V$ is given by the integral of the probability distribution function (PDF) of the noise, from +V to infinity (Stremler 1999, p. 463).

$$P(n_o > +V) = \frac{\sigma}{\sqrt{2\pi}} \int_{+V}^{\infty} \exp\left[\frac{\lambda^2}{2\sigma^2}\right] d\lambda \tag{5.21}$$

where σ is the rms noise voltage and λ is the variable of integration.

The integral in Eq. (5.21) cannot be solved analytically. Numerical or approximate solutions must be used, and one such expression is known as the Q *function,* Q(z). An alternative form is the *complementary error function*, erfc(x), which is closely related to Q(z). Fortunately, there are relatively simple approximate expressions available for these functions when the probability of an error is small – which is usually the case for a

workable digital link, where we expect bit errors to occur infrequently. The condition for probability of error to be small is that $V \gg \sigma$: That is, the sampled signal must be much larger than the rms noise at the receiver output. A value of $V > 3\sigma$ makes the following approximation valid (Couch 2007, p. 662).

$$Q(z) = \frac{\sigma}{\sqrt{2\pi}} \int_{+V}^{\infty} \exp\left[\frac{\lambda^2}{2\sigma^2}\right] \approx \frac{1}{z\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) \tag{5.22}$$

The probability that n_o exceeds $\pm V$ at the sample instant is given by

$$P(n_o > +V) = Q\left(\frac{V}{\sigma}\right) \tag{5.23}$$

The complementary error function erfc(x) can also be used to find the probability of an error with a Gaussian noise voltage. The probability that n_o exceeds $\pm V$ at the sample instant is given by

$$P(n_o > +V) = {}^1\!/_2 \operatorname{erfc}\left[\frac{V}{\sigma\sqrt{2}}\right] \tag{5.24}$$

Appendix C gives the relationship of the complementary error function erfc(x) to Q(z), and tables of values for both the complimentary error function and the Q function. Note that the Q function and erfc function are defined for a normalized rms value of the Gaussian variable of one. When applied to symbol error analysis, the Q function has an rms noise voltage of one volt. Errors occur whenever noise at the sample instant exceeds either $+V$ or $-V$ volts, depending on the transmitted symbol.

The probability of error for each transmitted data state is given by the Q(z) or erfc(x) functions in Eqs. (5.23) and (5.24). We should ensure that we send equal numbers of data 1 and data 0 states to make the probability of an error occurring in the 1 state the same as the probability of an error occurring in the 0 state. This usually requires a randomizer or *scrambler* to be inserted in the data stream at the transmitter to prevent the occurrence of long strings of data 1 or 0, which would violate the required condition. The scrambler also helps the symbol clock in the receiver to stay synchronized by providing frequent phase transitions in the received signal.

One symbol lasts for a period of T_s seconds. The power in the symbol waveform is $V^2/2R$ watts where R is the input resistance of the decision circuit. We will assume a resistance $R = 1\,\Omega$, as is commonly done in the analysis of communication signals. We will assume a constant amplitude V volts for the carrier waveform (ignoring the effects of the Nyquist SRRC filters on pulse shape), so the energy per symbol, E_s, is given by

$$E_s = {}^1\!/_2\, V^2 \times T_s \text{ joules} \tag{5.25}$$

Assuming that we have a matched filter receiver, the sampled signal voltage at the demodulator output is V volts where

$$V = \sqrt{2\, E_s / T_s} \text{ volts} \tag{5.26}$$

The rms noise power at the demodulator output is $N = \sigma^2/R = \sigma^2$ watts, relative to a resistance of 1 Ω. The noise is assumed to be white and therefore has a constant NPSD, N_o watts/Hz in the noise bandwidth B_n Hz of the receiver. In a receiver with ideal SRRC filters, $B_n = 1/T_s$.

The noise power spectral density (NPSD) is given by

$$N_o = \frac{N}{B_n} = \frac{\sigma^2}{B_n} = \sigma^2 \times T_s \text{ W/Hz} \tag{5.27}$$

Hence

$$\sigma = \sqrt{N_o/T_s} \text{ V rms} \tag{5.28}$$

Combining Eqs. (5.27) and (5.28) yields

$$\frac{V}{\sigma\sqrt{2}} = \sqrt{2\frac{E_s}{T_s} \times \frac{1}{2}\frac{T_s}{N_o}} = \sqrt{\frac{E_s}{N_o}} \tag{5.29}$$

The probability that a symbol error occurs is therefore

$$P_e = \frac{1}{2}\text{erfc}\left[\sqrt{\frac{E_s}{N_o}}\right] = Q\left[\sqrt{\frac{2\,E_s}{N_o}}\right] \tag{5.30}$$

5.3.2 BPSK Bit Error Rate

For BPSK a bit and a symbol are the same, so Eq. (5.30) can be written as

$$P_b = \frac{1}{2}\text{erfc}\left[\sqrt{\frac{E_b}{N_o}}\right] = \text{Q}\left[\sqrt{\frac{2\,E_b}{N_o}}\right] \tag{5.31}$$

The analysis in Chapter 4 provides methods by which the CNR in an earth station receiver or satellite transponder can be calculated for any satellite link. The results are in terms of the ratio of carrier power to noise power at the input to a demodulator, with the ratio CNR usually given in decibels. We need to relate the CNR for a receiver to the E_s/N_o ratio that provides us with a way to calculate the probability of a symbol error.

In a receiver with ideal SRRC filters, regardless of the value of the roll-off factor α, the noise bandwidth of the filter is equal to the symbol rate, which is the reciprocal of the symbol period

$$B_n = R_s = 1/T_s$$

or

$$B_n T_s = 1 \tag{5.32}$$

A result from matched filter theory states that the energy per symbol is the carrier power times the symbol duration if the transfer function of the receiving filter matches the spectrum of the received signal. A correctly designed digital radio link with SRRC filters meets this criterion, so we have

$$E_s = C \times T_s \text{ joules} \tag{5.33}$$

and the single sided NPSD N_o W/Hz is just the noise power N watts divided by the noise bandwidth B_n in hertz

$$N_o = N/B_n \text{ W/Hz} \tag{5.34}$$

Hence for the ideal conditions specified above where $T_s\, B_n = 1$

$$\frac{E_s}{N_o} = \frac{C\, T_s\, B_n}{N} = \frac{C}{N} \tag{5.35}$$

Applying the result of Eq. (5.35) we find that the bit error rate for a BPSK signal in an ideal SRRC filtered link is

$$P_{\text{e BPSK}} = \tfrac{1}{2} \exp\left[\sqrt{\frac{C}{N}}\right] = \text{Q}\left[\sqrt{\frac{2C}{N}}\right] \tag{5.36}$$

Note that the CNR value used in Eq. (5.36) is a linear power ratio, not a decibel value. Using decibel CNRs in bit error rate (BER) equations is a frequent source of error for beginning communications engineers.

Since coherent detection is the most efficient way of demodulating direct BPSK, Eq. (5.36) is the relation normally used to determine E_b/N_o, and hence the CNR that a satellite link must maintain to meet a specified bit error rate requirement.

5.3.3 QPSK Bit Error Rate

QPSK is simply two BPSK links operating on the same radio channel with their carriers in phase quadrature. The BER for each BPSK link is identical, and given by Eq. (5.36).

When the bit stream at the transmitter is split into two to drive the I and Q channel of a QPSK transmitter, the symbol rate on the link is halved. But the error rate remains the same as if the signal had been sent as a BPSK transmission at twice the symbol rate, with the same transmitter power. This is because BER is probability of error per bit, and the probability of a bit error is the same for all bits in a QPSK system, regardless of which channel (I or Q) they travel through.

So QPSK ends up with the same BER as BPSK when considered in terms of E_b/N_o.

The total energy per symbol of a QPSK signal is therefore twice that of either of the constituent BPSK signals, or a single BPSK signal sent over the same link with the same E_b/N_o ratio. Hence

$$E_{\text{s QPSK}} = 2 \times E_{\text{b BPSK}} \tag{5.37}$$

and therefore to obtain the same error rate for a QPSK signal that we can achieve with a BPSK signal in an RF channel with a noise bandwidth B_n Hz, we require

$$(\text{CNR})_{\text{QPSK}} = 2 \times (\text{CNR})_{\text{BPSK}} \tag{5.38}$$

Thus for QPSK, transmitted at a rate R_b bits/second in a channel with noise bandwidth B_n Hz

$$P_{\text{e QPSK}} = \tfrac{1}{2} \text{erfc}\left[\sqrt{\frac{E_b}{N_o}}\right] = \text{Q}\left[\sqrt{\frac{2E_b}{N_o}}\right] = \text{Q}\left[\sqrt{\frac{C}{N}}\right] \tag{5.39}$$

The analysis of the system performance of a radio link is always carried out in terms of CNR ratio, not E_b/N_o. Many communication system textbooks leave the BER results for radio links in terms of E_b/N_o, and state that the BER performance for BPSK and QPSK are the same. However, when BER is considered as a function of CNR, BPSK, and QPSK do not have the same BER. It takes twice as much transmitter power to deliver two BPSK data streams as to deliver one. Therefore, QPSK requires 3 dB higher CNR than BPSK to achieve the same error rate when transmitting at twice the bit rate in the same channel bandwidth. If a link is operated with QPSK, the CNR required for a given error rate is 3 dB higher than when the same link is operated with BPSK. The advantage of QPSK is that it can send twice as many bits per second relative to BPSK using a channel with a

Table 5.2 Short table of Q(*z*) values

z	Q(z)
0	0.5
2.0	2.28 E-2
3.0	1.35 E-3
4.0	3.17 E-5
4.7	1.30 E-6
5.0	2.87 E-7
6.0	1.00 E-9
7.0	1.28 E-12
8.0	6.22 E-16

Approximate values for bit error rate using the Q(z) function can be obtained from this table by linear interpolation between table entries. Approximate values for BER can be estimated from this table by interpolation to an accuracy that is consistent with CNR values calculated to the nearest 0.1 dB.

specified bandwidth. This advantage of QPSK can only be exploited if the CNR at the receiver is sufficiently high.

A short table of CNR and BER for BPSK and QPSK links operating without FEC is provided in Table 5.2. This table is adequate for determination of BER in most practical cases. For example, we can use the table for Q(*z*) in Appendix C to find the exact values of bit error rate for an ideal QPSK link, where $z = \sqrt{(\text{CNR})}$. A CNR of 15.0 dB in a QPSK link requires a value of $z = 5.62$ in Eq. (5.39). Interpolation between $z = 5.0$ and $z = 6.0$ gives $\text{Q}(z) = \text{BER} = 1.1 \times 10^{-7}$. If CNR increases to 15.1 dB, $z = 5.68$ and $\text{BER} = 9.3 \times 10^{-8}$. Since CNR values are calculated to the nearest 0.1 dB, BER values around CNR of 15.0 dB cannot be determined to better than 15% and more accurate analysis is not of any value. BER changes very rapidly with small changes in CNR, so accurate calculation of BER is often not needed.

Note that when CNR exceeds 15 dB in a BPSK link and 18 dB in a QPSK link the value of *z* is greater than 8.0 and the BER falls below 10^{-15}. At this BER, in a link with a 1 Gbps bit rate a single bit error occurs once every 11 days; at 1 Mbps single bit errors occur 31 years apart. These results are statistically correct, but not useful in practice, and the link is said to be essentially error free when bit errors occur this infrequently. When FEC is applied to the data stream, the link can be operated at a lower CNR such that errors occur frequently, but the vast majority of the errors are detected and corrected.

Example 5.5

A satellite link achieves a CNR in the receiver under clear air conditions of 14.0 dB (14.0 dB = power ratio of 25.) The receiver has a SRRC filter with a noise bandwidth of 1.0 MHz and a roll off factor of 0.3, with ideal correlation detection BPSK and QPSK demodulators.

What are the bit rate, symbol rate, occupied (absolute) bandwidth of the link, and BER when the link is operated:

i) with BPSK modulation
ii) with QPSK modulation?

If rain attenuation on the link causes the CNR of the received signal to fall by 3 dB, what are the new BER values for BPSK and QPSK modulations? Assume that ideal SRRC filters are used.

Answer

For all radio links using bandpass SRRC filters, the symbol rate is equal to the noise bandwidth of the SRRC filter. Thus, for both BPSK and QPSK the symbol rate for this link is

$$R_s = B_n = 1 \text{ Msps}$$

The occupied bandwidth of the RF signal is

$$B_{occ} = R_s(1 + \alpha) = 1.3 \times R_s = 1.3 \text{ MHz}$$

Probability of error can be found from Eqs. 5.36 and 5.39 using the tables in Appendix C, or by interpolation in Table 5.2.

i) BPSK
The bit rate R_b = symbol rate $R_s = 1$ Mbps.
Using the table in Appendix C,
BER in clear air $= P_{e\ BPSK} = Q(\sqrt{2\ CNR}) = Q(\sqrt{2 \times 25}) = Q(7.07) = 8.2 \times 10^{-13}$
Interpolation in Table 5.2 gives BER $= 1.19 \times 10^{-12}$.
Time between bit errors is 14.1 days calculating the result from Appendix C and 9.7 days using interpolation in Table 5.2.

ii) QPSK
The bit rate $R_b = 2 \times$ symbol rate $R_s = 2$ Mbps.

$$\text{BER in clear air} = P_{e\ QPSK} = Q(\sqrt{CNR}) = Q(\sqrt{25}) = Q(5) = 2.9 \times 10^{-7}$$

When rain attenuation reduces the received signal by 3 dB, the receiver CNR = 11 dB (Power ratio = 12.59). The resulting BER values are

i) BPSK

$$\text{BER in clear air} = P_{e\ BPSK} = Q(\sqrt{2\ CNR}) = Q(\sqrt{2 \times 12.59}) = Q(5.02) = 2.8 \times 10^{-7}$$

ii) QPSK

$$\text{BER in clear air} = P_{e\ QPSK} = Q(\sqrt{CNR}) = Q(\sqrt{12.59}) = Q(3.55) = 2.2 \times 10^{-4}$$

All the BERs are acceptable except the last value for QPSK. With a bit rate of 2 Mbps, and a BER of 2.2×10^{-4} there are hundreds of errors occurring every second. Forward error correction would be needed in the QPSK link to maintain an acceptable BER.

Using QPSK rather than BPSK over a link with a noise bandwidth defined by the SRRC filter in the receiver doubles the bit rate. However, the CNR must be higher to sustain an acceptable error rate. A decision on whether to implement BPSK or QPSK on a given link will rest on the CNR values that can be maintained, and the length of time for which the CNR might fall to levels at which an unacceptably high BER results. In many cases, the Federal Communications Commission (FCC) requires a bandwidth efficiency of one bit per hertz or higher, eliminating the use of BPSK with added FEC.

Figure 5.13 shows bit error rate for an ideal system with SRRC filters having a fixed bandwidth B Hz, carrying BPSK and QPSK signals. The QPSK system carries twice as

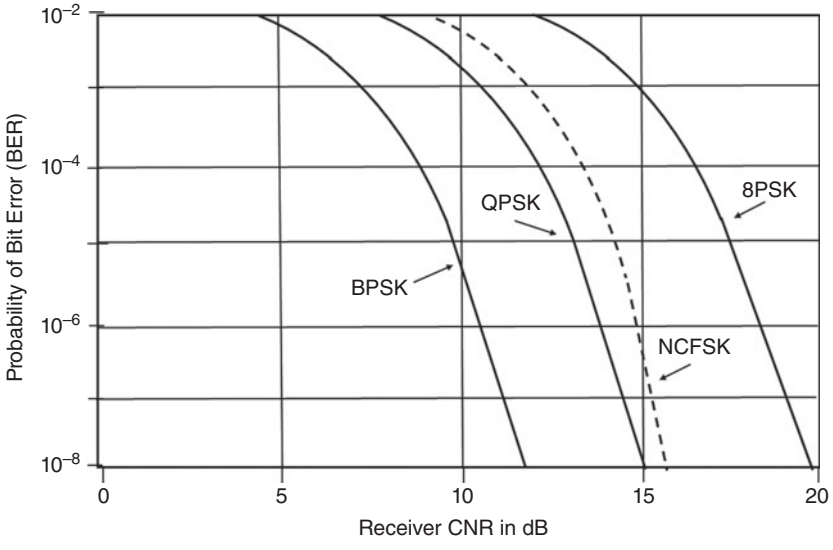


Figure 5.13 Theoretical bit error rate (BER) as a function of carrier to noise ratio (CNR) for links with ideal SRRC filters and no phase or timing jitter, resulting in zero ISI in the receiver. BPSK is binary phase shift keying, QPSK is quadrature phase shift keying, NCFSK is non-coherent detection frequency shift keying and 8PSK is eight-phase phase shift keying. FSK is rarely used on commercial satellite links, but a form called continuous phase frequency shift keying (CPFSK) has been used on some amateur satellites (Davidoff 2000).

much information (twice the number of bits carried by the BPSK system) but needs that extra 3 dB of CNR to achieve the same BER as BPSK. In a practical BPSK or QPSK radio link, which must have real filters, and inevitably suffers phase jitter in the carrier recovery circuit and timing jitter in the bit clock when the CNR is low, the ideal results shown in Figure 5.13 cannot be achieved.

An *implementation margin* must be added to the CNR to account for the difference between a real system and the ideal system for which the results of Figure 5.13 apply. In low bit rate systems, such as single channel per carrier (SCPC) channels in very small aperture terminal (VSAT) systems and low earth orbit (LEO) mobile satellite links, implementation margins as low as 0.2 dB have been reported in the literature. For high bit rate systems carrying multi-megabit per second QPSK data streams, implementation margins as high as 2 dB may be required. Hence BER for practical BPSK and QPSK satellite links is calculated from the following relationships using *effective* CNR

$$(\text{CNR})_{\text{eff}} = (\text{CNR})_{\text{o}} - \text{Implementation margin dB} \tag{5.40}$$

$$(\text{CNR})_{\text{eff ratio}} = 10^{(\text{CNR})_{\text{eff}}/10} \text{ as a ratio} \tag{5.41}$$

$$\text{BER}_{\text{BPSK}} = \tfrac{1}{2}\,\text{erfc}\sqrt{(\text{CNR}_{\text{eff ratio}})} = \text{Q}(\sqrt{2\text{CNR}_{\text{eff ratio}}}) \tag{5.42}$$

$$\text{BER}_{\text{QPSK}} = \tfrac{1}{2}\,\text{erfc}\left(\sqrt{\frac{1}{2}\text{CNR}_{\text{eff ratio}}}\right) = \text{Q}(\sqrt{\text{CNR}_{\text{eff ratio}}}) \tag{5.43}$$

Example 5.6

A satellite link uses a bandwidth of 10 MHz in a 52 MHz wide Ku-band transponder. The transmitter and receiver have SRRC bandpass filters with roll-off factor $\alpha = 0.2$. The overall $(\text{CNR})_o$ ratio for the carrier in the receiver is 16.0 dB in clear air, falling below 13.0 dB for 0.1% of an average year. The transmitter and the receiver have both BPSK and QPSK modulators and demodulators. The implementation margin for the BPSK link is 0.2 dB and for the QPSK link 0.4 dB.

Determine the bit rate that can be sent through the link with BPSK, and with QPSK. Find the bit error rate for each modulation in clear air conditions and for the 0.1% of the year conditions. Which modulation would you recommend for this system?

Answer

The symbol rate for the link is 10 Msps/1.2 = 8.33 Msps. With BPSK the bit rate equals the symbol rate, so $R_{\text{b BPSK}} = R_s = 8.33$ Mbps. With QPSK the bit rate equals twice the symbol rate, so $R_{\text{b QPSK}} = 2R_s = 16.67$ Mbps.

For the link using BPSK, the BER is found from Eq. (5.42)

$$\text{BER}_{\text{BPSK}} = \tfrac{1}{2}\,\text{erfc}\sqrt{(\text{CNR}_{\text{eff ratio}})} = \text{Q}\left(\sqrt{2\,\text{CNR}_{\text{eff ratio}}}\right)$$

In clear air $(\text{CNR})_{\text{eff}} = 16.0–0.2 = 15.8$ dB, hence $(\text{CNR})_{\text{eff ratio}} = 10^{1.58} = 38.0$

$$\text{BER}_{\text{BPSK}} = \text{Q}(\sqrt{2 \times 38.0} = \text{Q}(8.72)$$

Using the Q(z) table in Appendix C, the value of $z = 8.72$ is off the table, so $\text{BER} < 10^{-16} \approx 0$.

With BPSK, the link delivers 8.33×10^6 bits per second. With a BER of 10^{-16} a bit error will occur, on average, once every $1/(8.33 \times 10^6 \times 10^{-16}) = 1.2 \times 10^9$ seconds, assuming that bit errors occur at the same rate as symbol errors. This is a time of about 38 years, so in reality there will be no bit errors on this link. Anytime that $z > 8$ in the Q(z) expression the bit error rate on the link is effectively zero.

For 0.1% of the year $(\text{CNR})_{\text{eff}} \leq 13.0–0.2 = 12.8$ dB, and $(\text{CNR})_{\text{eff ratio}} \leq 10^{1.28} = 19.05$

$$\text{BER}_{\text{BPSK}} \geq \text{Q}(\sqrt{2 \times 19.05}) = \text{Q}(6.17)$$

Using the Q(z) table in Appendix C and approximating the value of Q(z) to $z = 6.2$, the BER will exceed 2.8×10^{-10} for 0.1% of an average year when BPSK is used as the modulation on this link. At a bit rate of 8.33 Mbps, a bit error occurs, on average, every seven minutes.

When the modulation on the link is changed to QPSK, the bit error rate will increase, as indicated by Eq. (5.43)

$$\text{BER}_{\text{QPSK}} = \tfrac{1}{2}\,\text{erfc}\left(\sqrt{\frac{1}{2}\text{CNR}_{\text{eff ratio}}}\right) = \text{Q}(\sqrt{\text{CNR}_{\text{eff ratio}}})$$

In clear air $(\text{CNR})_{\text{eff}} = 16.0–0.4 = 15.6$ dB, and $(\text{CNR})_{\text{eff ratio}} = 10^{1.56} = 36.31$

$$\text{BER}_{\text{QPSK}} = \text{Q}(\sqrt{36.31}) = \text{Q}(6.03)$$

Using the Q(z) table in Appendix C, the BER can be estimated for Q(6.0) as

$$\text{BER}_{\text{QPSK}} = 1 \times 10^{-9}$$

With QPSK, the link delivers 8.33×10^6 symbols per second (16.67 Mbps) and we will assume that a bit error occurs at the same rate as a symbol error. With $\text{CNR}_{\text{eff}} = 15.6$ dB there is a bit error, on average, once every $1/(8.33 \times 10^6 \times 10^{-9}) = 120$ seconds.

For 0.1% of the year $(\text{CNR})_{\text{eff}} \leq 13.0–0.4 = 12.6$ dB and $(\text{CNR})_{\text{eff ratio}} \leq 10^{1.26} = 18.2$

$$\text{BER}_{\text{QPSK}} \geq \text{Q}(\sqrt{18.2} = \text{Q}(4.27) \approx 10^{-5}$$

The BER will exceed 10^{-5} for 0.1% of an average year when QPSK is used as the modulation on this link. At a symbol rate of 8.33 Msps, there are 83 bit errors every second, on average, when CNR = 13.0 dB. FEC would be required to maintain the BER at an acceptable rate.

What is an acceptable bit error rate depends on the particular application. For financial transactions, a BER of 10^{-12} is typically required. Satellite systems do not often guarantee such a low error rate, so some form of error detection is needed, with retransmission of any data that are found to be in error. For general applications bit error rates of 10^{-8} to 10^{-6} are acceptable. Digital voice transmission using pulse code modulation (PCM) can withstand occasional error rates as high as 10^{-4}, which typically leads to a baseband SNR of 34 dB.

In this example, BPSK modulation has BER $< 10^{-16}$ for 99.9% of an average year, so the link is essentially error free. QPSK can deliver an acceptable error rate in clear air conditions, around 10^{-9}, but when the CNR starts to fall the BER increases quickly, so that with a drop of 3 dB in receiver CNR the BER falls to 10^{-5}. This is not adequate for general applications, but would suffice for digital voice links with a requirement that baseband SNR > 30 dB for 99.9% of the year. The obvious advantage of using QPSK is that twice as many voice channels can be carried by a 16.67 Mbps bit stream modulated with QPSK, compared to BPSK modulation, which can carry only 8.33 Mbps, within the available channel bandwidth (B_{occ}) of 10 MHz. QPSK is generally preferred over BPSK in most satellite links.

5.4 Digital Transmission of Analog Signals

The previous sections have discussed techniques for transmitting and receiving digital information via satellite. In the first half of the twentieth century, copper telephone lines carried the bulk of communication signals. The signals were either pulses used by the telegraph and teletypes (an automated form of the telegraph that connected electrical typewriters over telephone lines), or analog voltages from telephone handsets. The introduction of optical fibers and digital switches in telephone exchanges forced telephone systems to send digital data rather than analog signals. Voice signals from telephones and the output of a television camera are examples of analog signals – continuously fluctuating voltages that are proportional to a physical parameter; air pressure on a microphone with speech or music, and brightness and color of a scene viewed by a TV camera. To send analog signals over a digital link requires the conversion of the varying analog voltage level to a stream of digital words. The rate at which the analog signal must be sampled depends on its bandwidth, and the number of bits in the digital word is set by the allowable level of distortion caused by quantization. Samples of an analog signal are converted to digital words at the transmitting end of a link by an ADC

and recovered as a *quantized* version of the analog signal with a *digital to analog* (DAC) converter at the receiving end. *Quantization noise* is added to the analog signal when it is recovered from the digital words.

Appendix D describes the process of sampling and quantization that is fundamental to the conversion of analog signals into a stream of digital words. Results from Appendix D are used in this section to determine bit rates and performance of analog links that use digital transmission.

The Telegraph

Telegraph signaling predates the telephone and was the first example of digital transmission over copper wires. Samuel Morse and Ezra Cornell formed a company that developed the telegraph as the first commercial electrical communication system in the United States in the 1840s. Morse is well known for the Morse code, which was used by telegraph operators to send letters and numbers as sequences of short and long pulses called dots and dashes. However, what is generally called Morse code now is substantially different from the code that Morse devised (Worth 2006). Other telegraph systems were demonstrated in Britain and Germany before Morse's, but not developed commercially until later. Ezra Cornell formed the Western Union company with Hiram Sibley that connected cities across the United States with iron or copper wires strung on poles, and went on to found Cornell University with his share of the profits from the company.

The development of the telegraph occurred at the same time that railroads were being built across the United States. Telegraph lines could be run along the railroad right of way, and *repeater stations* were set up every 40 or 50 miles at a convenient depot. Early telegraph systems used a device called a register to mark the dots and dashes of a received signal on paper tape, but it was found that operators could listen to the clicks of the register and write out the message directly. Eventually relays replaced the telegraph operator as a way to forward messages and the device was called a *repeater*, a term that is still in use in the telephone industry for any devices that amplifies and retransmits a signal. In dry areas of the country telegraph signals could be sent up to 800 miles before requiring a repeater station, but in wet areas, 100 miles was the maximum length. Better insulation and more sensitive receivers allowed longer distance transmission, and by the 1860s transatlantic telegraph cables had been laid between Canada and Ireland. The capacitance of wires 2500 miles long resulted in slow rise and decay of the line pulses, restricting transmission rates to about 50 pulses per minute or eight words per minute with Morse code.

The telegraph had a profound effect on society. Before the introduction of the telegraph, the Pony Express was the only way to send messages quickly across the United States. The Pony Express used 120 riders, 184 stations, 400 horses, and several hundred personnel in 1861 to carry sacks of mail from Missouri to California. Pony Express riders rode day and night between stations set up every 15 miles, where they changed horses and continued for 75 miles before taking a break. It took 10 days for a message to cross the country from St. Joseph, Missouri to Sacramento, California. The first transcontinental telegraph line was completed on October 24, 1861. The Pony Express ceased business two days later.

5.4.1 Sampling and Quantizing

The basic processes in digital transmission of analog information are sampling, quantizing, encoding, and compression. The principles underlying sampling are routinely presented in beginning courses in communications theory, and are reviewed in Appendix D. See (Couch 2007, pp. 99–104; Shanmugam 1979, pp. 507–513) for further details, or search the internet for digital transmission of analog signals.

The *sampling theorem* (Nyquist 1924; Nyquist 1928) states that a signal may be reconstructed without error from regularly spaced samples taken at a rate f_s samples/second, which is at least twice the maximum frequency f_{max} Hz present in the signal. Instead of transmitting the continuous analog signal, we can transmit the samples. For example, in telephony, voice signals on satellite links are normally filtered at baseband to limit their spectra to the range 300–3400 Hz. Thus, one voice channel could be transmitted with samples taken at least 6800 times per second or, as it is usually expressed, with a minimum sampling frequency of 6800 Hz. Common telephone system practice is to use a sampling frequency of 8000 Hz. While transmitting the samples requires more bandwidth than transmitting the original waveform, the time between samples of one signal may be used to transmit samples of other signals. This is time division multiplexing (TDM).

The samples to which the sampling theorem refers are analog pulses whose amplitudes are equal to that of the original waveform at the time of sampling. If we send those (analog) pulses directly, the technique is called *pulse amplitude modulation* (PAM). The original waveform may be reconstructed without error by passing the samples through an ideal low-pass filter with a bandwidth f_{max} Hz. The bandwidth of the ideal rectangular low pass filter in the receiver is equal to the highest frequency in the analog signal, which is 3100 Hz in conventional telephone practice in the United States and 3.4 kHz elsewhere. Figure 5.14 illustrates the process of converting an analog waveform to a serial bit stream by a transmitter and recovery of the original analog waveform by a receiver. Appendix D discusses the process in more detail.

Analog pulses are subject to amplitude distortion, and they are also incompatible with baseband digital signals in which pulses take on only one of two possible values. Hence PAM is not used in communication links. Instead, the analog samples are quantized by an ADC and resolved into one of a finite number of possible values, then the quantized values are binary encoded and transmitted digitally. Thus each sample is converted into a digital word and represents the quantization value closest to the original analog sample. Quantization may be uniform or non-uniform depending on whether the quantized voltage levels are uniformly or non-uniformly spaced. At the receiver, the input waveform is sampled at the same rate as in the transmitter, which requires a clock recovery circuit, and then the samples are converted to a serial bit stream. The serial bit stream is converted to digital words that the digital-to-analog (DAC) converter outputs as a stepwise waveform. The output of the DAC is low pass filtered and the original input waveform is reconstructed. Appendix D explains the process in more detail.

A communications system that transmits digitally encoded quantized values is called a pulse code modulation (PCM) system, a rather antiquated name for a process that is not really a modulation or a code. When the PCM process is applied to voice and TV signals, they are often referred to simply as digital voice and digital TV. The standard digital word used in telephone systems has eight bits, so with sampling at 8 kHz, the bit rate of a digital telephone (PCM) channel is 64 kbps. This is called a DS0 signal.

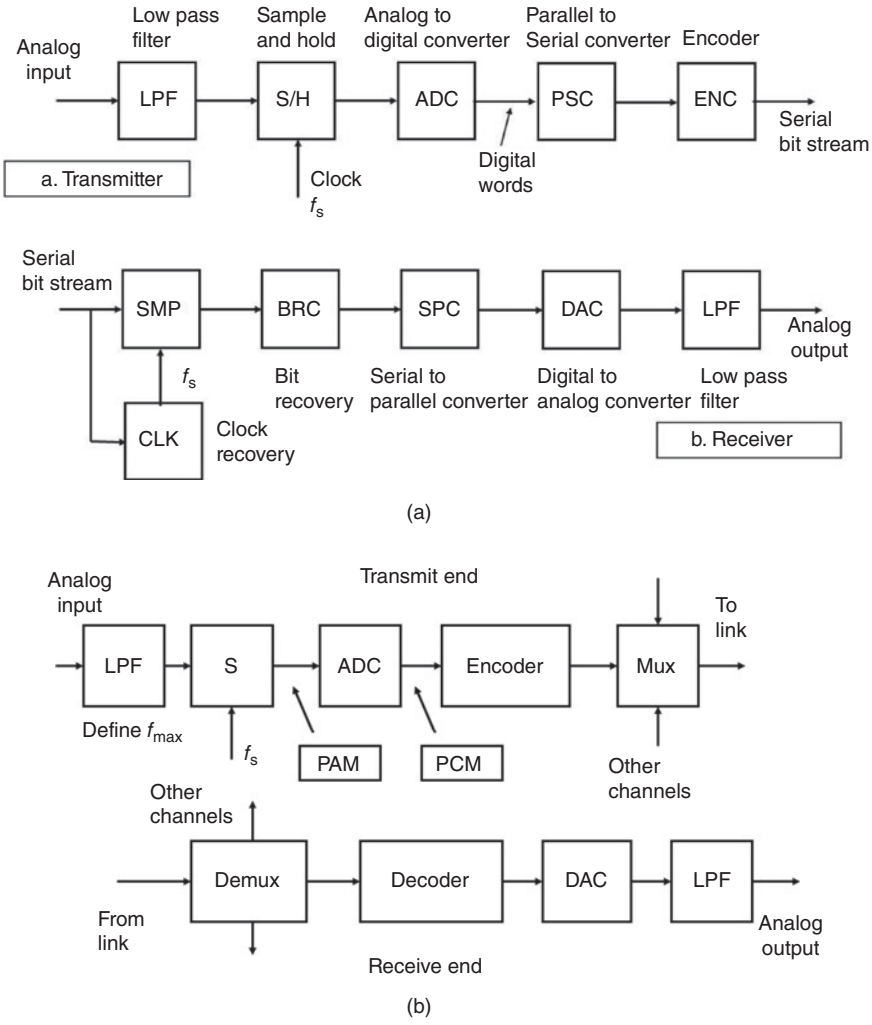


Figure 5.14 Block diagram of conversion of an analog signal to a digital bit stream and recovery at a receiver. (a) Transmitter. The analog signal is first low pass filtered to limit the maximum frequency and sampled, then converted to digital words by the ADC. The output of the ADC is converted to a serial bit stream. (b) Receiver. The input waveform is sampled by a bit clock synchronized to the bit rate of the transmitter. Samples are converted to bits and digital words which the DAC converts to a stepwise analog waveform. The low pass filter recovers the original analog waveform.

Voice signals are frequently compressed to reduce the bit rate so that a much smaller transmission bandwidth is required.

The origins of PCM voice transmission go back to the 1930s. A patent was issued in Paris to Alex Reeve in 1938 describing a PCM voice system and his US patent was granted in 1943 (IET 1979).

The first known implementation was by Bell Labs during WW II with a system code named SIGSALY.

There were no transatlantic cables capable of carrying voice signals at that time, so any voice links between Europe and the United States had to be by HF radio (short wave) using analog modulation – typically AM.

It is very difficult to encode an analog signal in such a way that it cannot easily be decoded. The WW II scrambler telephone inverted the frequencies in the audio speech signal, but was very easy to defeat.

The leaders of the western Allied forces in WW II were Winston Churchill in the UK and Franklin Roosevelt in the United States. Bell Labs built an encrypted voice system to create a secure radio link between the two leaders and their top military people so that they could converse by telephone without the enemy (Nazi Germany) being able to decode the intercepted radio transmission. It was assumed at the time that the Germans would intercept the radio signals but be unable to decode the transmissions.

SIGSALY was not a conventional PCM system as later used in digital telephony, but did incorporate many of the PCM telephone features including a six level quantizer, non-linear encoding, and a vocoder, and also introduced a randomized element that made the communications secure. The encoding and decoding equipment used vacuum tubes in 40 racks that weighed a total of 50 tons and occupied a whole room. The UK end of the link was under the US embassy in London and the US end of the link was at the Pentagon. The Bell labs workers were issued patents on their system in 1943, but these were not published until 1976 when details of the SIGSALY system were first released (Miller and Badgley 1943). To learn more, search for SIGSALY on the internet or consult references. Source: (SIGSALY, n.d.; Bennett 1983; Bennett 2009).

Commercial digital transmission of voice signals over telephone lines using PCM was first implemented by AT&T in the United States in the early 1960s. At that time, telephone exchanges (central offices in AT&T jargon) were connected by copper wires (twisted pairs). Two twisted pairs were needed to convey a two-way voice link between exchanges and as demand for telephone service increased the available twisted pairs became inadequate. Installing more copper wire, especially in underground conduits is expensive and time consuming. AT&T developed a 64 kbps PCM system that carried 24 voice channels on a single twisted pair, which was called T1. The T1 system uses a single synchronization bit added to the 192 bits from 24 voice channels that each supply an eight bit word, giving a *frame* of 193 bits and a bit rate of 1.544 Mbps. A pattern of synchronization bits is established over six frames, for example 100001, to enable the receiving end of the PCM link to synchronize to the start of the 193 bit frame.

Telephone links require additional information to be transmitted between the two ends of the link, such as on-hook and off-hook signals (indicating whether the user has lifted the handset or put it down, referring to the design of telephones in the 1920s), ring tone, and routing instructions. In the original T1 system, the least significant bit of each of the 24 voice channels of every sixth frame was *robbed* to create a separate *signaling channel* with a bit rate of 1.333 kbps. Later versions of the T1 system allocate one of the 24 voice channels for signaling, leaving 23 voice channels in the T1 frame. For more details of T1 digital voice systems refer to Appendix D or (Couch 2007; Haykin 2001; Lathi and Ding 2009).

Expansion of the connection between two telephone exchanges can be implemented by adding a PCM unit at each exchange and substituting 24 PCM voice channels for the single analog voice channel of a twisted pair. Transistors had come into use by the

1960s allowing the size and power consumption of the equipment to be greatly reduced compared to a system with vacuum tubes. T1 links could also be used to send digital data, and the 1960s saw the rapid development of digital computers, which could use the T1 links to interconnect machines. As a result, the T1 bit rate of 1.544 Mbps became a de facto standard for data links. A hierarchy of T rates was developed by AT&T by multiplexing T1 signals; for example, T2 multiplexed four T1 signals to create a serial bit rate of 6.176 Mbps.

In 2018, only the United States, Canada, and Japan employed T1 technology for voice transmission.

In the rest of the world the E1 system is used for PCM voice transmission. The E1 system creates a frame of 32 PCM words, one from each of 30 voice channels. This gives a 256 bit frame, and frames are transmitted at the rate of 8000 frames per second. The resulting bit rate is 2.048 Mbps, which is a digital transmission standard, regardless of whether digital voice or digital data are being transmitted. The additional two channels of a voice communication system are used for synchronization and control. The reader interested in the history and practice of telephone systems should refer to any of the texts on communications systems (Couch 2007; Haykin 2001; Lathi and Ding 2009). Alternatively, searching the internet for T1 or E1 PCM systems will yield many web sites that discuss these topics.

Signal Intelligence From Radio Intercepts

Interception of radio signals was widely used in WW II to gain intelligence about the enemy's plans and actions. Telegraph signals were encrypted before transmission and had to be decrypted to extract intelligence. The best known example from WW II is the German Enigma system, which was used by the German Navy, Army, and Air Force to send radio traffic between headquarters and ships at sea, army units in the field, and between command centers and airfields. Enigma signals were transmitted by hand using a Morse code key and frequency shift keying (FSK). Listeners at intercept stations marked paper tapes with the detected bits for subsequent decoding. British workers at Bletchley Park created a method of decrypting Enigma encoded signals using a mechanical device known as a Bombe, which contributed significantly to the successful conduct of the war by the British and American forces (Smith 2011; Kahn 1991). Less well known is that the Lorentz encoded teletype system used by the German High Command was also broken at Bletchley Park during WW II using a digital computer known as Colossus (Copeland 2006; Gannon 2006). The computer was designed by Alan Turing and built by Post Office engineer Tommy Flowers using 1500 vacuum tubes. Because the decryption work of Bletchley Park was a closely held secret until 1977, neither Turing nor Flowers got the recognition they deserved for building the world's first digital computer.

Interception of telephone and radio signals as a means of gathering intelligence, termed SIGINT, expanded rapidly after WW II. It was realized toward the end of WW II that intelligence gathered from intercepted radio traffic was much more reliable than trying to use networks of spies who collected human intelligence (HUMINT), since spies often worked for both sides. Allied military strategy after 1943 was based almost entirely on radio intercepts; the most useful information that spies could obtain was the code books that allowed decryption of Enigma and Lorentz encoded signals. After the end of WW II fast digital computers enabled intelligence agencies to process millions of phone calls and data transmissions, looking for specific words. Satellites were launched that

listened to cellular and microwave link traffic and continue to be a significant part of the space segment. Every country reserves the right to monitor all communication traffic going into and out of the country, and many countries also monitor all internal electronic communications.

5.4.2 Signal-to-Noise Ratio in Digital Voice Systems

Thermal noise causes bit errors in digital communication links, as discussed in Section 5.2. In a PCM system, the digital data is converted back to a baseband analog signal at the receiver. We need to know the signal-to-noise ratio (SNR) that corresponds to a given probability of a bit error occurring in the digital data at the receiver. The analysis is straightforward when only one bit error occurs in each PCM word; provided the BER is below 10^{-4} and we have seven or eight bits per word, the likelihood of two bit errors occurring in one word is very small. We will assume this to be the case in the analysis that follows.

When a bit is in error in a PCM word, the recovered sample of the baseband analog signal will be at the wrong level. A noise impulse of amplitude V_n and duration T_s, the period of one sample, is added to the original analog signal. The bit that is in error may be located in any position in the PCM word of N bits. If the least significant bit is in error, V_n is small and equal to A, the ADC step size; if it is the most significant bit that is in error, V will be large and equal to $2^{N-1} \times A$.

The resulting average SNR in the baseband analog channel caused by random bit errors is $(\text{SNR})_t$, the subscript t standing for thermal noise, which is assumed to be the cause of the random bit errors

$$(\text{SNR})_t = \frac{3\,LM^2}{1 + 4(M^2 - 1)P_b} \tag{5.44}$$

where P_b is the probability of a single bit error, and M is the number of levels in the ADC (Couch 2007, p. 144).

Quantization noise results from the received signal having a series of stepped levels instead of a continuous waveform, which leads to a quantization $(\text{SNR})_q$ given by

$$(\text{SNR})_q \approx 6\ \text{N dB} \tag{5.45}$$

when linear quantization is used (Couch 2007, p. 146).

We can combine thermal noise from Eq. (5.44) with quantization noise from Eq. (5.45) to find the baseband SNR in a PCM link

$$(\text{SNR})_{\text{PCM}} = \frac{2^{2N}}{1 + 4P_b \times 2^{2\text{N}}} \tag{5.46}$$

When P_b is small, for example, less than 10^{-6}, the quantization noise will dominate and $(\text{SNR})_{\text{PCM}} \approx 2^{2\text{N}}$. For $N =$ eight bits, quantization noise limits the baseband SNR to 48 dB. When P_b is larger, thermal noise dominates; for example, with $P_b = 10^{-4}$ and $N = 8$, the term $(4P_b \times 2^{2\text{N}}) \gg 1$, and thermal noise limits baseband SNR to $\frac{1}{4}P_b = 34$ dB.

The corresponding SNRs in analog telephone systems were set by *noise objectives*. Analog wireline telephone systems were designed to deliver a SNR of 50 dB or better in the baseband channel under ideal conditions. This was particularly challenging in long telephone links, for example across the United States, when up to 50 microwave radio links in series were needed. Noise power adds up along the chain of microwave

links with analog signals, whereas bit errors add up with digital signals, making it much easier to engineer very long digital links. A link in which the analog SNR had fallen to 30 dB was considered to be in outage, although speech is still intelligible below this SNR. The corresponding bit error rates for a digital voice link are $P_b \leq 10^{-6}$ for good quality and $P_b \geq 10^{-4}$ for the link to be in outage. For BER below 10^{-6}, quantization noise is dominant. However, when the bit error rate is around 10^{-5} there is a bit error roughly once per second in a 64 kbps PCM voice link. The bit error will be heard as a click, so the numerical value of the baseband thermal noise SNR is not really meaningful in this case.

Figure 5.13 in Section 5.2 shows that BER rises very quickly when the CNR of a digital radio system falls. For example, a QPSK satellite link with an overall receiver $(CNR)_o = 16.5$ dB and an implementation margin of 0.5 dB has a BER of 10^{-10}. In a speech channel using 64 kbps bit rate, a BER of 10^{-10} gives one bit error every two days; the channel is essentially error free and quantization noise will set the channel SNR at a subjective value of 48 dB with linear encoding. If the receiver CNR falls by 3 dB, due to rain attenuation in the path and the accompanying increase in receiver noise temperature, for example, the BER will increase to 4×10^{-6}, giving a thermal noise SNR of $¼P_b = 54$ dB. This is close to the quantization SNR of 48 dB. The link therefore operates in one of two regimes. For about 99% of the time on a typical satellite PCM voice link, when clear air conditions prevail, there are no bit errors and the baseband SNR will be 48 dB from quantization noise. For the remaining 1% of the time when attenuation occurs on the link, the baseband SNR will be below 48 dB, but will only fall below 40 dB for a very small fraction of the time when the BER exceeds 4×10^{-4}. The techniques described in Chapter 4 are used to design satellite communication links so that bit error rates can be maintained above 4×10^{-4} for all but a very small percentage of the time.

Non-linear encoding is used on digital telephone links to avoid the *quiet talker problem*. Voice levels in telephone links vary from +2 dB for a loud talker to −30 dB for a quiet talker, referenced to a nominal signal level of 0 dBm. The gains of amplifiers along a telephone link are set so that a 1 kHz sine wave (*a tone*) transmitted at a power level of 0 dBm, corresponding to an rms voltage of 0.775 V in the standard resistance of 600 Ω used in wireline telephone links, is delivered to the other end of the link at 0 dBm. The quiet talker produces a range of voltages that cover only the lower levels of the ADC, resulting in much higher quantization noise because of the smaller number of bits that are generated. Non-linear encoding provides more levels at the low voltage end of the ADC range, in exchange for fewer levels, and therefore larger voltage steps, at the higher end. Although this means there is more quantization noise for the loud talker, the effect is not noticeable. There are two commonly used non-linear ADC functions in wireline telephony: μ-law and A-law. (See Appendix D for details.)

Speech compression is widely used in digital voice links to reduce the required bit rate. Cellular telephones typically compress voice signals to 4.8 kbps and some military systems use rates as low as 2.4 kbps (Rappaport 2002, pp.429–436). Appendix D includes a simplified explanation of how speech compression works. An internet search for speech compression will yield many articles and papers on the various techniques. One widely used speech compression technique is *linear predictive encoding* (LPC), implemented as CLPC (*code book excited LPC*) in global system mobile (GSM) cell phones. Speech compression is achieved by the use of DSP integrated circuits or software running on a fast microprocessor. The DSP IC contains a fast microprocessor and a large memory, and may be executing many millions of operations a second. The analog speech waveform is sampled and digitized, typically at 64 kbps for telephony, and then processed by the DSP compression IC to reduce the bit rate. The low bit rate signal is transmitted to a

decoding IC in the receiver, which regenerates the 64 kbps bit stream and thus recovers the original speech waveform. What the listener to a cellular phone is hearing is not the voice of the speaker at the transmitting end of the link, rather the output of a *vocoder*, which is an artificial voice generator. The challenge in developing speech compression systems is to provide a natural sounding voice at the receive end of the link, which works well in any language and provides speaker recognition and intelligibility.

Speech compression techniques are termed lossy because some of the information in the original analog signal is lost in the compression and decompression processes. The theoretical SNR values no longer apply, and a *mean opinion score* (MOS) with a value between 0 and 5 is used to quantify the quality of the recovered audio. A MOS is obtained by having a panel of typical users listen to male and female voices reading lists of words and phrases. The lists include words that are easily confused; for example, ban and van, cam and can, where the difference between the consonants *b* and *v*, *m* and *n* can be difficult to discern when the speaker is speaking in English. However, other languages have different sounds that can be confused, so designing a speech compression system that works well in all languages is challenging.

Two factors are evaluated when determining a MOS: intelligibility, and the quality of the recovered speech. The listeners write down what they hear and their opinion of the quality of the speech, and the panel's results are used to create the MOS. Face to face conversation should score a MOS of five. An analog wireline telephone link without compression should score around 4.5, and a standard PCM link around 4.3. Cellular telephones and other wireless devices using heavy compression score closer to 3.0. The process of determining MOS scores has been automated to replace human subjects with electronic measuring equipment. However, the equipment does not necessarily measure all of the defects in a communications link that can be annoying to users (Tektronix® white paper 2009).

The development of vocoders led to the availability of text to speech conversion, as used in telephone answering machines and weather forecasts broadcast by the National Weather Service in the United States. Voice to text conversion followed, allowing computers to respond to voice commands, and the widespread use of (annoying) telephone trees by commercial entities that try to replace a human agent who answers your telephone call with an inanimate machine. The most valuable use of the technology is real time translation of text between different languages, so that, for example, a French listener hears in French what an English speaker is saying in English, and vice versa.

5.4.3 Digital Television

Digital television has replaced analog TV for DBS-TV, cable TV, and terrestrial broadcasting of all television signals in the United States and many other countries. A television signal is made up of two separate parts: the *video* signal, which creates the picture at the receiver, and the *audio* part that carries the accompanying sound signals. The video and audio signals are digitized separately, and then multiplexed into a serial bit stream for transmission. The audio signal is typically digitized with more bits than a telephone channel to provide sound of good quality.

Video signals are divided into three components: the *luminance* component describes the brightness of the scene viewed by the TV camera and two *chrominance* components that provide information about the colors present. Any color can be created by

combining two colored light sources in different amounts. A digital color TV receiver has a screen made up of millions of light emitting diodes (LEDs), which are grouped in threes, typically producing red, green, and blue light. The appropriate combination of these three colors produces white light, or any other color. When light is reflected from a surface, for example a painting, the primary colors are different from the case of transmitted light. The primary colors for reflected light are red, yellow, and blue.

Television in the United States dates back to the early 1950s, with all transmissions being in black and white as only the luminance of the scene was transmitted. A television camera scanned the scene to be viewed at a rate of 30 frames per second consisting of 525 lines (the *raster*). The frames were transmitted at 60 Hz as alternating half frames taken from the odd and even numbered lines to reduce flicker on the TV screen. The result was an analog signal with a bandwidth of 4.2 MHz, which was transmitted using vestigial sideband amplitude modulation in an RF bandwidth of 6.0 MHz. Terrestrial broadcast channels were established in the vhf and uhf bands with 6 MHz spacing. In the early 1950s the FCC formed the National Television Standards Committee (NTSC) to determine a way to add color to the analog TV signal without requiring a wider RF channel bandwidth. The NTSC solution added two *chrominance* signals to convey color information in the form of quadrature amplitude modulation (QAM) of a sub-carrier centered at 3.58 MHz and overlaid on the higher frequency portion of the luminance signal. The choice of sub-carrier frequency was set by complicated considerations designed to minimize the interference that the chrominance signals would cause in the luminance signal. The bandwidth of the chrominance signals was less than 1 MHz, so color definition in the NTSC system was poor compared to movies or printed pictures. The analog TV system came to be known as NTSC-TV, and lasted in the United States as the terrestrial broadcast TV standard for over 50 years until broadcast stations changed to digital television in 2009. For a full description of the NTSC system see reference (Couch 2007, pp. 613–632) or search the internet for *NTSC TV*.

In the 1990s, the FCC formed the Advanced Television Standards Committee (ATSC) to set the standards for digital TV. A high definition (HD) digital TV displays 1920 pixels horizontally and 1080 pixels vertically on the screen for a total of approximately two million pixels. A frame refresh rate of 30 Hz would require the transmission of 60 million pixels per second, and three components must be transmitted, the luminance and two chrominance signals. If we allocate eight bits to each component the transmission rate would be 16.6 Gbps. This is not compatible with transmissions in an RF bandwidth of 6 MHz, as demanded by the FCC to avoid changing the broadcast TV channel allocations. In the ATSC video system, the analog outputs of a color TV camera (red, blue, green, or three equivalents) are combined to produce the luminance and two color signals. Luminance is sampled at 13.5 MHz and the two color components at 6.75 MHz, then converted to eight bit digital words giving a raw bit rate of 216 Mbps. The rate is reduced to 166 Mbps by omitting the time gap between frames.

To overcome the high bit rates required when a video signal is sampled and digitized, compression techniques have been developed that reduce the bit rate by a factor of 40 or more. The most important video compression techniques are MPEG-2 and MPEG-4, where MPEG stands for the *Motion Pictures Experts Group,* an industry standards body. MPEG compression techniques are used in satellite television, digital video discs

(DVDs), and terrestrial broadcast TV. The MPEG system divides the picture into 8 by 8 pixel blocks and takes a *discrete cosine transform* (DCT) of each block. Only the significant coefficients of the DCT are then transmitted, which greatly reduces the bit rate needed to send each 8×8 pixel block. MPEG-2 and MPEG-4 also use frame to frame comparison of the video signal to determine which blocks within each frame need to be transmitted because a change has occurred between frames. Compression factors of up to 75 can be achieved with full motion video using MPEG-2 techniques. If degradation of the picture is allowed, even higher compression ratios are possible, but definition is degraded and motion becomes jerky or blurred because fewer frames are sent per second. MPEG-2 was key to the development of digital satellite television systems in the 1990s, with a transition to MPEG-4 for transmission of HDTV. The ATSC system uses MPEG-2 or MPEG-4 compression to reduce the transmitted rate to 19.4 Mbps for terrestrial broadcasting. The DVB-S standard used for satellite television employs MPEG-2 for standard definition TV transmissions and the DVB-S2 standard uses MPEG-4 for HDTV with rates that can be as low as 2 Mbps for a standard definition television signal.

Similar compression techniques are used for the transmission of pictures over the internet using JPEG (*Joint Picture Experts Group*) standards. The human eye is less sensitive to noise and distortion in a video signal than the human ear is to distortions in audio signals. This allows video and picture transmissions to use fewer bits and higher quantization noise levels than audio transmissions. One major disadvantage of MPEG-2 compression is that it introduces a delay of about a second in the transmission of the TV signal. The delay is of no significance in recorded material, but is apparent in two-way live transmissions such as interviews between a TV studio and a reporter in the field.

The digital video broadcast standard used in DBS-TV of standard definition video signals (DVB-S) has an average transmission rate of between 2 and 4 Mbps for live video and 1.6 Mbps for pre-recorded material. This allows up to 10 video signals to be transmitted by a single 27 MHz bandwidth transponder. For more details of DBS-TV see Chapter 10.

5.5 Time Division Multiplexing

In TDM a group of signals take turns using a channel. This contrasts to FDM where all the signals occupy the channel at the same time but on different carrier frequencies. Since digital signals are precisely timed and can consist of groups of short pulses with relatively long intervals between them, TDM is the natural way to combine digital signals for transmission. Additional bits must be added to the data stream when TDM is used so that data bits can be identified correctly and the receiver can synchronize to the packets or frames containing data. Precise timing is needed in the receiver to make sure that the correct bits are extracted.

Figure 5.15a illustrates the transmitting end of a TDM link for n data channels. The data channels D1 through Dn (often called packets) are multiplexed into a serial bit stream and two signaling channels S1 and S2 are added to form a frame of $n + 2$ channels. A cyclic redundancy check (CRC) is applied to the frame and the result added at the end of the frame as a CRC value. If the data channels are carrying PCM voice, there are 30 voice channel of eight bits with two eight bit signaling channels and no CRC. This gives a 32 channel E1 frame of 256 bits. With standard 64 kbps PCM voice channels, the transmission rate is 2.048 Mbps. Because voice channels must be transmitted in near

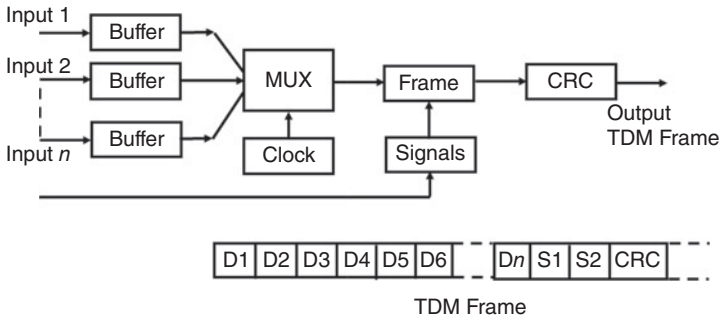


Figure 5.15a Transmit end of a TDM link for *n* channels. Incoming data is buffered and clocked into the multiplexer (MUX) to create a serial bit stream of data blocks, D1, D2 … D*n*. Bits from two signaling channels (S1, S2) are added to create the TDM frame and a cyclic redundancy check (CRC) is added to the end of the frame. If the TDM system carries voice channels, the CRC is omitted. Data blocks are often called packets.

real time, it is not possible to go back and correct errors. If the channels are carrying data rather than voice, the number of bits per channel in the frame can have any value, and the signaling blocks are replaced by control and address blocks.

Figure 5.15b illustrates the receiving end of the TDM link for the TDM frame generated in Figure 5.15a. The TDM frames are clocked into an input buffer and the CRC value is checked. The signaling blocks are extracted and sent to output buffers, and the data channels are then demultiplexed and sent to output buffers. If the CRC value was correct, data is released from the output buffers to the separate output lines. The clock in the receiver is synchronized to the incoming data stream and feeds a counter that drives the demultiplexer (DeMUX).

5.5.1 Comparison of GEO Satellites and Optical Fibers for Point to Point Links

The rapid development of optical fiber systems allowed very high speed digital signals to be transmitted over long distances. Optical fibers have very large bandwidths compared to radio links, and transmission has been demonstrated at rates exceeding 100 Gbps. Optical fibers operated as part of a *synchronous optical network* (SONET) have rates that are multiples of the base rate OC-1 at 51.84 Mbps (Optical fiber 2018). Since

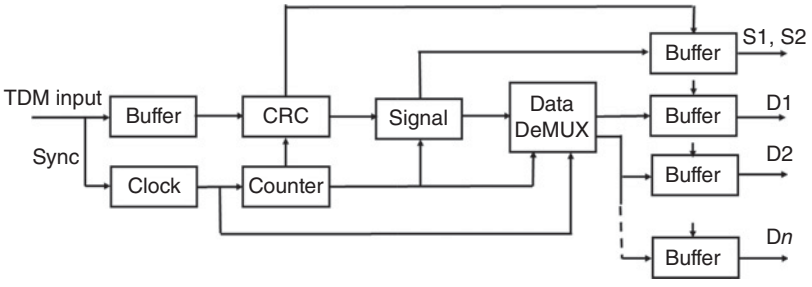


Figure 5.15b Receive end of the TDM link for the frames created in Figure 5.15a. The clock is synchronized to the incoming data stream and feeds a counter that drives the demultiplexer. The cyclic redundancy check value (CRC) is checked first. If the CRC indicates one or more errors are present in the frame, the output buffers are blocked. S1, S2 are the signaling output channels. D1, D2 … D*n* are the data output channels.

Table 5.3 Fiber optic cable transmission standards

Designation	OC-1	OC-3	OC-12	OC-24	OC-48
Bit rate	51.84 Mbps	155.52 Mbps	622.08 Mbps	1.244 Gbps	2.488 Gbps
Designation	OC-192	OC-768	OC-1920	OC-3840	
Bit rate	9.953 Gbps	39.813 Gbps	99.532 Gbps	200 Gbps	

optical fiber cables rarely contain a single fiber, but more usually between ten and hundreds of fibers, the capacity of an optical cable system is very large. The largest and heaviest GEO satellites in orbit in 2018 had a maximum capacity of 140 Gbps, equivalent of 48 OC-28 circuits or 14 typical optical fibers operating at 10 Gbps. For point to point communication, satellites cannot compete financially with optical fibers once they are laid; that is why the majority of earnings from commercial operation of satellites comes from broadcasting or point to multi-point transmission. Table 5.3 shows some of the standard bit rates for fiber optic cables; all are multiples of 51.48 Mbps, indicated by the number following the OC designation. Interconnection of satellite links and terrestrial circuits is feasible at rates up to OC-12, especially with the later generation of Ka-band GEO satellites using higher order modulations.

5.5.2 Channel Synchronization in TDM

We have made the tacit assumption that all incoming PCM channels are synchronized with each other and running at the same bit rate, and therefore can be readily multiplexed into a higher bit rate stream. This condition would hold if the voice channels had reached the originating earth station in analog form and had been digitized by ADCs running on a common clock. But if the channels come into the station in digital form, their synchronization cannot be guaranteed. They may need to be resynchronized for TDM transmission by a technique called pulse stuffing.

In pulse stuffing, at the transmitter the incoming words for each channel flow into an elastic buffer. There is one such buffer per channel, and each buffer can hold several words. The multiplexer reads words out of the buffer slightly faster than they come in. Periodically the multiplexer will go to the buffer and find less than a full word remaining. When that happens it inserts a dummy word called a *stuff word* into the frame in place of the word it would have taken from the buffer. At the same time it places a message on the signaling channel that states that a stuff word has been inserted. When the DeMUX at the other end of the link receives the message it ignores the stuff word. When it is time for the next frame to be sent the buffer will have more than a full word waiting for transmission.

Pulse stuffing is normal on satellite links that transmit digital signals between different terrestrial TDM systems, because the TDM systems may not be synchronized. The satellite link is run at a bit rate slightly higher than either of the terrestrial TDM systems that it joins. Stuffing bits and words are added to the satellite data stream as needed to fill empty bit and word spaces.

5.6 Packets, Frames, and Protocols

Digital data, from whatever source consists of streams of ones and zeroes. In a communication system, the ones and zeroes may be digital words generated from analog

voice or video sources, or they might be data from one computer on the internet destined to another computer. All ones and zeroes look alike, so *frames* or *packets* are used to enable the receiving end of a link to identify what is being sent and to synchronize to the data. Frames tend to be longer than packets, but both serve the same purpose and form part of a *protocol*. A protocol is defined as a code of conduct for diplomatic relationships between countries, but as used by computer engineers a protocol defines a standard procedure for regulating data transmission. There are many different protocols in existence for different applications. TCP/IP (transmission control protocol/internet protocol) allows any device to access the internet by defining how to set up a connection, how to send or receive data, and how to disconnect at the end of a session. Other well known protocols are Bluetooth, used for short range radio communications with frequency hopping, IEEE 802.11 for wireless data communications in unlicensed frequency bands, and GSM used in cellular phones. The protocols used on satellite links are usually proprietary to the users of the link, and in particular, TCP/IP cannot be used over GEO satellite links because the delay exceeds a 60 ms limit that is part of the TCP/IP protocol.

In a voice telephone system, protocol is controlled by the persons making and receiving the telephone call, so there is no set procedure. However, there are good and bad procedures. If you want to find a specific answer to an enquiry, you hope that a real live person will answer the telephone and provide you with the information you want. Instead, you may hear a recorded voice that gives you six options, selectable by pressing numbers one through six on the telephone keypad, none of which connect you to a person who can answer your enquiry. A further option is to press zero, which takes you back to the beginning of the same sequence. That is a familiar and annoying procedure, implemented by organizations that are trying to save money by not employing a sufficient number of customer service agents.

Data transmission systems must be designed to not require human intervention. For example, requesting a page from a web site requires a well defined protocol that can first connect your computer to the internet, send your request to the correct web site, instruct the web site which page you want, and then arrange for that data stream to be sent to your personal internet address, and finally disconnect your computer from the internet. The process is necessarily complex and is the subject of many texts and papers that fall in an overlapping area between computer science and electrical engineering.

The structure of all data communication systems is defined by the International Systems Organization (ISO) in the seven layers of the Open Systems Interconnect (OSI) model. A simplified version of the model is shown in Table 5.4. The lowest level is called the *physical layer*, which is where bits are sent between transmitters and receivers. That

Table 5.4 The ISO-OSI seven layer model

Layer 7	Application	Human-machine interfaces
Layer 6	Presentation	Encryption, compression
Layer 5	Session	Authentication, permissions
Layer 4	Transport	End to end error control
Layer 3	Network	Controls transmissions over the network
Layer 2	Data link	Frames and packets are defined
Layer 1	Physical	Bits are transmitted and received

Figure 5.16 Generic packet structure. The SYNC block is used to synchronize the bit clock in the receiver, FLAG (or a unique word) indicates the start of the packet, ADDR is where the transmitter and receiver addresses can be inserted, if needed, and CNTL is where control bits are inserted. The DATA block can be of fixed length, or variable length in which case the number of data bits will be given in CNTL. CKSM is a checksum used by the receiver to verify that errors have not occurred in the packet. Frequently, a cyclic redundancy check (CRC) is used in preference to a checksum because a CRC can find multiple errors in a packet.

is where radio links such as satellite communications exist, and is the province of communications engineers. In the second layer, the data link layer, are the protocols that define the way data is collected into frames or packets. The network layer may include error detection and correction methods, which are important to the performance of satellite links. We will not be concerned with the details of the OSI model, which are largely the province of computer engineers and are implemented in software running in the terminals at each end of a communications link (Microsoft Support 2017). The OSI model is rarely used by communication engineers because the many parts of a real digital communication system do not align well with the seven layers of the OSI model.

Frames and packets have a common set of blocks that fulfill specific functions. Figure 5.16 shows a generic packet that is not related to any particular communication system; it simply illustrates the general way in which packets can be constructed. The generic packet has a variable number of eight bit data words in the data section and the modulation is QPSK.

The block marked SYNC contains a string of ones and zeroes to help the receiver synchronize the local carrier used in the QPSK demodulator and also to synchronize the bit clock that extracts data. The SYNC block is commonly used in frames, when frames arrive from different transmitters as in TDMA system. If a series of packets is received from the same transmitter, SYNC can be omitted after the first packet because the receiver will stay synchronized. The SYNC block is followed by FLAG, for example, the eight bit sequence 10000001. Flag identifies the start of the packet, and also the end of the packet because it occurs at the beginning of the next packet. At the transmit end of the link, the data stream is examined to see whether the 10000001 sequence occurs in the data stream. If it does, a different sequence is substituted and control bits are added to inform the receiver of the change. Alternatively, the transmitter may not allow a sequence in the data containing six consecutive ones or zeroes. Flag is then easily identified because it cannot be part of a data stream. In some packets or frames, flag takes the form of a long *unique word*, a sequence of bits that can readily be identified at the start of the packet or frame and may be as long as 48 bits. Although the unique word sequence might occasionally occur in the data stream, the receiver is designed to require several repetitions of the unique word in successive frames before it locks to the unique word location.

ADDR typically contains the address of the transmitter sending the packet and the address of the destination to which the packet is sent. In the internet, addresses are URLs, but in a VSAT network addresses may be defined in a completely different way, and may not be present in every packet. In DBS-TV the ADDR block can contain a unique address for each customer.

The customer's address is used to send messages and programming table updates to individual receivers because customers can buy different packages of programming and therefore have different tables. Messages can be sent for display on the TV receiver; for example, if a customer fails to pay their bill, a message will be sent saying *call the customer service center*. The receiver may eventually be turned off if the bill remains unpaid for too long.

CNTL is a control block of a specified number of bits, which provide information to the receiver about the data that follows; for example, the type of packet because not every packet carries data, the number of words in the packet if it does carry data, etc. The data block, DATA in Figure 5.16, contains either a fixed or a variable number of words, or a fixed number of bits, as indicated in the CNTL block.

At the end of the generic packet is a checksum (CKSM) or a CRC. The check sum word is created at the transmitter by counting the number of ones in the packet. The receiver makes the same count; if the numbers disagree the receiver can reject the packet, request a retransmission of the packet, or simply ignore the error. Those options are discussed in the next section on error control.

5.7 Error Control

Satellite links are subject to propagation impairments such as rain attenuation in the earth's atmosphere, which cause the CNR at the receiving end of the link to fall. As we saw in Section 5.2, the bit error rate on a digital link increases very rapidly as the CNR falls. This is illustrated in Figure 5.13 for the specific cases of BPSK and QPSK modulation, where CNR must be maintained above 10.6 dB for BPSK and 13.6 dB for QPSK to ensure the BER is below 10^{-6}. The target value for BER depends on the application; 10^{-6} is a typical value, but digital voice systems may allow BER to fall to 10^{-4} while the transmission of financial data typically requires BER not to exceed 10^{-12}. Satellite links can be designed to operate with almost any CNR greater than 0 dB in the receiver, but the error rate at the receiver output will be high when the CNR is low. The error rate can be improved by applying forward error correction (FEC) to the data stream. Additional *coding bits* are added to the data that allow the receiver to detect and correct bit errors. The topic of FEC codes is complex and specialized, largely the province of mathematicians, and will be discussed here only so far as it relates to satellite communication systems.

Claude Shannon, working at Bell Labs in the 1940s and 1950s, developed information theory that shows there is a minimum CNR of −1.6 dB below which bits cannot be recovered without error, regardless of the FEC method that is used (Shannon 1948). The limit can be approached by using FEC techniques such as *turbo coding* and *low density parity check* (LDPC) *coding*, which make use of iterative techniques to correct errors. LDPC coding is used in the DVB-S2 standard for high definition DBS-TV transmission by satellite.

5.7.1 Error Detection and Correction

There are several ways in which the problem of ensuring that data output by a receiver has an acceptable rate of errors, depending on the application. Error detection is always easier than error correction and requires fewer coding bits, so some systems only detect errors without attempting to correct them. When an error is detected it can be ignored

and the rate at which errors occur can be calculated and used as a guide to the quality of the link. Too many errors in a given time may cause the link to shut down. The digital transmission of analog data such as voice, music, or video requires the reconstruction of the analog waveform in the receiver. When a single isolated error is detected, its location can be flagged and interpolation between adjacent samples can be used to fill in for the errored sample. In systems that require accurate transmission of data, errors that are detected in a packet will trigger a request for a retransmission of that packet. This is known as ARQ, *automatic repeat request*. ARQ inserts a delay into the reception of data and cannot be applied successfully to voice or video links over a GEO satellite. The TCP/IP protocol used for data transfer to and from the internet employs ARQ with powerful error detecting codes to ensure that all information is transferred correctly.

GEO satellites links cannot employ the TCP/IP protocol, as noted earlier, because of the long round trip delay on a GEO satellite link. The protocol of the satellite link must include an ARQ capability such as *continuous transmission* ARQ or *selective repeat* ARQ, which numbers all packets and sends a request from the receive end to the transmit end when a packet error is detected. The transmitter then inserts a second copy of the errored packet into the transmitted packet stream. This results in packets arriving at the receive end of the link out of sequence and this requires buffering of at least 500 ms in the receiver to get the packets back into the correct order, increasing the latency of the transmission. More details of error detection and correction on satellite internet access can be found in Chapter 11.

Forward error correction (FEC) coding is widely used to correct errors at the receiving end of a digital link, and also to correct errors that occur when digital data is recovered from a memory. The latter application was the incentive for much of the work on error detection and correction coding in the 1960s, rather than for telecommunications. The concept of correcting errors in a bit stream is attributed to Richard Hamming, a colleague of Claude Shannon, working at Bell Labs. Hamming created the first *linear error correcting block codes* and showed how to implement the process of detecting and correcting errors (Hamming 1950, pp. 147–160).

In the material that follows, no attempt is made to explain information theory or the mathematics that lies behind the generation and decoding of error detecting and correcting codes.

Those topics are the subject of many text books and papers, published mainly in mathematical journals (Lin and Costello 1983; Drury et al. 2000; Krouk and Seminov 2011). What follows is sufficient information for a beginning satellite communications engineer to understand how error control is applied in satellite links, and the cost of applying FEC coding in terms of data transmission rates.

The transmission of information over a satellite communication system always results in some degradation in the quality of the information. In analog links the degradation takes the form of a decrease in SNR. In analog radio systems wideband FM is used to trade bandwidth for power and achieve a large baseband SNR improvement. Terrestrial radio broadcasts use FM with a 200 kHz RF channel employed to transmit 15 kHz bandwidth audio signals. With an additional technique called pre-emphasis that reduces high frequency noise in the audio signal at the receiver, a SNR improvement of 39 dB is obtained provided the CNR is above 13 dB. Hence, provided the received CNR is greater than 13 dB, good quality audio with SNR > 50 dB will be output by the receiver. In digital links we measure degradation of the information content of a signal in terms of the bit error rate. By using PSK, we can trade bandwidth for signal power and achieve low

bit error rates with low CNRs. The RF bandwidth required to transmit a digital voice signal using QPSK is comparable to the RF bandwidth needed with FM, and requires a minimum CNR of 13 dB for a baseband audio SNR of 48 dB if FEC is not applied to the signals.

A fundamental difference between analog and digital signals is that we can improve the quality of a digital signal by the use of error correction techniques. No such technique is available for analog signals since once the information is contaminated by noise, it is extremely difficult to remove the noise, as we cannot in general distinguish between the signal and the noise electronically. In a digital system, we can add extra *coding bits* to our data stream prior to transmission, which can tell us when an error occurs in the data on reception and can also point to the particular bit or bits that have been corrupted. Systems that can only detect errors use *error detection*. Systems that can detect and correct errors use FEC. Systems that have only error detection must make a decision about what action to take when an error is detected. The options are to ignore the error, to flag the error, to send a block of information again, or to estimate the error and replace the corrupted data. Which option is selected depends on the nature of the signal that is transmitted. Collectively, these techniques are known as *error control*. In advanced digital satellite communication links FEC may be switched in and out on demand, or the code rate changed, depending on the measured bit error rate or CNR at a terminal.

Some confusion surrounds the term *coding,* since it is applied to several different processes, not all of which are concerned with error detection and correction. In the popular sense, *coding* is used to describe the rearrangement of information to prevent unauthorized use. This process is known technically as *encryption*. It is widely used on digital signals that are sent by cable and radio links. Encryption of digital signals is achieved by convolving the data bits with a long code sequence to destroy the intelligibility of the baseband data. To recover the information, the code sequence used in the encryption process must be known to the recipient; this information is contained in the *key* to the code, which may be changed at frequent intervals to maintain good security. We will not be concerned any further in this chapter with encryption.

Coding is also a name applied to many processes that change data from one form to another. For example, PCM changes analog data into binary words for transmission over a digital link. It is fundamental to the transmission of voice and video signals by digital techniques, and uses a device commonly called a *codec,* short for coder-decoder. The term *coding* is also applied to devices that scramble a digital data stream to prevent the occurrence of long strings of 1s and 0s data bits.

Throughout this chapter we shall use the term *coding* to refer to error detection or error correction. This implies that additional (coding) bits are added to the data stream prior to transmission to form an error-detecting or error-correcting code. It is possible, in theory, to generate codes that can detect or correct every error in a given data stream. In practice, there is a trade-off between the number of coding bits added to the information data bits and the rate at which information is sent over the link. The *efficiency* of a coding scheme is a measure of the number of coding bits that must be added to detect or correct a given number of errors. In some FEC systems the number of coding bits is equal to the number of data bits, resulting in a halving of the data rate for a given channel transmission rate. This is called *half rate FEC*. The loss of communication capacity is traded for a guaranteed lower error rate. This technique is widely used in internet access systems and DBS-TV where the links have low CNRs.

The reduction in BER at the baseband output of the receiver is roughly equivalent to a 3 dB improvement in CNR when the additional bandwidth required to transmit the half rate encoded signal is taken into account.

A 3 dB improvement in CNR can be obtained by increasing the diameter of the receiving terminal antenna by 41%, but it is an expensive and unwieldy option compared to inserting half rate FEC into the terminal's bit stream. Consequently, all satellite terminals that tend to have low CNRs (VSATs, satellite telephones, DBS-TV terminals, internet access) make use of FEC to improve the bit error rate on the links to and from the small terminal.

The DVB-S and DVB-S2 standards allow for different FEC rates to be applied to the signal.

With the DVB-S standard the modulation is QPSK and any of five FEC code rates 1/2, 2/3, 3/4, 5/6, and 7/8 can be used. In direct to home satellite TV broadcasting using the DVB-S standard one code rate is selected, for example, rate 3/4, and is not changed. This is simple to implement but inefficient, because under clear air conditions the error rate on the link may be very low, and FEC is needed for only a small part of the time when rain attenuation affects the link. However, DBS-TV receivers can be designed so that the code rate can be changed by sending command signals to all receivers. In the DVB-S2 standard, there are 28 combinations of modulation and FEC rate (European Television Standards Institute 2009). With a return link from the receiver to the transmitting station, a combination of modulation and code rate can be selected to optimize performance on that specific link. Thus links with clear air conditions can operate at the highest efficiency and only require a lower order modulation and more FEC bits when attenuation affects that link.

Error control for DBS-TV systems is discussed in Chapter 10, and for internet access systems in Chapter 11. Systems having two-way connections, such as internet access can establish a virtual circuit between the gateway station and each customer. This allows the coding and modulation to be changed for individual receivers that are suffering from a propagation disturbance by adding more FEC coding or reducing the order of modulation, while other links operate at maximum efficiency. The data rate on affected links will be reduced during the propagation event, but this is preferable to losing the connection. Since most propagation disturbances are caused by rain affecting the link between the satellite and the customer, the user of the internet access system is likely to be aware of a possible slowing of the data rate when it is raining. Without a return connection, as in most DBS-TV systems, ACM cannot be implemented. However, if the DBS-TV satellite has multiple beams, downlink transmitted power can be increased in beams that serve regions affected by rain attenuation, and different coding and modulation can be applied to different beams. Beams covering regions of the country where there are no propagation disturbances can deliver the highest bit rates using lower transmitted power.

With digital data such as internet access, file transfers, and email, some measures must be taken to guard against errors, and the end user will normally determine how this achieved. As noted earlier, financial transactions and records are required to be transmitted with a BER of 10^{-12}. Few communications links guarantee such low error rates and a customer sending financial data over a satellite link must use an ARQ protocol to ensure that the probability of an undetected bit error becomes extremely low. For example, the TCP/IP protocol used for internet access incorporates selective repeat ARQ to ensure error free transmissions.

Links operating at frequencies above 10 GHz are subject to increases in BER during propagation disturbances. These links are designed with a margin of a few decibels so that the BER falls below an acceptable level, typically 10^{-6}, for only a small percentage of any month or year. The total time for which the margin is exceeded by propagation effects will be less than 0.5% of any month in a well designed system. During the remaining 99.5% of the month, the CNR of the received signal will be well above threshold, and very low BER will result. There may, in fact, be no errors for long periods of time and billions of bits can be transmitted with complete accuracy. Under these conditions, FEC and error-detection systems do nothing for the communication system. However, unless we can detect a falling CNR directly, or an increase in BER, coding may have to be applied all the time to be certain it is available when CNR approaches threshold. To that extent, coding is an insurance against the possibility of bit errors; for most of the time it is unnecessary, but when it is needed, it proves invaluable. The error rate objective for the DVB-S2 system is the loss of no more than one packet every eleven days. This is achieved with a BER of 10^{-12} at the output of the receiver demodulator by a double layer of FEC coding. (See Chapter 10 for details.)

Common carriers, who supply communication links to users on a dial-up or leased basis, do not generally apply FEC to their links, nor do they define the protocols to be used. These are user-supplied services and must be defined by the user for the data to be sent. In such cases, the error detection and correction equipment will be located at the customers' premises, whereas the earth station may be a long distance away and accessed via terrestrial data links. The situation may be very different in a single-user network such as a direct broadcast television system, where the coding and format of the transmitted data are specified by the company that operates the uplink and satellites. A similar situation can arise in carefully controlled systems such as a military communication system where the link operator specifies the user's earth station and operating parameters in detail.

5.7.2 Channel Capacity

In any communication system operating with a noisy channel, there is an upper limit on the information capacity of the channel. Shannon examined channel capacity in mathematical terms, and his work led to significant developments in information theory and coding (Shannon 1948).

For an additive white Gaussian noise channel, the capacity H in bits per second is given by

$$H = B \log_2 \left(1 + \frac{P}{N_o B}\right) \text{ bps} \tag{5.47}$$

where B is the channel bandwidth in hertz, P is the received power in watts, and N_o is the single sided NPSD in watts per hertz.

Eq. 5.47 is commonly known as the Shannon-Hartley law. We can rewrite Eq. (5.47) specifically for a digital communication link by putting $H = 1/T_b$ where T_b is the bit duration in seconds. The energy per bit is E_b joules, giving

$$E_b = PT_b = P/H \text{ joules} \tag{5.48}$$

Then substituting $E_b/N_o = P/(HN_o)$ in Eq. (5.48) yields

$$\frac{H}{B} = \log_2\left(1 + \frac{E_b H}{N_o B}\right) \text{ bits/Hz} \tag{5.49}$$

The ratio H/B is the *spectral efficiency* of the communication link, the ratio of the bit rate to the bandwidth of the channel. Regardless of the bandwidth used, the E_b/N_o ratio cannot go below −1.6 dB (ln 2) if we are to operate at capacity without errors. This is known as the Shannon bound. It sets a lower theoretical limit on the E_b/N_o ratio we can use in any communication link. A link operating with $H < B$ is said to be *power limited* because it does not use its bandwidth efficiently. Powerful FEC schemes such as *turbo codes* (Berrou, Glavieux, and Thitimajshima 1993; Pyndiah 1998; Drury et al. 2000) and LDPC codes (Ryan and Lin 2009, pp. 201–248; Richardson et al. 2001, pp. 619–637) allow links to operate down to a CNR of 0 dB at acceptable bit error rates. This is 1.6 dB above the Shannon limit, leaving little room for further improvement.

When $H > B$, the link is said to be *bandwidth limited,* implying that we could increase capacity by using the available transmitter power in a wider bandwidth. Shannon's theory assumes essentially zero bit errors; to achieve a bit error rate of 10^{-6}, a QPSK link requires a theoretical E_b/N_o ratio of 10.6 dB with a spectral efficiency of 2 bits/Hz. Equation (5.49) predicts a bound of $E_b/N_o = 1.77$ dB for this case. What coding, in particular FEC can do for us is to improve the link performance under conditions of low CNR, such as during periods of rain attenuation, so that the BER of the link does not rise excessively. This takes us closer to the Shannon capacity in the region of low CNRs while not increasing excessively the bandwidth required for transmission. A theoretical spectral efficiency of eight bits/Hz can be achieved with $E_b/N_o = 15.0$ dB. A satellite link could use 256-QAM modulation to achieve this spectral efficiency, but would need CNR of 30.3 dB ($E_b/N_o = 21.3$ dB) to give a bit error rate of 10^{-6}. This is well below the theoretical efficiency for the Shannon bound.

5.7.3 Error Detection

Error detection coding is a technique for adding coding bits to a data stream in such a way that one or more errors in the data stream can be detected. One coding bit is added for every n data bits; this allows a single error within that *block* of n bits to be detected, whatever the number of bits in the block A simple example of an error detecting code system that has been in use for many years is the single bit parity applied to the seven bit ASCII (American standard code for information interchange) code. The ASCII code is widely used for transmission of alphanumeric data over the internet, telephone lines, and radio links.

The seven bit ASCII code consists of 128 characters that have internationally agreed interpretation and represent the alphabet in uppercase and lowercase letters, numerals 0–9, and a set of useful commands, symbols, and punctuation marks. An eighth bit, the parity bit, is used for detection of error in the seven data bits of the character. For example, in a system using *even parity,* the parity bit is 0 when the sum of the data bits is even, and 1 when the sum is odd. Thus the sum of the data bits plus the parity bit is always made even, or 0 in modulo-2 arithmetic. Figure 5.17 shows an example of even and odd parity coding. In *odd parity,* the sum of the data bits plus the parity bit is always odd, or 1 in modulo-2 arithmetic. Errors in the seven data bits, or the parity bit, are detected at the receiving end of the link by checking the eight received bits of each character for

	Data bits	Parity bits	Sum (Modulo-2)
Even parity	0 1 0 1 1 0 1	0	0
Odd parity	0 1 0 1 1 0 1	1	1

(a)

	Received codeword	Sum of bits	Error detected?
No error	0 1 0 1 1 0 1 0	0	No
One error	0 1 0 1 1 0 **0** 0	1	Yes
Two errors	0 1 **1** 1 1 0 **0** 0	0	No
Three errors	0 **0 1** 1 1 0 **1** 0	1	Yes

(b)

Figure 5.17 Example of error detection using a single parity bit with a seven bit ASCII word. (a) Even and odd parity. (b) Error detection with a single parity bit, even parity. Errored bits are underscored.

conformity with the parity rule. In modulo-2 arithmetic, $0 \oplus 0 = 0$, $0 \oplus 1 = 1$, $1 \oplus 0 = 1$, and $1 \oplus 1 = 0$. This is the exclusive OR function of digital logic. Similarly, $0 \otimes 0 = 0$, $0 \otimes 1 = 0$, $1 \otimes 0 = 0$, and $1 \otimes 1 = 1$. All the codes that we will be considering are binary, and modulo-2 arithmetic will be used throughout this section.

Suppose we have a system using even parity, which transmits the character *A* in ASCII code, as illustrated in Figure 5.17. At the receiving end of the link we check the eight bits by modulo-2 addition. If the sum is 0, we conclude that the character is correct. If the sum is 1, we detect an error. Should two bits of the character be corrupted, the modulo-2 sum is 0, so we cannot detect this condition.

We can easily calculate the improvement in error rate (assuming that we discard corrupted characters) that results from adding a parity bit to a seven bit word. For example, let the probability of a single bit error occurring in the link be p, and let us suppose that p is not greater than 10^{-1}. The probability $P_e(k)$ of k bits being in error in a block of n bits is given by the binomial probability function

$$P_e(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{5.50}$$

where p is the probability of a single bit error occurring, and

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{5.51}$$

where *n!* etc. indicates a factorial.

For example, with single parity and the seven bit ASCII characters, we have one parity bit, which allows us to detect one error, and seven data bits. Two errors cannot be detected, although three can. The most likely error that goes undetected is a two bit error. The probability that there are four or more errors in the eight bit word is much

lower than the probability of two bit errors, provided the BER is no higher than 10^{-2}, so when a single parity bit is used, the probability of an undetected error occurring in an ASCII word is approximately P_{wc} where

$$P_{wc} = P_e(2) = \binom{8}{2} {p_c}^2 (1 - p_c)^6 \tag{5.52}$$

where p_c is the single bit error probability for the eight bit word (i.e., the BER on the link). When p_c is small, $(1 - p_c)^6 \approx 1$ and

$$P_{wc} \approx \binom{8}{2} {p_c}^2 = 28 {p_c}^2 \tag{5.53}$$

If we had not used parity, the probability P_{wu} of a single error in the seven bit word with bit error probability p_u is

$$P_{wu} = \binom{7}{1} p_u (1 - p_u)^7 \approx 7 p_u \tag{5.54}$$

Thus the improvement in rate of undetected errors for the ASCII words is approximately $4p$, provided $p_c \approx p_u$.

Example 5.7

A data link transmits seven bit ASCII words at a bit rate of 1 Mbps, with a single parity bit. The probability of a bit error on the link is $p = 10^{-3}$. Find the probability of an undetected error when uncoded data is transmitted and when a single parity bit is added to each seven bit word. What is the probability of an undetected bit error when the BER on the link is 10^{-5}? How many undetected character errors would be present if a 500 page textbook were transmitted over this link, using single parity? Assume that there are an average of 330 words on a page, with a word having five letters followed by a space, giving 1980 characters per page.

Answer

Using the result in Eq. (5.54), the probability of error for uncoded seven bit words is 7×10^{-3}. If we add a single parity bit, the probability of an undetected error is given by the result of Eq. (5.53) as

$$P_{wc} \approx 28 {p_c}^2 = 28 \times 10^{-6}$$

With a BER of 10^{-5}, the undetected character error rate is 7×10^{-5} for no parity and 28×10^{-10} with single bit parity. Based on 1980 characters per page (including spaces) and 500 pages in the book, there are approximately 1 million characters in the text and 8 million bits. Without error detection the bit error rate is 7×10^{-5}, so there would be 560 typos in the text caused by transmission over the link. With single bit parity error detection the error rate is 28×10^{-10}, so the text could be transmitted 44 times before a single undetected character error occurred. At a bit rate of 10 Mbps the entire book can be transmitted in 0.7 seconds without parity and 0.8 seconds with single bit parity. The single undetected error – which will appear in the text as a single typo – would occur (statistically) after 35 seconds when single bit parity is used.

However, with single bit parity there is still a single error in one in every 44 books, consisting of a single incorrect character. The error has not been corrected, but could be flagged. For example, the word *antenna* might appear as *anten&a* because of a bit error in the sixth character of the word *antenna*. The error could be flagged by adding asterisks to each end of the word to give **anten&a** indicating that there is a known error in that word. By using the *find* capability of a word processor, all errors in the text would be easy to correct by searching for asterisks.

The above example illustrates how powerful even a single parity bit can be in the detection of bit errors in a link with low BER. Some caution is needed in making the assumption that the bit error rate is the same for uncoded and coded transmissions. When we added the single parity bit to a seven bit ASCII character, the transmission bit rate went up from seven to eight bits per character. Either it will take longer to transmit the book, or the bit rate on the link must be increased, which will result in an increase in BER because the channel bandwidth must be increased; this increases the noise power in the receiver and reduces the CNR, so not all the expected decrease in character error rate will be achieved in this case.

Example 5.7 above for parity error detection is a simple example of *block error detection*. We have transmitted our data as *blocks,* in this case eight bit blocks consisting of seven data bits and one coding parity bit. In general, we can transmit n bits in a block, made up of k message bits plus r parity bits. There are two ways in which the blocks of data can be formed. In a packet transmission system the block might have between 64 and 2048 bits, and CKSM or CRC bits are added to the block to generate an error condition at the receiver if there are errors in the block. A CRC is a sequence of bits generated by an algorithm operating on the data block. A CRC can typically determine the number of errors in a data block, whereas a CKSM can only tell whether one or more errors occurred. Alternatively, the blocks may be members of a set of n bit *codewords*. Coding schemes in which the message bits appear at the beginning of the codeword, followed by the parity bits, or parity bits followed by message bits, are called *systematic block codes*.

Transmissions that contain error control bits are specified by *code rate*. Code rate is the ratio of the uncoded bit rate to the coded bit rate, that is, code rate = k/n. If we add as many parity bits as there are message bits so that $k = n/2$, we have *half rate* coding. Other ratios such as 7/8, 3/4, and 2/3 are widely used.

5.7.4 Error Correction With Linear and Cyclic Block Codes

Linear block codes are codes in which there are 2^n possible codewords containing k message bits and $(n - k)$ coding check bits. In a typical systematic linear block code the first k bits of the codeword are the message and the remaining $(n - k)$ bits are the parity bits, or possibly the other way round with the parity bits first. A codeword with n bits of which k bits are message data is written as (n, k). Early work on single error correction linear block codes was done by Hamming at Bell Labs in the 1950s (Hamming 1950). A subset of linear block codes called *binary cyclic codes* has been developed for which implementation of error-correction logic is relatively easy. The codes can be generated using shift registers, and error detection and correction can also be achieved with shift registers and some additional logic gates. A large number of binary cyclic codes have been found, many of which have been named after the people who first proposed them. The best known are the Bose-Chaudhuri Hocquenghem codes (*BCH codes*), which were

independently proposed by three workers at about the same time in 1959–1960 (Bose and Ray-Chaudhuri 1960, pp. 68–79; Hocquenghem 1959, pp. 147–156). Other examples of block codes in widespread use are the Reed-Solomon codes, used on CDs, DVDs, and DBS-TV signals. Most of the block codes were developed in the late 1950s for error detection and correction with early computer memories, which were prone to cause data errors in the recording and recovery of data. Subsequently, the codes were applied to digital transmission of data.

Suppose we have a transmitted code word that has n bits, of which k bits are message bits and (n-k) bits are parity bits. There are 2^k valid codewords within the set of 2^n codewords. For example, if we create a (7, 4) linear block code that has four message bits and three parity bits, there are a total of $2^7 = 128$ codewords in the set, but only $2^4 = 16$ codewords are valid. If we receive an invalid codeword we know that an error has occurred on transmission, although we do not necessarily know how many bits have been corrupted, or which of the message bits are wrong. One way to correct errors in received codewords is to compare the received codeword with the valid set of codewords and select the correct codeword that is closest to the received codeword.

Some linear block codes are better for error detection or correction than others. There are some basic theorems that define the capabilities of linear block codes in terms of the *weight, distance,* and *minimum distance*. The weight, w, of a codeword C is the number of nonzero components of C (i.e., the number of logical ones). The distance, d, between two codewords C_1 and C_2 is the number of components by which they differ. The minimum distance of a block code is the smallest distance between any pairs of codewords in the entire code. With a single error detection code, the number of errors that can be detected in a code with minimum distance $d_{\min}$ is $(d_{\min} - 1)$. The number of errors that can be corrected is $\frac{1}{2}(d_{\min} - 1)$, rounded to the next lowest integer if the number is fractional. Thus it is always easier to detect errors than to correct them with linear block codes, a feature that is exploited in the Reed-Solomon codes used in CDs, DVDs, and DBS-TV systems.

There are no rules for forming block codes, so useful codes are found by inspiration and a lot of trial and error, using weight, distance and minimum distance for guidance. A good error detection code is one that detects as many bit errors as possible in a codeword of given length, preferably even when the bit errors are sequential (a *burst error*). BCH codes have good burst error correction capability, and are used with turbo coding and LDPC coding when links are operated at very low CNR, as is the case with DVB-S2 satellite television.

5.7.5 Convolutional Codes

Convolutional codes are generated by a tapped shift register and two or more modulo-2 adders wired in a feedback network. The name is given because the output is the convolution of the incoming bit stream and the bit sequence that represents the impulse response of the shift register and its feedback network (Ryan and Lin 2009, pp.147–193). A decoder for convolutional codes keeps track of the encoder's state transitions and reconstructs the input bit stream. Transmission errors are detected because they correspond to a sequence of transitions that could not have been transmitted. When an error is detected, the decoder begins to construct and keep track of all the possible tracks (sequences of state transitions) that the encoder might be transmitting. At some point, which depends on its speed and memory, the decoder selects the most probable

track and outputs the input bit sequence corresponding to that track. The other tracks that it had been carrying are discarded. One of the best algorithm used for decoding is named for A. J. Viterbi and for this reason convolutional codes are sometimes called Viterbi codes (Heller and Jacobs, 1971, pp. 268–278; Forney 1973, pp. 268–278). The improvement in bit error rate when FEC is applied to a bit stream is called *coding gain*. For example, coding gain with half rate FEC using convolutional coding with constraint length seven gives 6–7 dB of CNR improvement, so in a QPSK link the theoretical CNR for BER = 10^{-6} is typically 6.6–7.6 dB with half rate convolutional encoding, compared to a theoretical CNR of 13.6 dB without coding.

Coding gain must be used with care in satellite links, because to obtain coding gain we must increase the bit rate on the link, which results in a lower E_b/N_o because the bandwidth of the receiver has to be increased, resulting in lower CNRs and thus a higher BER. This is particularly true when we apply half rate FEC to a link. If we want to achieve the full coding gain we can send only half as many message bits. If we double the bandwidth of the link in order to send the same data rate, the CNR will fall by 3 dB and the effective increase in CNR will be only 3 dB instead of 6 dB.

5.7.6 Turbo Codes and LDPC Codes

Turbo codes and LDPC codes are the most powerful available codes for forward error correction. LDPC coding was selected for the DVB-S2 standard over turbo coding for its better error correcting performance and lower decoder complexity. However, at CNR close to 0 dB, turbo codes appear to have better performance.

The turbo code was first proposed by C. Berrou, R. Pyndiah, and P. Thitimajshima from the Ecole Nationale Superieure des Telecommunications de Bretagne (ENST), in France (Berrou et al. 1993). Turbo codes have been demonstrated that can achieve a BER of 10^{-6} with a received signal at E_b/N_o of 0.7 dB, an improvement in coding gain of 1 dB over the most powerful concatenated and interleaved convolutional codes. The ENST researchers were looking for codes that could be decoded in hardware, rather than by software, as an engineering solution that could be implemented in an integrated circuit.

The basic form of turbo code generator uses two component codes, separated by an interleaver. Message bits are read into an interleaver by row and then simultaneously read out by rows and by columns into two separate encoders that use either block coding or convolutional coding. One encoder is driven by the row message bits and the other by the column message bits from the interleaver, so that an entirely different bit sequence is applied to each encoder, but both encoders are sending the same message bits. Since the row output of the interleaver is the original data stream, one encoder has an input, which is the original message bit sequence and the other encoder input is an interleaved version of the message bits. The outputs of the two encoders are combined by multiplying the two bit sequences (modulo two). Alternatively, the two outputs can be added and sent sequentially. Turbo codes based on convolutional codes are usually known as CTCs (convolutional turbo codes) and those based on block codes as BTCs (block turbo codes).

At the receiver, the incoming symbol stream is sampled to create a *soft input* to the two decoders. A soft decoder creates a digital word from each sample using an ADC so that information about the magnitude as well as the state of the received symbol is retained. Recovered bits are given a weighting in the decoding process according to the confidence level from the sampling process. For example, in a BPSK receiver, we expect

to recover symbols with sample magnitude +V or −V volts. We have greater confidence that a symbol with a magnitude close to the expected value of +V or −V volts is correct than for a received symbol with a magnitude close to zero volts because we can guess that a significant amount of noise has been added to a received symbol with a value close to zero volts. The symbol with a value close to zero volts is much more likely to be in error. The magnitude of the received symbols is retained through the decoding process, which is one of the strengths in turbo codes. The process is known as *soft input soft output* (SISO) decoding.

The soft input symbol stream is input to a de-interleaver of the same size as the interleaver used in the transmitter and then read out by row and by column into two SISO decoders corresponding to the two encoders at the transmitter. The outputs of the decoders are two versions of the original message data, one of which was interleaved and the other direct. The outputs from the decoders are compared, using the soft output values to apply confidence levels to the decoded bits. The decoding process is then repeated to obtain a better estimate of the original transmitted data. The characteristics of the encoding schemes and the decoding methods is such that repeated processing of the interleaver output through the decoders reduces the number of errors remaining in the recovered data, thus improving the BER. The disadvantage of this approach is that many iterations in the decoders creates latency in the bit stream and the soft decoder must run at a clock rate many times higher than the bit rate of the message data. The overall performance of turbo codes can be improved by adding an outer layer of Reed-Solomon encoding, just as with concatenated convolutional encoding.

Much research has been carried out to optimize the two codes used by a turbo encoder and the soft decoding process at the receiver. The aim is to achieve the highest coding gain with the smallest latency and the least complex decoder. With lower speed bit streams such as compressed digital speech, it is possible to use fast DSP integrated circuits to perform the decoding process, making turbo coding attractive for cellular telephone systems. The additional coding gain achieved by turbo codes and the possibility of operating at CNRs as low as 0 dB makes turbo coding a good choice for a fading radio channel. With higher bit rates FPGAs are needed to perform parallel processing. Commercial turbo code products in the form of coder and decoder boards became available in the late 1990s, and are now available as single chip codecs.

The Jet Propulsion Lab (JPL) in the United States has developed turbo codes for use on its deep space research spacecraft (Andrews et al. 2007, pp. 2142–2156). The signals from spacecraft at interplanetary distances are very weak, so powerful error correction coding at very low received CNRs is essential. The turbo code implemented for the Voyager spacecraft uses a 16 384 bit interleaver and 10 iterations in the decoder giving a half rate code with a BER of 10^{-5} at a received SNR of 0.7 dB. Given Shannon's limit of −1.6 dB for error free recovery of data, turbo codes are approaching this theoretical performance limit.

5.7.7 Implementation of Error Detection on Satellite Links

Error detection is invariably a user-defined service, forming part of the operating protocol of a communication system in which the earth-satellite-earth segment may only be a part. It allows the user to send and receive data with a greatly reduced probability of error, and a very high probability that uncorrected errors are identified and located within a block of data, so that the existence of an error is known even if the exact bit or

word in error cannot be determined. The penalty for the user is a reduced transmission rate, just as in FEC. However, the transmission rate is reduced only when errors occur, so there is no bit rate penalty under normal conditions when no errors occur, unlike FEC. Error detection is readily accomplished using the coding techniques described in the previous sections. In many communications systems it is not sufficient simply to detect an error; it must be corrected. However, some systems such as speech and picture transmission may simply count errors to determine whether the link is above an operating threshold. Excessive error counts may cause the link to automatically shut down. When a succession of pictures is transmitted, for example weather maps, the previous value of a pixel can be used when a new value is in error. When an error-detection code or CRC is used and the error must be corrected, a retransmission of the block of data containing the error must be made so that the correct data are acquired at the receiving terminal. If an error is detected in the block, a *not acknowledge* (*NAK*) signal is sent back to the transmit end of the link, which triggers a retransmission of the erroneous block of data. This is called an ARQ system.

In terrestrial data communication systems, it is common practice for the receiving terminal to send an *acknowledge* (ACK) signal to the transmitting end whenever it receives an error-free block of data. Such protocols works well on terrestrial data links with short time delays, but the long transmission delays in GEO satellite communication systems make it highly undesirable to send ACK signals for every error free block that is received, so internet access via GEO satellites must use a different access protocol. Often a satellite terminal receives data using a terrestrial protocol such as TCP/IP and generates the acknowledgments, then transmits the data over the satellite link using a different protocol.

There are three basic techniques for retransmission requests, depending on the type of link used. In a one-way, *simplex link,* the ACK or NAK signal must travel on the same path as the data, so the transmitter must stop after each block and wait for the receiver to send back a NAK or ACK before it retransmits the last data block or sends the next one. This is called *stop and wait* ARQ. With a one-way delay of 240 ms on a GEO satellite link, the data rate of such a system will be very low as it is possible to send only two blocks of data per second. Satellite links usually establish two-way communication (duplex channel). The ACK and NAK signals are sent on the return channel while data are sent on the go channel. However, if the data rate is high, the acknowledgment will arrive long after the block to which it relates was transmitted.

In a *continuous transmission* system using the *go-back-N* technique, data are sent in blocks continuously and held in a buffer at the receive end of the link. Each data block is checked for errors as it arrives, and the appropriate ACK or NAK is sent back to the transmitting end, with the block number appended. When a NAK(N) is received, the transmit end goes back to block N and retransmits all subsequent blocks. This requires the transmitter on a GEO satellite link to hold at least 480 ms of data, to allow time for the data to reach the receive end and be checked for errors and for the acknowledgment to be sent back to the transmit end. Since there is a delay in transmission only when a NAK is received, the *throughput* on this system is much greater than with the stop-and-wait method. Throughput is the ratio of the number of bits sent in a given time to the number that could theoretically be sent over an ideal link.

If sufficient buffering is provided at both ends of the link, only the corrupted block need be retransmitted. This system is called *selective repeat* ARQ. Figure 5.18 illustrates the principle of selective repeat ARQ. Block 3 is corrupted in transit resulting in the

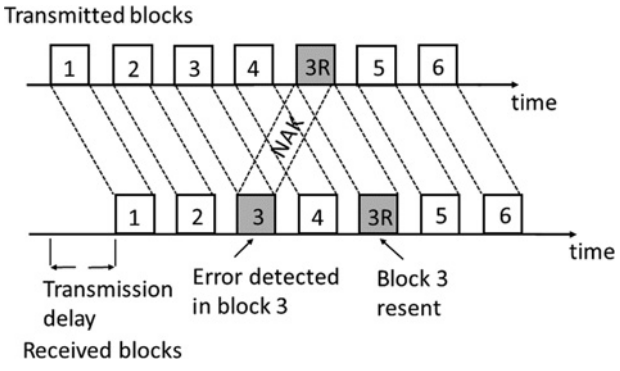


Figure 5.18 Selective repeat ARQ. An error is detected at the receiver in block 3. A not acknowledge signal (NAK) is sent by the receiver to the transmitter requesting a resend of block 3. A buffer is needed at both ends of the link to ensure blocks are sent from the receiver in the correct order, adding latency to the transmissions.

receiver sending a NAK(3) message to the transmitter, but block 4 is transmitted before the NAK(3) message is received. On receiving a correct version of block 3, the receiver buffer substitutes it for the corrupted version and releases the data for retransmission. In systems handling data rates of megabits per second, the buffer requirements for continuous transmission systems become quite large because of the delay on the link via a geostationary satellite. The internet protocol TCP/IP uses selective repeat of blocks that contain errors. The relatively short delays on terrestrial paths used for internet traffic allow the use of selective repeat ARQ. The TCP/IP protocol cannot be used over a geostationary link without modification because it was designed for terrestrial links with short delays and sets a maximum wait time of 60 ms. The protocol times out before repeat transmissions arrive and no replies are received when a GEO satellite link is used. The main disadvantage of go back N and selective repeat ARQ is that they add *latency* to the transmission of data. In real time applications such as gaming and stock trading these delays may make satellite links unattractive.

Figure 5.18 illustrates how selective repeat ARQ operates. Blocks of data are numbered and sent to the receiver, where they arrive with a delay. Blocks are checked for errors at the receiver. No action is taken if a block is received correctly, but when an error is detected in a block, a NAK message is sent back to the transmitter with the number of the block that is in error. The transmitter stops the transmit sequence, retransmits the errored block, and then continues with the sequence of blocks. In a GEO satellite link with a round trip delay of 500 ms, many blocks will have been transmitted before the NAK message arrives at the transmitter. At least 500 ms of buffering is required at both ends of the link to allow resequencing of the received data, which adds significant latency the delivery of data.

Example 5.8

Calculate the frequency of retransmission, throughput, and buffer requirements of a satellite link capable of carry data at rates of (a) 24 kbps and (b) 1 Mbps when a block length of 127 bits is used and the one-way path delay is 240 ms, for a bit error rate of 10^{-4} and a double error detecting code (127, 120), using the following ARQ schemes.

1. Stop and wait.
2. Continuous transmission with transmit buffer only (go-back-N).
3. Continuous transmission with buffers at both ends of the link (selective repeat).

Comment on the three ARQ schemes and their suitability for a GEO satellite communication link.

Answer
For a 127-bit code block, the probability of one or two errors is given by Eq. (5.50)

$$P_{\mathrm{e}}(k) = \binom{n}{k} p^{\mathrm{k}}(1-p)^{n-\mathrm{k}}$$

where $k = 2$, $n = 127$, and p is the probability of a single bit error, which is 10^{-4} in this example. The probability of one or two errors being detected in a block of 127 bits is

$$P(\text{one or two errors}) = 127 \times 0.9875 \times 10^{-4} + \frac{1}{2}(127 \times 126) \times 0.9876 \times 10^{-8} = 0.01262$$

Thus, on average, one in every 79 received blocks has a detectable error.

1. Stop-and-wait.
   We send 127 bits and wait for an acknowledgment, which takes 0.485 seconds at 24 kbps and 0.4801 seconds at 1 Mbps. We therefore send only about two blocks each second at either data rate, since the time is dominated by waiting for an acknowledgment. After 79 blocks have been sent, on average, we detect an error; that is after about 39.5 seconds. The average data rate on the link is approximately 254 bps for both transmission bit rates.
2. Go-back-N.
   a. The time required to send 79 blocks of 127 bits at 24 kbps is 0.418 seconds.
      The 79th block has a detected error: while the NAK signal goes back to the send end to request a retransmission of block 79, a further 91 blocks arrive. These are discarded when the retransmission arrives 0.48 seconds later, starting at block 79. This slows the throughput by about 54%, to about 11.2 kbps. The buffer at the transmit end must hold 0.48 seconds worth of bits, approximately 11 600 bits.
   b. At a bit rate of 1 Mbps, the time to send 79 blocks is 0.01 seconds. We then wait 0.48 seconds for the retransmission. The average data rate is then 20.46 kbps. The transmit buffer must hold 490 kbits of data.
3. Selective repeat system.
   a. The only time lost in a selective repeat system is in the retransmission of blocks which have errors. On average, one block in every 79 has to be retransmitted, so the rate efficiency of the system is 78/79, or 98.73%. At 24 kbps, the average data rate is 23.70 kbps, and the buffer needed is 11 600 bits.
   b. At 1 Mbps, the average data rate is 987.3 kbps, and the buffer must hold 490 kbits. Clearly, selective repeat ARQ has far better throughput than either of the other two techniques.

5.7.8 Concatenated Coding and Interleaving

Sophisticated error correction and detection systems are used on some satellite links to overcome burst errors and the effects of low CNRs. Burst errors can occur when the signal is temporarily blocked, or the CNR has become unusually low and the BER is very high. Forward error correction codes are limited in the number of sequential bit errors that can be corrected, so long strings of incorrect bits will not be corrected by the FEC

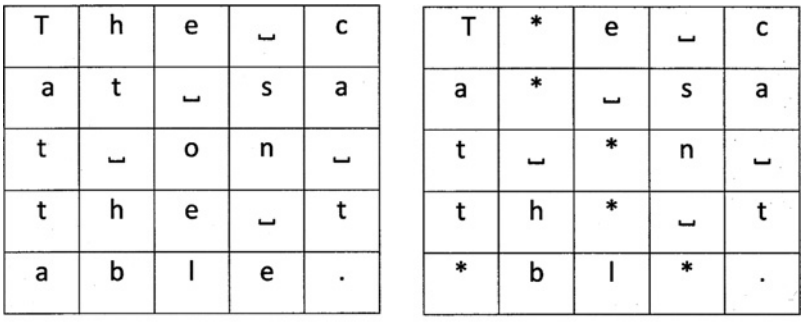


Figure 5.19 Illustration of interleaving. Letters are used in this example to show how the interleaver works. In practice, data entries are ones and zeroes. (a) Data are read into the transmit end interleaver by row and out by column. Spaces are represented here by the symbol (␣) (b) At the receiving end, the received data are read into the de-interleaver by column and out by row. Errors are shown by asterisks (*).

decoder in the receiver. Burst errors can be overcome by using *interleaving*. Interleaving is usually applied ahead of error control coding where a single coding operation is used. When two separate error control coding operations are placed in series, the process is called c*oncatenation*. Two coding operations can be applied in sequence to a data stream to decrease the probability of undetected errors occurring because of low CNR on a satellite link. Concatenated coding can achieve coding gains up to 9 dB. With concatenated coding, the interleaver may be placed between the two stages of error control coding. This is the strategy used in audio compact disk recordings, video DVDs and also in direct broadcast satellite TV signals using the DVB-S and DVB-S2 standards in satellite TV systems (Haykin and Moher 2009, pp. 208–212).

The purpose of interleaving is to spread out the errors that occurred in a burst and thus to make it easier for an error correction system to recover the original data. Figure 5.19 shows a simple interleaver with 5 rows and 5 columns. Bits are read in to the rows and read out by columns, which spreads out the bits in the resulting bit stream. Interleaving can be illustrated more easily using letters of the alphabet rather than binary data.

Consider the following message: *The␣cat␣sat␣on␣the␣table.* (Space is inicated by the symbol ␣) Let us suppose that we transmit the message with single bit error detection coding, and that two and three bit burst errors occurs so that this message, without interleaving, is received as: *The␣***␣sat␣**␣the␣*able*. Where the * indicates an error. Because the English language is highly redundant, and only certain letter combination make valid English words, we can guess at the message based on what we received, although we might guess *dog* or *cup* instead of *cat*, or *cable* instead of *table*. Now consider what happens when we use the 5×5 interleaver at the transmitter, as shown in Figure 5.19.

The message is read into the interleaver by row, as shown. The message is read out by column to give: *Tattaht␣hbe␣oel␣sn␣eca␣t*. When this message is sent over the satellite link, it is received with bit errors in the same positions as previously. We receive the following message:

*Tatt***␣hbe␣**l␣s␣*ca␣t*. The message is read into the de-interleaver at the receiver by column, and read out by row. The message at the receiver is then:

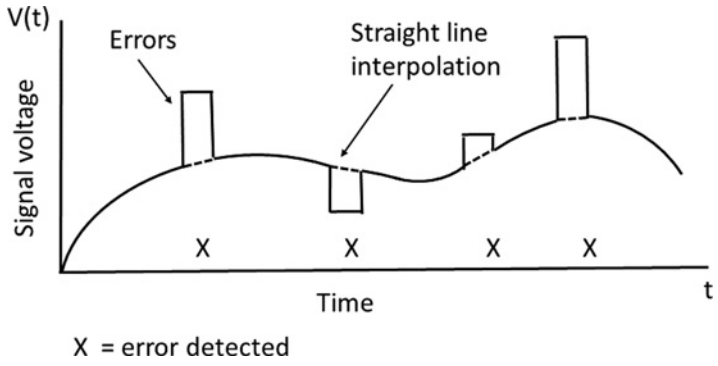


Figure 5.20 Illustration of interpolation to fill in missing data in an analog waveform. Errors are detected in digital words at the times marked by X. Interpolation between the words on either side of the error is used to reconstruct the analog waveform.

*T*e␣ca*␣sat␣*n␣th*␣t*bl**. We could correct the message with a single character error correction FEC code, or we could make an improved guess at the true message based on the received message which now has no burst errors. Since there are no burst errors here, a single (character) error correction code would recover the original message correctly. The high level of redundancy in the English language ensures that the message: *T*e␣ca*␣sat␣*n␣th*␣t*bl**. can be read with little error – although we cannot be certain whether *ca** is *cab*, *cat*, *cam*, *can*, *cap*, or *car*. However, the rest of the message is obvious after de-interleaving. There is no such redundancy in a stream of bits. Ones and zeroes are equally likely; indeed the bit stream is usually scrambled to make certain that this is the case, so we cannot guess at the correct bit or byte when an error occurs.

Interleavers used in digital communication systems work on the same principle at the simple example shown in Figure 5.19 to break up burst errors and make possible correction of the received data. A key element in the digital transmission of analog data is that *analog interpolation* can be used to reconstruct an analog waveform containing single word errors. In the example above using letters, interpolation between the correctly received characters is much easier when the errors occur singly. We can make a much better guess at the message received after interleaving than when there is no interleaving. The same principle holds true for binary transmission of analog data. Interpolation of an analog waveform is simple when only one sample (a digital word) is missing. Figure 5.20 shows an example of interpolation in an analog waveform where errors have been detected but not corrected. The locations of errors are marked with an X. Straight line interpolation introduces a small amount of distortion into the analog waveform.

This approach is used very successfully on audio compact disks, DVDs, and in digital video broadcast transmissions. Compact disks store audio data bits as changes in the reflectivity of the disk. The bits are about one micron long, so a scratch or a speck of dust on the surface of the CD will cause a lengthy burst error. A long interleaver is used to spread out the burst error and this makes reconstruction of the analog waveform possible for burst errors as long as a thousand bits.

Figure 5.21 shows the coding and decoding structure of DBS-TV signals conforming to the DVB-S and DVB-S2 standards. In the DVB-S version, shown in Figure 5.21a, the digital signal is first encoded using a (204/188) Reed-Solomon code, then interleaved with a 17×11 interleaver with one byte per cell, and finally encoded with a rate three

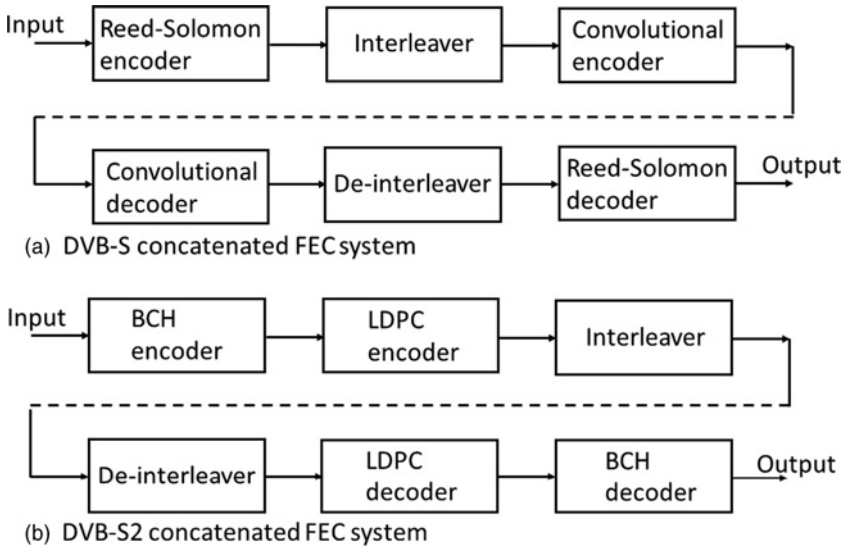


Figure 5.21 Concatenated FEC structure used in the DVB-S and DVB-S2 satellite television standards. (a) Structure for DVB-S standard. (b) Structure for DVB-S2 standard with 8-PSK modulation. Interleaving is not used in all transmissions. Reed-Solomon: (204,188) error detection code, BCH: Bose-Chaudhury Hoquenghem burst error correcting code, LDPC: Low density parity check code.

quarters FEC code, using a convolutional code. At the receiver, the convolutional code is used to correct some errors, which improves the BER, but when the CNR in the receiver is low not all the errors will be corrected. The signal is de-interleaved, which spreads out any uncorrected errors so that the probability of a burst error longer than a couple of bits becomes small. The Reed-Solomon code is then used to correct as many errors as possible and then to detect the location of the remaining errors and flag video words with errors. The video decoder looks at the digital words on each side of a word that has been flagged with an error and replaces the flagged word with a new word that is found by interpolation between the two neighboring words. The resulting error in the analog video waveform is small. Thus the Reed-Solomon coding, which has a high code rate and uses both error correction and error detection properties allows errors in the video signal to be corrected. The coding gain is typically 6–7 dB with BER $= 10^{-6}$ at the output of a convolutional decoder, and very few errors are left in the analog waveform after Reed-Solomon decoding and analog interpolation. Greater coding gain can be achieved using a combination of BCH and low density parity code (LPDC) codes as in the DVB-S2 standard.

In the DVB-S2 standard, the FEC codes employed and their order is different, as illustrated in Figure 5.21b. The data are first encoded using an inner LDPC code, which has good performance with very low CNR signals, and then encoded again with a BCH code, which has good burst error correction capability. Interleaving is not used when QPSK modulated signals are transmitted, but is applied with 8-PSK modulation. The bit stream is than interleaved with a much longer interleaver than is used in the DVB-S standard. At the receiver, the de-interleaver spreads out burst errors, the LDPC decoder corrects most of the errors, and the BCH decoder corrects or locates any remaining burst errors. Chapter 10 contains more details on the DVB-S and DVB-S2 standards.

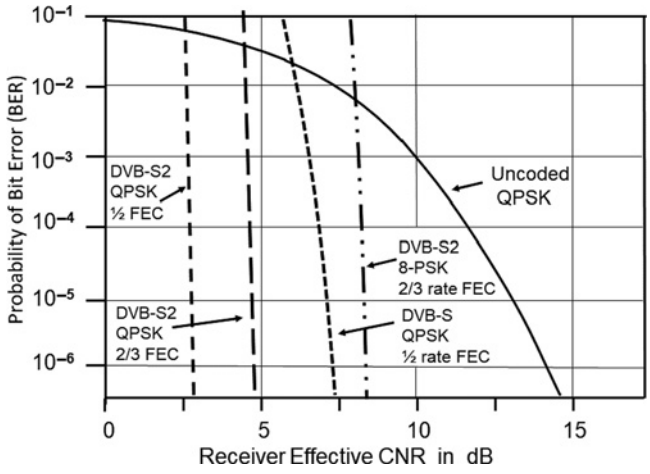


Figure 5.22 BER performance with DVB-S and DVB-S2 standard direct broadcast TV transmissions. A 1.6 dB implementation margin is included in the effective CNR. Before the implementation margin is applied, the performance of the DVB-S2 links is close to 1 dB above the Shannon limit. The ½FEC and ⅔FEC (or ½ rate FEC and ⅔ rate FEC) indicate the forward error correction coding applied to the data stream.

One of the features of the DBS-TV systems is that the software in the customers' receivers can be modified via the satellite. Packets that are tagged as software can be sent to all receivers, and changes made to the signal processing. This allows the system to change the coding and decoding methods to implement improved error control strategies. Television program transmission has to be stopped while software upgrades are made, so software changes are usually made in the early hours of the morning following a warning message that all TV programs will be suspended.

The performance of direct broadcast satellite TV links using the DVB-S and DVB-S2 standards that incorporate sophisticated error correction techniques is illustrated in Figure 5.22, with a 1.8 dB implementation margin applied. Comparison with the uncoded QPSK case shows that coding gain is about 6 dB for DVB-S signals and 9 dB for DVB-S2 signals at an output bit rate of 10^{-6}. The design objective for DVB-S2 links is to achieve an output BER below 10^{-11} when the BER at the demodulator output is between 10^{-2} and 10^{-1}. This is called the *quasi error free* (QEF) objective, and results in one video frame in 10^7 having errors. With MPEG compression a single bit error in a transmitted frame can cause multiple pixel errors in the display. Note that the near vertical lines for DVB-S2 signals mean that the output of the receiver collapses for a very small change in received signal CNR – typically less than 1 dB. If rain with sufficient attenuation moves into the signal path, the picture on the TV set will suddenly vanish, usually replaced by a text that states "Searching for satellite signal." See Chapter 10 for further details of the DVB-S standards and how this impressive performance is obtained.

5.8 Summary

Digitized analog signals may conveniently share a channel with digital data, allowing a link to carry a varying mix of voice and data traffic. While baseband digital signals

are often visualized as rectangular voltage pulses, careful pulse shaping is required to prevent intersymbol interference (ISI) and to permit reasonably distortionless transmission through the limited bandwidth of a transponder. SRRC filters are used in the design of digital radio links to create the required zero ISI waveforms at the receiver. The ideal SRRC filter does not exist, so real filters which approximate the SRRC filter shape must be used, or an approximation to the ideal SRRC waveform must be generated, with consequent non-zero ISI at the receiver output. Equalizers must also be used in digital transmitters and receivers to compensate for changes in signal spectra that can cause ISI.

SRRC filters can be implemented in digital form, in the time domain, using FIR or IIR digital filters. FIR filters are preferred in radio systems. SRRC waveforms can be created digitally in a transmitter and digital modulations can also be implemented digitally. Software radio receivers use DSP to implement each of the blocks of a conventional hardware receiver. All of these techniques must use truncated forms of the SRRC waveform and limited structures in digital filters, which leads to some ISI at the receiver and ensures that an implementation margin must be allowed when calculating satellite link performance.

The common digital modulation schemes used on digital satellite links are QPSK and m-PSK or m-APSK. In these modulations, an incoming data stream sets the phase of a sinusoidal carrier to two (BPSK), four (QPSK) or m values. The performance of a digital link is described by its bit error rate. Digital links are designed to meet bit error rate requirements in the same way as analog links are designed to deliver minimum SNR values.

Analog signals must be digitized for transmission over a digital link. This involves sampling the signal at a rate that is at least twice the highest frequency present and converting the sample values to digital words. Standard practice with telephone channels is to use non-uniform quantization with a sampling rate of 8 kHz, and to transmit eight bit words giving a serial transmit bit rate of 64 kbps. This system for transmitting digital speech is often known by the old name of PCM. Adaptive differential pulse code modulation (ADPCM) allows the speech to be transmitted at 32 or 16 kbps. LPC can further reduce the bit rate of digital speech signals to 4.8 kbps and is used in some LEO satellite telephone systems. Television signals are transmitted digitally using MPEG-2 and MPEG-4 compression techniques to reduce the bit rate to a manageable level.

Digital data is sent in packets that contain additional bits to identify the start of the packet, addresses for the source and destination of the packet, control information that identifies the type of packet and its contents, and a CKSM or CRC. Protocols describe the way in which a two-way system connects to other users; the best known is TCP/IP, which is used to connect to the internet. TCP/IP cannot be used over GEO satellite links because the round trip delay ($\geq$480 ms) exceeds the 60 ms limit set by the protocol. A different protocol is used for the satellite portion of the link and TCP/IP transmission is restored at the receiving earth station.

The transmission of data over a satellite communication link is likely to result in errors occurring in the received data, for at least a small percentage of time, because of noise added by the transmission system. Many links guarantee only 10^{-6} bit error rate and may not achieve this accuracy during periods of rain or other propagation disturbances. Bit errors contribute to the baseband SNR when digital speech is sent, but it is rarely necessary to correct bit errors in speech; the listener can make such corrections because

of redundancy in speech and for error rates up to 10^{-4} speech remains intelligible. When data are sent over a link, the receiving terminal does not know in advance what form the data take and can only detect or correct errors if extra coding bits are added to the transmitted data.

Coding of data provides a means of detecting errors at the receiving terminal. Error detecting codes allow the presence of one or more errors in a block of data bits to be detected. Error correcting codes allow the receiving terminal equipment to locate and correct a limited number of errors in a block of data. When error detection is employed, a retransmission scheme is often needed so that the data block can be sent again when it is found to be in error. The long round trip delay ($\geq$480 ms) in a satellite link makes simple stop-and-wait systems unattractive for real time data transmission. Throughput can be increased by providing data storage at both ends of the link and using selective repeat ARQ in which corrupted data blocks are retransmitted by interleaving them with subsequent data block transmissions. However, latency is increased because buffering is required at both ends of the link.

Forward error correction (FEC) provides a means of both detecting and correcting errors at the receiving terminal without retransmission of data. FEC codes add parity check bits to the data bits in a way that allows errors to be located within the received data stream. In general, twice as many errors can be detected by an FEC code as can be corrected. FEC has the advantage over error detection that a single unit at each end of the link (a codec) can insert and remove the FEC code bits, and make corrections as required. FEC is used on all satellite links where the CNR at the receiver is likely to be low. This includes satellite telephones, VSAT terminals, and DBS-TV.

Interference tends to cause burst errors in which many sequential bits are corrupted. Special burst-error correction codes are available with the capability of correcting errors in a number of adjacent bits. Scrambling and interleaving of data bits, and interpolation of analog waveforms are other ways in which the effect of burst errors can be reduced.

Exercises

5.1 A Ku-band satellite uplink has a carrier frequency of 14.0 MHz and carries a symbol stream at $R_s = 20$ Msps. The transmitter and receiver have SRRC filters with $\alpha = 0.20$.
- **a.** What is bandwidth occupied by the RF signal?
- **b.** What is the frequency range of the transmitted RF signal?
- **c.** Modulation is QPSK with 2/3 rate FEC coding. What is the bit rate of the signal?

5.2 A satellite transponder has a bandwidth of 36 MHz. A group of earth stations has ideal SRRC filters with $\alpha = 0.25$. What is the maximum bit rate that can be sent through this transponder with
- **a.** BPSK modulation
- **b.** QPSK modulation
- **c.** 8-PSK modulation?

5.3 A data stream at 100 Msps is to be sent via a satellite using QPSK. The receiver IF frequency is 200 MHz. Find the RF bandwidth needed to transmit the QPSK

signal when ideal SRRC filters with $\alpha = 0.25$ are used. What is the bit rate of the signal with QPSK modulation and FEC coding rates of

a. ¾
b. ⅔
c. ½?

5.4 Repeat Question #4 for 8-PSK modulation.

5.5 A satellite link has overall CNR in the receiver under clear sky conditions of 15.0 dB. The transmitter and receiver have ideal SRRC filters with a noise bandwidth of 1.0 MHz and a roll off factor of 0.25. The link has an implementation margin of 0.5 dB. FEC encoding is not applied to the transmitted signal.
What are the bit rate, symbol rate, occupied bandwidth of the link, and BER when the link is operated with QPSK modulation?
If rain attenuation on the link causes the overall CNR of the received signal to fall by 3 dB, what is the new BER value?

5.6 A satellite link uses a bandwidth of 5 MHz in a 52 MHz wide Ku-band transponder. The transmitter and receiver have SRRC bandpass filters with roll-off factor $\alpha = 0.2$. The overall $(CNR)_o$ ratio for the carrier in the receiver is 11.0 dB in clear air, falling below 8.0 dB for 0.1% of an average year. The transmitter and the receiver have both 8-PSK and QPSK modulators and demodulators. The implementation margin with BPSK modulation is 0.2 dB, and for QPSK modulation is 0.4 dB. FEC coding produces a coding gain in the receiver of 6.0 dB.
Determine the bit rate that can be sent through the link with BPSK, and with QPSK. Find the bit error rate for each modulation in clear air conditions and for the 0.1% of the year conditions. Which modulation would you recommend for this system?

5.7 A data link transmits at a bit rate of 1 Mbps using a FEC code that can correct one error and two errors, but not three or more errors. The probability of a bit error on the link is $p = 10^{-3}$.

a. Find the probability of an undetected error.
b. What is the probability of an undetected bit error when the BER on the link is 10^{-5}?
c. How many bits can be sent, on average, before an undetected error occurs when the bit error rate is
   i) 10^{-3}
   ii) 10^{-5}?

5.8 A group of 50 VSAT stations operate in a star network with a central gateway station using FDMA. Each VSAT station has transmitter with an effective isotropically radiated power (EIRP) of 42 dBW. Signals transmitted by the VSAT stations (the inbound link) have a bit rate of 64 kbps with QPSK modulation and half rate FEC encoding. At the gateway station, the overall CRN ratio for a single VSAT channel is 17.0 dB in clear sky conditions. The outbound link is a TDM data stream at a bit rate of 3.2 Mbps, with QPSK modulation and half rate FEC encoding. The CNR ratio for the outbound TDM signal in a VSAT station receiver

is 18.0 dB. The implementation margins for the inbound and outbound links is 1.0 dB.

The stations share the transponder using FDMA, with 8 kHz guard bands between the edges of the RF signals. The SRRC filters used in all links have $\alpha = 0.35$. To minimize intermodulation between signals, the transponder is operated with 3 dB output back off.

a. Calculate the RF bandwidth occupied by one VSAT transmission, and the RF bandwidth occupied by the gateway station transmission.

b. The inbound and outbound links share the same transponder with a guard band between the links of 250 kHz, what is the total bandwidth needed in the transponder?

c. After FEC decoding is applied in the receivers of both the inbound and outbound links, the threshold for a BER of 10^{-6} is 10.6 dB, corresponding to a coding gain of 6.0 dB. What is the BER at the gateway station receiver output for one VSAT signal in clear sky conditions? What is the BER at a VSAT station receiver output in clear sky conditions?

d. What is the average time between bit errors at the VSAT receiver output and at the gateway receiver output when the BER is 10^{-6}?

e. The time between bit errors in any link must not be less than 100 hours. Calculate the threshold BER for each link (inbound and outbound) and estimate the overall CNR required at the VSAT receiver output and at the gateway receiver output to meet this requirement.

f. A second group of 50 VSAT stations is added to the star network. The outbound TDM data rate is increased to 6.4 Mbps, but no other changes are made to the system parameters. What are the new overall CNR ratios in the VSAT receiver and the gateway receiver for the conditions you found in part (e)?

5.9 What are the advantages of software radio design compared to radios built from hardware components? Why are software radios important when handheld two-way radios are needed?

What is the difference between a software radio that digitizes an IF signal and a direct conversion digital radio that processes baseband signals?

5.10 Write a software program or create a spread sheet that calculates the waveform at the output of an SRRC filter with $\alpha = 0.3$ from Eq. (5.18). Use ten samples per bit of the signal bit stream and truncate the waveform to six periods.

Generate the SRRC filter output for the bit sequence 10101111010. How does this waveform differ from the waveform you would expect with ideal SRRC filters?

5.11 Explain the cause of quantization noise in a PCM voice link. How does quantization noise differ from thermal noise when heard over a telephone link? What is a MOS for compressed digital speech? Why is a MOS used to quantify performance of a digital telephone rather than quantization noise $(CNR)_q$?

5.12 BCH and Reed-Solomon block coding is used in many satellite links. What are the advantages of using these codes rather than simpler Hamming block codes or more complex convolutional codes?

5.13 Why are interleavers used in some digital satellite communication links and also in CD optical disc recordings?

References

Andrews, K.S. et al. (2007). The development of turbo and LDPC codes for deep-space applications. *Proceedings of the IEEE* 95: 2142–2156.

Antoniou, A. (2018). *Digital Filters – Analysis, Design, and Signal Processing Applications*. New York, NY: McGraw-Hill.

Baudot, E. (2018). https://en.wikipedia.org/wiki/%C3%89mile_Baudot (accessed 16 February 2018).

Bennett, W.R. (1983). Secret telephony as a historical example of spread-Spectrum communications. *IEEE Transactions on Communications* COM-31 (1): 99.

Bennett. (2009). SIGSALY http://www.nsa.gov/about/cryptologic-heritage/historical.../sigsaly-start-digital.shtml (accessed 22 May 2018).

Berrou, C., Glavious, A., and Thitimajshima, P. (1993). *Near Shannon Limit Error-Correcting and Decoding: Turbo Codes*, 1064–1070. ICC.

Bose, R.C. and Ray-Chaudhuri, D.K. (1960). On a class of error correcting binary group codes. *Information Control* 3: 68–79.

Copeland, J.B. (ed.) (2006). *Colossus: The secrets of Bletchley Park's code-breaking computers*. Oxford, UK: Oxford University Press.

Couch, L.W. (2007). *Digital and Analog Communication Systems*, 7e. Upper Saddle River, NJ: Pearson Education Inc.

Davidoff, M. (ed.) (2000). *Amateur Satellite Handbook*. Newington, CT: American Radio Relay League.

Drury, G., Marhavian, G., and Pichavano, P. (2000). *Coding and Modulation for Digital TV*. Dordrecht, NL: Kleuwer Publishing Co.

European Television Standards Institute (2009). Digital Video Broadcasting DVB-S2. ETSI EN 302 307 V1.2.1 (2009–08).

Forney, F. (1973). The Viterbi Algorithm. *Proceeding of the IEEE*, vol. 61, 268–278.

Gannon, P. (2006). *Colossus: Bletchley Park's Last Secret*. London, UK: Atlantic Books.

Hamming, R.W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal* 29 (2): 147–160.

Haykin, S.S. (2001). *Digital Communications*, 4e. Hoboken, NJ: Wiley.

Haykin, S.S. and Moher, M. (2005). *Modern Wireless Communications*. Upper Saddle River, NJ: Pearson Education Inc.

Haykin, S.S. and Moher, M. (2009). *Communication Systems*, 5e. New York, NY: Wiley.

Heller, J. and Jacobs, I. (1971). Viterbi decoding for satellite and space communication. In: *IEEE Trans on Communications*, vol. 19, 835–848.

Hocquenhem, A. (1959). Codes Corecteurs d'Erreur. In: *Chiffre*, vol. 2, 147–156.

IET (1979). History of pulse code modulation. *Proceedings of the IEEE* 126 (9): 889–892.

Kahn, D. (1991). *Seizing the Enigma: The Race to Break the German U-Boat Codes*, 1933–1945. Annapolis, MD: Naval Institute Press.

Krouk, E. and Seminov, S. (eds.) (2011). *Modulation and Coding techniques in Wireless Communications*. Chichester, UK: Wiley.

Lathi, B.P. and Ding, Z. (2009). *Modern Digital and Analog Communication Systems*, 4e. Oxford, UK: Oxford University Press.

Lin, S. and Costello, D.J. Jr. (1983). *Error Control Coding: Fundamentals and Applications*. NJ, Prentice-Hall: Englewood Cliffs.

Microsoft Support (2017). https://support.microsoft.com/en-us/help/103884/the-osi-model-s-seven-layers-defined-and-functions-explained (accessed 8 February 2018).

Miller R. and Badgley B. (1943). N-ary Pulse Code Modulation. U.S. Patent 3,912,868 filed 1943, Awarded 1976.

Nyquist, H. (1924). Certain factors affecting telegraph speed. *Journal of the A. I. E. E.* 43: 124. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7505.1924.tb01361.x (accessed 27 June 2018).

Nyquist, H. (1928). Certain topics in telegraph transmission theory. *AIEE Transactions* 47 (2): 617–644.

Optical fiber (2018). https://en.wikipedia.org/wiki/Optical_Carrier_transmission_rates. (accessed 7 June 2018).

Pratt, T. and Bostian, C.W. (1986). *Satellite Communications*. New York, NY: Wiley.

Pratt, T., Bostian, C.W., and Allnutt, J.E. (2003). *Satellite Communications*, 2e. Hoboken, NJ: Wiley.

Pyndiah, R.M. (1998). Near-optimum decoding of product codes: Block turbo codes. *IEEE Transactions on Communications* 46 (8): 1003–1010.

Rappaport, T.S. (2002). *Wireless Communications*, 2e. Upper Saddle River, NJ: Prentice Hall.

Reed, J.H. (2002). *Software Radio-A Modern Approach to Radio Engineering*. Upper Saddle River, NJ: Prentice Hall.

Richardson, T.J. and Urbanke, R.L. (2001). Design of capacity-approaching irregular low-density parity-check codes. *IEEE Transactions on Information Theory* 47: 638–656.

Ryan, W. and Lin, S. (2009). *Channel Codes: Classical and Modern*. Cambridge, UK: Cambridge University Press.

Shanmugam, S. (1979). *Digital and Analog Communication Systems*. New York, NY: Wiley.

Shannon, C.E. (1948). A mathematic theory of communications. *Bell System Technical Journal* (Part 1): 379–423; Part 11, pp. 623–656.

SIGSALY (n.d.). http://www.nsa.gov/about/cryptologic-heritage/historical-figures-publications/publications/wwii/sigsaly-story (accessed 10 February 2018).

Simulink® (2019). https://www.mathworks.com/products/simulink.html (accessed 10 January 2019).

Smith, M. (2011). *The Secrets of Station X: How the Bletchley Park Codebreakers Helped Win the War*. London, UK: Biteback Publishing.

Stremler, F.G. (1999). *Introduction to Communication Systems*, 3e. Upper Saddle River, NJ: Pearson Education Inc.

Taylor, F.J. (2011). *Digital Filters – Principles and Applications with MatLab*. Holboken, NJ: Wiley.

Tektronix white paper (Mean Opinion Score Algorithms for Speech Quality Evaluation) (2009). https://www.tek.com/document/whitepaper/mos-technology-brief-mean-opinion-score-algorithms-speech-quality-evaluation (accessed 20 February 2018).

ViaSat (2017). http://www.viasat.com/sites/default/files/media/documents/ecc3100_skyphy_receiver_asic_datasheet_11_web.pdf (accessed 16 February 2018).

Worth, R. (2006). *The Telegraph and Telephone*. Milwaukee, Wi: World Almanac Library http://www.tek.com/dl/MOSTechnologyBrief_CCW-24142-0_WP_NM.pdf.

Ziemer, R.E. and Tranter, W.H. (2015). *Principles of Communications*, 7e. Hoboken, NJ: Wiley.

6

Modulation and Multiple Access

6.1 Introduction

A radio wave – called a *carrier wave* – contains little information of its own. The carrier has a frequency, a magnitude, and a phase angle; one or more of these parameters must be varied with time to convey information. An unmodulated carrier is called a CW signal, for *continuous wave*. The earliest radio transmissions, pioneered by Marconi and others around 1900, were made by turning the carrier on and off to send short and long bursts of the carrier in the form of Morse code. Information is transmitted on radio waves by modulating the carrier in proportion to the signal that is to be transmitted. The information carrying signal is called a *baseband signal*. Typical baseband signals are *audio* (voice or music), *video* (television), or *data* (bits). The term baseband is used because the frequency content of the signal extends from close to zero hertz to some upper limit.

The baseband signals must be modulated onto the radio frequency (RF) carrier to convey information as a radio signal. An RF carrier has the general form

$$v(t) = A\cos(\omega t + \varphi) = A\cos(2\pi f t + \varphi) \tag{6.1}$$

where A = magnitude (volts), $\omega = 2\pi f$ is the angular frequency (rad/s), f is the carrier frequency (hertz), and φ = phase angle (radians). We can vary any (or all) of the three parameters to impress information on the carrier. If we modulate a parameter of the RF carrier in direct proportion to the voltage of the baseband signal, we have *analog modulation*.

For example, varying the amplitude of the carrier produces amplitude modulation (AM).

If we vary a parameter of the RF carrier between two or more discrete states, we have *digital modulation*. The most basic form of digital modulation is *binary*, where the carrier is switched between two states to represent the binary states 1 and 0.

6.1.1 Analog Modulation

In analog modulation, the modulating signal (the baseband signal carrying the information to be sent on the radio wave) is proportional to a physical quantity, for example a voltage derived from a microphone that is proportional to the sound pressure on the microphone's transducer. The RF carrier has only three parameters – amplitude, frequency, and phase – so an analog modulation system must vary one of those three

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

parameters (or possibly two at the same time). The main analog modulation methods are: amplitude modulation (AM), frequency modulation (FM), and phase modulation. Analog modulation is usually AM or FM, and both are in widespread use for broadcasting. Phase modulation is rarely used directly in analog form. AM is the oldest form of modulation, and its use today is largely restricted to broadcasting in the radio frequency bands below 30 MHz. It is also the easiest to produce in a radio transmitter and to demodulate in a radio receiver, so it was employed by the first broadcast radio systems in the 1920s. AM is not the best modulation technique for sound broadcasting, as it has poor noise performance, and its use is confined to the lower RF frequencies where interference from manmade noise is greatest.

In the 1930s, Edwin Howard Armstrong pioneered FM broadcasting, which produces better quality transmissions than AM, with much less static and interference. The audio (baseband) bandwidth employed for sound broadcasting was widened from the 300–3500 Hz range used in AM radio to 100 Hz to 15 kHz for FM, giving much better fidelity to broadcasts of music. The bandwidth occupied by AM broadcast signals is typically 8 kHz with a channel spacing of 10 kHz. Throughout the world, sound radio broadcasting uses the vhf band, specifically 88–108 MHz and FM. The bandwidth of an FM broadcast signal is typically 180 kHz, with a channel spacing of 200 kHz. Analog modulation is now largely obsolete, and confined to the world of radio broadcasting; however, radio broadcasting continues to serve a large population of listeners, particularly those in automobiles. One other radio system that still uses AM is ground to air communication with aircraft. In the United States, the band 118–136 MHz is used by air traffic control to communicate with pilots. The channel bandwidth is 25 kHz, a legacy of two way vhf radios of the 1960s. Outside the United States, channel bandwidths of 8 kHz are used in many countries, greatly increasing the number of radio channels available.

The major advantage of FM over AM is the improvement in baseband signal to noise ratio (SNR) that can be achieved in an FM receiver relative to an AM receiver for a given carrier to noise ratio (CNR) in the received RF signal. In a broadcast AM receiver, the baseband SNR is always less than the RF CNR. In broadcast FM, the RF bandwidth is much greater than the baseband bandwidth. Conventional FM broadcasting uses a baseband bandwidth of 15 kHz and an RF bandwidth of 180 kHz. The characteristics of FM modulation are such that the wide bandwidth of the RF signal suppresses noise at the demodulator output giving a 26 dB improvement in baseband SNR over the received signal CNR. A further 13 dB subjective improvement in baseband SNR is achieved with pre-emphasis in the transmitter and de-emphasis in the receiver. (Couch 2007; Haykin and Moher 2009) FM demodulators output more noise at higher baseband frequencies than at lower baseband frequencies. De-emphasis in the receiver progressively reduces the output of the FM demodulator at higher baseband frequencies to suppress the high frequency noise. A compensating pre-emphasis circuit in the transmitter ensures that the baseband signal integrity is maintained by creating constant gain from transmitter to receiver for the baseband signal across its entire frequency range. The action of the pre-emphasis circuit in the transmitter is to change frequency modulation to phase modulation at higher baseband frequencies, making broadcast FM a hybrid of frequency and phase modulation.

The net improvement in SNR for broadcast FM by using wideband FM and pre-emphasis and de-emphasis is 39 dB, so with a receiver RF CNR of 13 dB, the baseband SNR is 52 dB. This is regarded as a good quality audio signal. A broadcast AM receiver with CNR 13 dB would output a SNR less than 13 dB and would be unusable. For details

of analog modulations refer to any of the many classic texts on communication theory (Couch 2007; Haykin and Moher 2009; Lathi and Ding 2009).

6.2 Digital Modulation

In digital modulation, a parameter of the carrier is varied between two or more discrete states. As in analog modulation, our choice is restricted to varying the amplitude, frequency, or phase of the carrier wave in response to the state of the modulating signal. Historically, digital modulation was called *shift keying*. The basic digital modulations are ASK, FSK, and PSK. Equation (6.1) shows that a carrier wave has three parameters, so we can have three digital modulations:

Modulate A – amplitude shift keying (ASK)
Modulate ω or f – frequency shift keying (FSK)
Modulate φ – phase shift keying (PSK)

Digital modulation is usually FSK, PSK, or a combination of ASK and PSK. ASK is inefficient, and rarely used except for radio telegraphy – Morse code transmissions. ASK and PSK can be combined to form APSK (*amplitude phase shift keying*) and QAM (*quadrature amplitude modulation*), which are discussed in Section 6.3.2. Virtually every new development in radio communication is based on digital modulation. The power of digital processing, especially *digital signal processing* (DSP) makes digital modulation a natural choice.

6.2.1 Phase Shift Keying

While any parameter of a carrier waveform – amplitude, frequency, or phase – may be digitally modulated, phase modulation is almost universally used for satellite links. An m-PSK modulator puts the phase of a carrier into one of m states according to the value of a modulating voltage. The phase state of the transmitted signal represents a *symbol*, which can convey more than one bit. Two-state or biphase shift keying, is called BPSK (sometimes called *phase reversal keying* PRK) and four-state or quadriphase PSK is termed quadrature phase shift keying (QPSK). BPSK transfers one bit per symbol and QPSK transfers two bits per symbol. Other numbers of states and some combinations of amplitude and phase modulation are also possible. For direct to home satellite TV applications, the digital video broadcast – satellite (DVB-S) standard employs QPSK with two bits per symbol, and the DVB-S2 standard recommends QPSK and 8-PSK. 8-PSK conveys three bits per symbol and maintains a constant voltage signal and phase steps of 45°. Figure 6.1 shows phasor diagrams for BPSK, QPSK, and 8-PSK. Each of the phase states in Figure 6.1 has the same magnitude, which is a significant advantage when the signal is transmitted through a non-linear transponder.

Signals received from satellites are usually weak with significant random noise (*additive white Gaussian noise*, AWGN) added to the signal. The logic associated with the demodulator in the receiver must decide which symbol to output based on the received phase. In a QPSK demodulator, the received phase state can vary by up to 45° from the expected state before an error occurs, whereas with 8-PSK the tolerance is only 22.5°. This is illustrated in Figure 6.1 by the shaded regions. A symbol error will occur when additive noise, represented by the phasor v_n in Figure 6.1 adds to the signal phasor v_i to

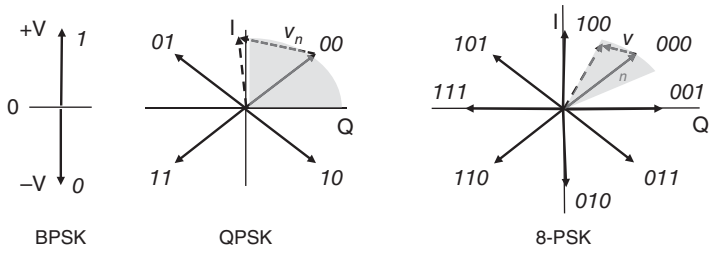


Figure 6.1 Phasor diagrams for BPSK, QPSK, and 8-PSK. The transmitted signal has a constant magnitude v_i volts and the phasor represents a symbol in each case. The assignment of logical words to each phase state is arbitrary, but the order is Gray coded so that the most likely error caused by additive noise changes only one bit in the symbol. The shaded areas in the diagrams for QPSK and 8-PSK illustrate the phase regions within which the symbol is received correctly.

give a resultant received phasor v_r that lies outside the shaded region. The most likely symbol error is where v_r has a phase error between $\pm 45°$ and $\pm 90°$ for QPSK, or between $\pm 22.5°$ and $\pm 45°$ for 8-PSK, resulting in the received signal phase lying in the adjacent symbol region. The modulator will output an incorrect symbol when this occurs. The sequence of logical words assigned to the phase states is *Gray coded* so that the most likely symbol error causes only a single bit error. Gray coding is used in the mapping of digital words to the symbols in a signal *constellation*, like those shown in Figures 6.1–6.3. The most likely error to occur in any symbol is that noise is added to the symbol voltage in the receiver in such a way that the symbol is interpreted as an adjacent symbol in the constellation, because this error requires the smallest amount of noise to add to the symbol compared to changing two or more bits. The purpose of Gray coding is to ensure that adjacent symbols in the constellation differ by no more than one bit. This can be seen in Figure 6.2 for the QPSK and 8-PSK constellations, and in Figure 6.3 for the 16-QAM and 16-APSK constellations. A much larger noise voltage must be added to the signal to cause the phase angle of the received signal to cross two boundaries and cause a two bit error. The probability that noise reaches this magnitude is low compared

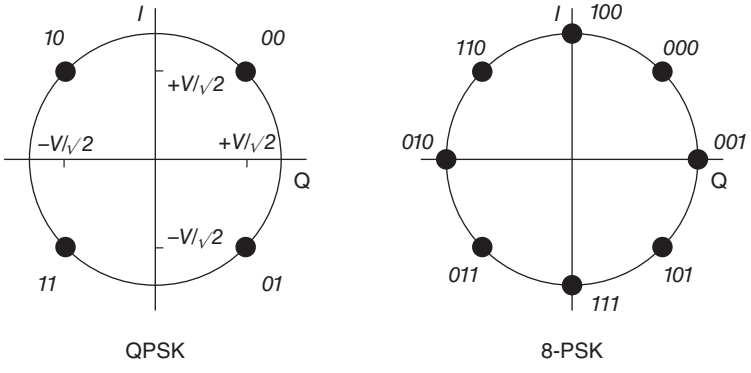


Figure 6.2 Constellation diagrams for QPSK and 8-PSK. The dots represent the tips of the phasors shown in Figure 6.1. Because the dots lie on a circle in both QPSK and 8-PSK, the magnitude of the modulated wave is constant at V volts. The assignment of symbols to each phase state is Gray coded: adjacent states always differ by only one bit.

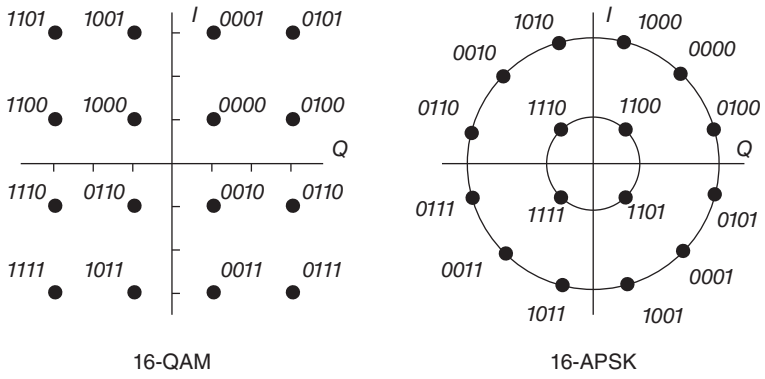


Figure 6.3 Constellation diagrams for 16-QAM and 16-APSK. The Gray coding applied in each case is not unique; other Gray coded sequences can be used, and the assignment of symbols to phase states is also arbitrary. 16-APSK is the preferred choice when the transmitted signal must pass through a non-linear transponder, but 16-QAM has the advantage that a lower CNR is required for a given bit error rate than with 16-APSK.

to the probability that the noise causes a single bit error. Hence, the bit error rate (BER) is generally assumed to be equal to the symbol error rate provided Gray coding of the symbol sequence is employed. The Gray coded sequences shown in Figures 6.1–6.3 are not unique; other Gray coding sequences can be found in the literature (Gray 1953).

It is evident from Figure 6.1 that it takes much less additive noise to cause a bit error with 8-PSK than with QPSK. We could guess that the value of v_n to cause an error with 8-PSK is about half that needed to cause an error with QPSK, so for a given noise power in the receiver the voltage of a received signal with 8-PSK modulation for a given error rate must be about twice that for QPSK. Since power is proportional to voltage squared, we need approximately 6 dB greater CNR when employing 8-PSK than when QPSK is used. In an ideal communication system with AWGN, the CNR for a symbol error rate of 10^{-6} with QPSK modulation is 13.6 dB and for 8-PSK modulation it is 18.5 dB.

Representation of Sine Waves Using Phasor Diagrams

Any sine wave has the generic form given in Eq. (6.1)

$$v(t) = A\cos(\omega t + \varphi) = A\cos(2\pi f t + \varphi)$$

where A is the magnitude of the wave, ω is its angular frequency in rad/s, f is its frequency in hertz and φ is the phase angle of the sine wave relative to a reference. (Although Eq. (6.1) is written as $A\cos(\omega t + \varphi)$ the waveform is always described as a sine wave, not a cosine wave.) Sine waves at the same frequency differ in amplitude and phase and can therefore be described by any variable that has two parameters. With a fixed frequency of f hertz, Eq. (6.1) has two parameters that can be varied, A and φ, but is in a form that is often not convenient for analysis because drawing sine waves is a chore. Another way to describe the sine wave is with a complex exponential function such as

$$v(t) = \frac{A}{2}\exp(j\omega t + \varphi) + \frac{A}{2}\exp(-j\omega t + \varphi)$$

Complex arithmetic is widely used in communication systems texts as a mathematically convenient way to analyze sine waves. The most useful description of the sine wave for analysis of modulated waves is with two sine waves in phase quadrature

$$v(t) = A_{\mathrm{i}} \cos \omega t + A_{\mathrm{q}} \sin\ \omega t$$

where the cosine wave is regarded as the in-phase component with magnitude A_{i} volts and the sine wave is a quadrature component with a phase shift of 90° with respect to the in-phase component and a magnitude of A_{q} volts. The magnitude of the wave is then

$$A = \sqrt{{A_{\mathrm{i}}}^2 + {A_{\mathrm{q}}}^2} \text{ volts}$$

and the phase angle

$$\varphi = \tan \frac{A_{\mathrm{q}}}{A_{\mathrm{i}}}$$

relative to the in-phase component.

We can now describe any sine wave with a *phasor diagram* that has *x* and *y* axes for the in-phase and quadrature components of the sine wave, as shown in Figure 6.1. For example, the QPSK signal in Figure 6.1 has four phase states at phase angles of 45°, 135°, 225°, and 315° relative to the in-phase I axis. Each state has the same magnitude of *V* volts, represented by the length of the line in Figure 6.1. The QPSK signal can be generated by phasor addition of the in-phase component with magnitude $\pm V/\sqrt{2}$ volts and the quadrature component of $\pm V/\sqrt{2}$ volts to create a phasor with magnitude

$$A = \sqrt{[\pm V/\sqrt{2}]^2 + [\pm V/\sqrt{2}]^2} = V \text{ volts}$$

It is easy to see from the phasor diagram for the QPSK signal in Figure 6.1 how a QPSK modulator can be implemented by adding the outputs of an in-phase intermediate frequency (IF) carrier to a quadrature IF carrier. This is how a QPSK modulator can be built in practice.

Modulations can be represented graphically by *constellation diagrams*. A constellation diagram shows the magnitude and phase of the symbols that are transmitted by dots with the digital word corresponding to each symbol alongside. The dot represents the tip of a phasor from the origin of the diagram to the dot, with the length of the line representing the voltage of the symbol and its angle from a zero degree reference phase showing the phase of the symbol. Constellation diagrams for QPSK, and 8-PSK are shown in Figure 6.2. BPSK, QPSK, and 8-PSK all transmit symbols with a constant voltage level and avoid the compression problem of non-linear transponders.

6.2.2 QAM and APSK

Higher order modulations allow more bits to be carried by each symbol, but require an increasingly high CNR in the receiver to maintain a given BER when compared to modulations with fewer bits per symbol. Terrestrial microwave links and cable TV systems employ QAM, which combines multiple voltage levels with QPSK. 256-QAM carries

eight bits per symbol and 1024-QAM carries 10 bits. The advantage of higher order modulations is that more information is transferred in a given radio frequency bandwidth because it is the symbol rate that is related to bandwidth. Hence given a channel with a bandwidth of B Hz that carries R_b bits per second with BPSK modulation, the same channel can carry $2R_b$ bits per second with QPSK and $3R_b$ bits per second with 8-PSK. The DVB-S2 standard includes 16-APSK (sometimes referred to as 4 + 12 PSK) and 32-APSK, modulations that combine more than one amplitude level with PSK (ETSI 2009). These modulations are typically restricted to links that are not direct to home television (DTH-TV) because of the higher CNR and better phase stability required in the receiver, such as point to point video links and electronic news gathering systems.

Two examples for modulations that transmit four bits per symbol are shown in Figure 6.3. 16-APSK and 16-QAM both employ two voltage levels and eight phase states to generate 16 symbols. In 16-APSK the phasors take only two possible voltage values, V volts and $3V$ volts. In 16-QAM there are three voltage levels, with the disadvantage that the symbols at the corners of the square constellation require considerably more transmitted power than the other symbols. 16-APSK is a better choice than 16-QAM with a non-linear transponder. The power assigned at the uplink transmitter to the outer ring of symbols in the 16-APSK constellation can be increased to compensate for compression in the satellite transponder. This keeps the space between all symbols the same resulting in the same error rate for each symbol. If uncompensated compression occurs, the error rate for the symbols with larger voltages will be higher than for the symbols with smaller voltages. The main advantage of m-QAM over m-APSK is that the CNR required at the receiver for a given BER is lower with m-QAM.

For most geostationary earth orbit (GEO) satellite communication links, 8-PSK and 16-PSK are preferred over the equivalent 8-QAM and 16-QAM modulations. The advantage of m-QAM over the equivalent m-PSK is that a lower CNR is required in the receiver to achieve a given BER. However, most GEO links drive the satellite transponder high power amplifier (HPA) into its non-linear region, toward saturation. This reduces the voltage spacing between the higher voltage levels in QAM signals and results in an increase in BER. In dedicated satellite links such as DTH-TV, non-linearity in the transponder can be overcome by pre-distortion of the signal transmitted by the uplink earth station. However, this further increases the power that must be transmitted when m-QAM modulation is employed rather than m-APSK. Where a baseband processing transponder is used on the satellite, QAM can be employed on the uplink because the earth station HPA can be kept within its linear region, and the modulation used on the downlink can be different from the uplink. For example, 16-QAM could be used for the uplink and 16-PSK for the downlink.

Any type of PSK can be *direct* or *differential,* depending on whether it is the state of the modulating voltage or the *change* in state of the modulating voltage that determines the transmitted phase. Whether direct or differential, a PSK modulator causes the phase of a carrier waveform to go to one of a finite set of values. The transition time plus the time spent at the desired phase constitute a fixed time interval called the *symbol period;* the transmitted waveform during the interval is called a *symbol*. The set of all symbols for a particular modulation type is called its *alphabet*. Thus BPSK has a two-symbol alphabet and QPSK has a four-symbol alphabet.

In the digital modulation process, a stream of incoming bits determines which symbol of the m available in the alphabet will be transmitted. Mathematically, N_b bits are

required to specify which of m possible symbols is being transmitted where N_b and m are related by

$$N_b = \log_2 m \tag{6.2}$$

As defined by this equation, N_b is the number of *bits per symbol* for the m-PSK modulation scheme. Standard practice is to make m a power of 2 so that N_b will be an integer.

Simple FSK is rarely used on satellite links, although a version known as continuous phase frequency shift keying (CPFSK) has been used by some amateur satellites (Davidoff 1998). FSK has the advantage that a coherent detector is not required in the receiver, making for simpler receiver design, but does not perform as well as PSK when comparing BER for a given CNR. A variant of CPFSK known as *minimum shift keying* (MSK) is used for single channel per carrier (SCPC) satellite links because it produces a narrower spectrum than conventional PSK modulators that require tight SRRC (*square root raised cosine*) filtering to avoid adjacent channel interference. MSK is also used in the global system mobile (GSM) cellular telephone system. See Chapter 5 for a discussion of SRRC filtering.

Table 6.1 shows the theoretical CNR and *energy per bit to noise power spectral density ratio* E_b/N_o required to achieve a symbol error rate of 10^{-6} with a number of different modulations in a perfect link with ideal SRRC filtering, no phase jitter or interference, and only additive white Gaussian noise (AWGN). CNR and E_b/N_o are measures that are used to determine the performance of a digital communication link in terms of symbol error rate. Refer to Chapter 5 for a detailed discussion of these relationships.

It should be remembered that no practical link can achieve the BER values shown in Table 6.1 for the given CNR. There are no ideal SRRC filters and phase jitter is always present in receivers, especially at very low CNRs. Interference from adjacent channels, other satellites, and terrestrial transmitters is often present in a satellite link. An implementation margin must be added to the theoretical CNR to achieve a given BER in practice. Although QAM has the lowest CNR for a given BER, the circular constellations of the PSK modulations lead to their being preferred in many satellite links. The CNR penalty for using 8-PSK rather than 8-QAM is 0.9 dB, which can be tolerated in exchange for the constant magnitude of all the phase states in 8-PSK. The penalty with 16-APSK compared to 16-QAM is even lower at 0.5 dB.

Table 6.1 Theoretical CNR and E_b/N_o ratio required to achieve BER of 10^{-6} in an ideal link

Modulation	CNR (dB)	E_b/N_o ratio (dB)
BPSK	10.6	10.6
QPSK	13.6	10.6
8-QAM	17.6	10.6
8-PSK	18.5	14.0
16-QAM	20.5	14.5
16-APSK	21.0	15.0
16-PSK	24.3	18.3
32-QAM	24.4	17.4

Error rates for PSK and QAM signals can be calculated from the following formulas (Rappaport 2002).

For m-PSK, the probability of a symbol error is given by

$$P_e = 2\,Q[\sin(\pi/2m) \times \sqrt{4E_s/N_o}] \tag{6.3}$$

For m-QAM the probability of a symbol error is given by

$$P_e = 4[(1 - (1/\sqrt{m})) \times Q\sqrt{3E_{av}/((m-1)N_o)}] \tag{6.4}$$

6.2.3 Generation and Demodulation of BPSK Signals

Phase shift keyed waveforms can be created by multiplying the carrier wave by the digital baseband signal. The simplest PSK waveform is BPSK, which can be generated by multiplying the carrier by +1 or −1, corresponding to the logical values 1 and 0 (or 0 and 1). Denoting the modulating signal as $u_i(t)$, where u_i has a value of +1 or − 1, the BPSK waveform is

$$V_{BPSK} = V_c \cos(\omega_c t + u_i \pi/2) = u_i V_c \sin \omega_c t \tag{6.5}$$

where V_c is the magnitude of the PSK signal in volts and ω_c is its angular frequency in radians per second. Figure 6.1 shows the phasor diagram for a BPSK signal, with $V_c \cos \omega_c t$ as the in phase carrier. The BPSK waveform has phase reversals, which lead to a wide sinc shaped spectrum that must then be filtered by a SRRC band pass filter (BPF) before transmission over a radio link. The process was illustrated in Figure 5.5 in Chapter 5, which is repeated here as Figure 6.4. The baseband signal is a non-return-to-zero (NRZ) digital waveform $u_i(t)$ which is the modulating input to a multiplier driven by the carrier $V_c \cos(\omega_c t)$. Equalization with an $x/\sin x$ equalizer is required because the input to the SRRC filter is an NRZ waveform, not a sequence of impulses. The carrier is typically at a convenient IF that is subsequently up converted to the required RF frequency. The spectrum at each stage of the modulator is shown in Figure 6.4c.

The BPF in the transmitter is an SRRC filter and the equalizer after the BPF is required to obtain the correct zero intersymbol interference (ISI) waveform with the transmitted spectrum $|v_4(t)|$ in Figure 6.4c. The BPF in the receiver is an identical SRRC filter. (Repeat of Figure 5.5 in Chapter 5.)

Demodulation of a BPSK waveform in a radio receiver is achieved by multiplying the BPSK signal with a locally generated carrier, after filtering with a bandpass SRRC filter, as illustrated in Figure 6.5. The RF signal is usually down converted to a convenient IF before filtering and multiplication by a replica of the unmodulated carrier wave. The locally generated carrier must be phase locked to the received signal, which is achieved with a phase locked loop (PLL). The output of the PLL is a waveform $V_1 \cos(\omega t)$ with V_1 set to unity. Hence, the output of the multiplier in Figure 6.5 is

$$V_6(t) = Vu_i \cos(\omega t) \times \cos(\omega t) = \frac{1}{2} V u_i(t) + \frac{1}{2} V u_i(t) \cos(2\omega t) \tag{6.6}$$

where the trigonometric identity

$$\cos^2 x = \frac{1}{2} + \frac{1}{2}\cos(2x)$$

has been applied. The low pass filter (LPF) following the multiplier removes the double frequency term and outputs the baseband signal $u_i(t)$. Figure 6.5 shows the blocks of

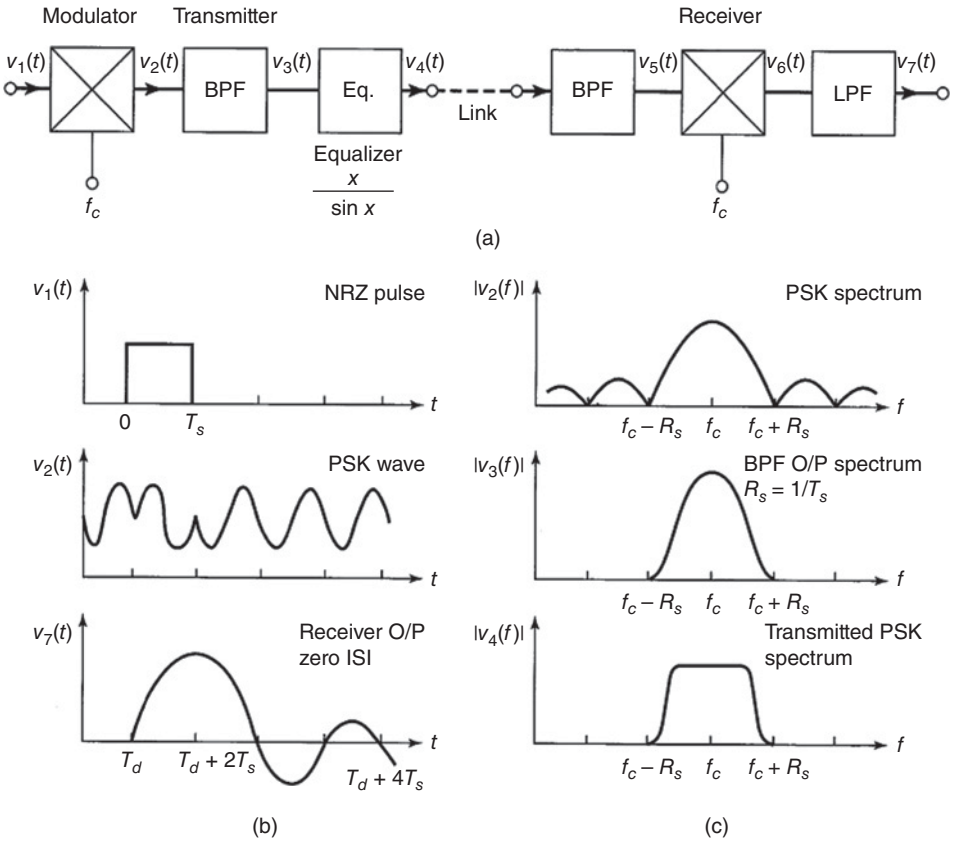


Figure 6.4 Simplified diagram of a BPSK transmitter and receiver showing modulation and demodulation of a single NRZ pulse. The blocks typically operate at an intermediate frequency. The input NRZ waveform is multiplied by an IF carrier to give the PSK waveform v_2(t) in Figure 6.4b. The output of the demodulator and receiver LPF is a zero ISI waveform $v_7(t)$.

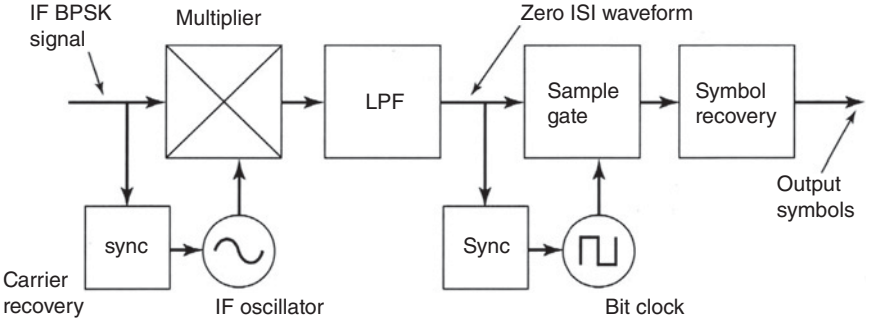


Figure 6.5 Block diagram of a BPSK demodulator. The blocks shown correspond to a hardware version operating at the intermediate frequency of the receiver. The first sync block is a PLL that recovers the IF carrier from the incoming noisy BPSK signal and locks in the IF oscillator. The LPF following the multiplier removes the double frequency component in Eq. (6.6). The second sync block is a PLL that locks to the bit rate of the received signal and synchronizes the bit clock that drives the sample gate. The symbol recovery block outputs digital symbols.

a hardware implementation of a BPSK demodulator. In a digital receiver, the IF BPSK signal is sampled with I and Q analog to digital converters (ADCs) and the operations of all the subsequent blocks is then performed digitally.

In a conventional BPSK radio receiver employing hardware IF devices, the multiplier in the modulator and the demodulator is typically a *balanced mixer*, which is an electronic switch using diodes or field effect transistors (FETs). Specialized integrated circuit balanced mixers are also available taking the form of variable gain amplifiers where the gain can be varied from +G to –G. Details of balanced mixer design can be found in most texts on communication systems, for example (Couch 2007, pp. 257–263; Ziemer and Tranter 2015, pp. 168–170). In a digital radio transmitter, PSK waveforms can be generated directly, as discussed in Chapter 5. This avoids the need for hardware SRRC filters which require physically large inductors that are not compatible with small handheld radios. In the receiver, the IF SRRC filter can be implemented as a finite impulse response (FIR) digital filter. Many digital receivers employ direct conversion from the output of the IF stage to baseband using high speed sampling and analog to digital conversion of the IF signal, allowing implementation of the SRRC filter at baseband. See Chapter 5 for details of these techniques.

To recover the data bits u_i the receiver uses a *correlation detector*, which is the equivalent of a *matched filter*. A matched filter has a transfer function with a magnitude response that is identical to the magnitude of the spectrum of the signal it is required to process, and maintains the highest possible CNR in the receiver. A correlation detector multiplies the received signal by a replica of the unmodulated transmitted signal, integrates the result over the symbol period, and samples the output of the integrator at the end of the period (Ziemer and Tranter 2015, p. 469). The practical implementation of a correlation detector for BPSK signals is shown in Figure 6.5. The replica of the unmodulated transmitted signal is a locally generated carrier that creates a coherent (in phase) sine wave at the carrier frequency. The BPSK demodulator of Figure 6.5 multiplies the received signal by the local carrier in a mixer, and then uses a LPF, rather than an integrator to recover the demodulated waveform. Ideally, the recovered time waveform at the LPF output is a zero ISI waveform, which has a maximum value in the center of the received symbol period.

Multiplying the BPSK signal by a coherent carrier, followed by low pass filtering is the key to recovering the baseband signal. The BPSK signal is defined in Eq. (6.5), without shaping by the Nyquist filters of the transmitter and receiver. When multiplied by a locally generated, in-phase carrier, sin $\omega_c t$ the multiplier output is $v_o(t)$ where

$$v_o(t) = u_i V_c \sin(\omega_c t) \times \sin(\omega_c t) = u_i V_c \times \frac{1}{2}\left(1 - \cos^2(\omega_c t)\right) \tag{6.7}$$

The LPF removes the double frequency term generated by $\cos^2(\omega t)$ leaving an output signal $v_r(t)$.

$$v_r(t) = \frac{1}{2} u_i V_c \tag{6.8}$$

Thus the BPSK demodulator has recovered the modulating signal u_i. The magnitude factor V_c is removed in the IF stages of the receiver by limiting the magnitude of the received signal, exactly in the same way that a broadcast FM receiver limits a received FM signal prior to demodulation.

At the center of each symbol interval, the output of the demodulator is sampled and a decision circuit decides whether $v_r(t)$ is greater or less than 0 V (i.e., whether the sample is a positive or a negative voltage) and thus determines whether the transmitted signal u_i was a $+V$ and represented a data one or whether it was a $-V$ and represented a data zero. This is called a *hard decision* and contains no information on the magnitude of the sample. If the sample is much smaller than the expected value, it is possible that an error has occurred. Forward error correction techniques such as turbo codes and low density parity check (LDPC) codes make use of *soft decision* sampling, which provides information about the magnitude of the sample and improves the correction capability of the FEC decoder.

The reference carrier that drives the multiplier (mixer) of the BPSK demodulator shown in Figure 6.5 must be derived from the received signal. This is accomplished with a *carrier recovery circuit*. One such circuit applies the BPSK signal to both inputs of a multiplier (mixer) to obtain the square of the BPSK signal. This has the effect of stripping off the modulation, but creates a double frequency signal component. The BPSK signal given by Eq. (6.5) is squared to give

$$v(t)^2 = u_i^2 V_c^2 \sin^2(\omega_c t) \tag{6.9}$$

Since u_i is either +1 or −1, $(u_i)^2$ is always +1. We will ignore the V_c^2 term since it is easy to limit an AC waveform to a predetermined magnitude. Expanding the $\sin^2(\omega_c t)$ term gives an output from the squarer circuit of

$$v_o(t) = \frac{1}{2}[1 - \cos(2\omega_c t)] = \frac{1}{2} - \frac{1}{2}\cos(2\omega_c t) \tag{6.10}$$

We can extract the double frequency term with a BPF tuned to $2f_c$ Hz and then divide it by two, which is easily accomplished with a PLL, to give a reference carrier at the correct frequency, and with the correct phase angle. (The PLL provides a 90° phase shift that converts $\cos(\omega t)$ to $\sin(\omega t)$.) Other techniques for carrier recovery, such as the Costas loop, are also used in coherent PSK receivers (Ziemer and Tranter 2015, p. 561).

Most carrier recovery loops for BPSK have a 180° phase ambiguity; that is, when the loop is locked the phase of the recovered carrier may differ by 180° from the correct value. This has the effect of interchanging logical ones and zeroes and causes the demodulated bit stream to be the complement of what was transmitted. With QPSK, the loop can lock up in four different states, with offsets of +90°, 180°, and −90°. There are several ways to eliminate the ambiguity; one is to use differential encoding in which adjacent symbols have the same phase if the modulating voltage is a logical 1 and are 180° out of phase if the modulating voltage is a logical 0. This may be realized by a binary phase shifter that toggles between 0° phase shift and 180° phase shift each time the modulating bit is a 0. Incoming logical 1 values have no effect.

Differential modulation is more error prone than direct modulation, since an error on a single bit in a differential system will cause one or more subsequent bits to be interpreted incorrectly. See reference (Haykin and Moher 2009, pp. 152–154) for a detailed analysis of differential PSK. Most practical satellite systems avoid differential encoding and check the status of the recovered carrier phase periodically by transmitting a known (unique) word or flag at the start of each frame or packet. Logic in the receiver looks for the flag or unique word (UW). If the unique word is received correctly, then the recovered carrier phase is correct. If the unique word is the complement of the known word (1 s and 0 s interchanged), than the recovered carrier phase is off by 180° (in a BPSK

receiver) and the demodulated data stream must be complemented before it is sent to the end user.

6.2.4 Generation and Demodulation of QPSK Signals

Generation of QPSK waveforms requires two BPSK modulators and two carrier waveforms in phase quadrature as illustrated in Figure 6.6. A single IF oscillator is used in the transmitter to generate an *in-phase carrier* and a 90° phase shift of the in-phase signal generates a *quadrature carrier*. The modulator driven by the in-phase carrier is called the I channel and the modulator driven by the quadrature carrier is called the Q channel. Thus all QPSK transmitters and receivers have two channels, I for in-phase and Q for quadrature. The magnitude of the modulated wave is V volts in each case. A QPSK carrier with magnitude V volts is generated by adding a sine wave with magnitude $V/\sqrt{2} = 0.707V$ volts from the I channel to a sine wave with magnitude $V/\sqrt{2} = 0.707V$ volts from the Q channel.

In QPSK the phase, ϕ, of the carrier is set by the modulator to one of four possible values. We may write the result as

$$v(t) = V/\sqrt{2}\ \cos(\omega_c t - \varphi) \tag{6.11}$$

where φ takes on the values $\pi/4$, $3\pi/4$, $5\pi/4$, and $7\pi/4$ rad (45°, 135°, 225°, 315°). Using trigonometric identities to expand Eq. (6.11) we obtain

$$v(t) = V/\sqrt{2}\ [\cos(\omega_c t)\cos\varphi + \sin(\omega_c t)\sin\varphi] \tag{6.12}$$

The first term in the square brackets is a BPSK signal in phase with the carrier, and is called the I channel. The second term is a BPSK signal in quadrature with the carrier

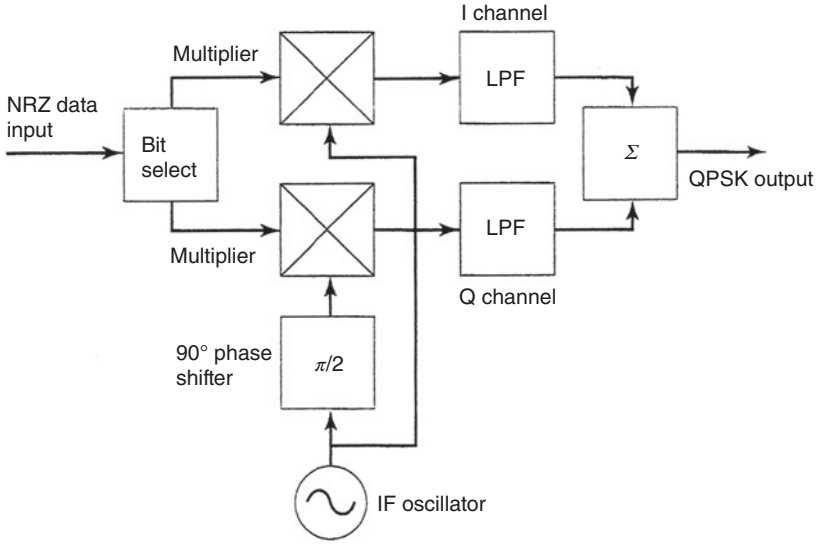


Figure 6.6 Block diagram of a QPSK modulator implemented in hardware form.

and is called the Q channel. Thus a QPSK waveform may be generated by combining two BPSK waveforms in quadrature. We may write the result as

$$v(t) = u_i V \cos(\omega_c t) + u_q V \sin(\omega_c t) \tag{6.13}$$

where u_i represents a binary data stream modulating the I channel and u_q represents a binary data stream modulating the Q channel. For both of these signals a logical 1 corresponds to u_i or $u_q = +1$, and a logical 0 corresponds to u_i or $u_q = -1$, giving the QPSK phasor diagram seen in Figure 6.1. The bits u_i and u_q are selected alternately from the input bit stream. For example, u_i may represent the odd-number bits and u_q the even-number bits in an incoming bit stream. In this case one binary data channel enters the QPSK modulator and the outgoing symbol rate is equal to half of the incoming bit rate.

QPSK modulators and demodulators are basically dual-channel BPSK modulators and demodulators. The I channel processes the u_i bits and uses the reference carrier; the Q channel processes the u_q bits and uses a 90° phase shifted version of the reference. Figures 6.6 and 6.7 show generalized block diagrams of a QPSK modulator and demodulator. More detailed information is available in (Haykin and Moher 2009, p.228; Rappaport 2002, pp. 302–304).

Figure 6.7 shows a block diagram of a QPSK demodulator. The bit select block sends alternate bits to the I and Q channels where they modulate in-phase and quadrature IF carriers. The LPFs remove double frequency components generated by the multiplication process and the two modulated carriers are added to give a QPSK signal. The modulator must be followed by an SRRC filter to achieve the correct spectrum for transmission.

The unmodulated carrier of the incoming noisy QPSK signal at IF is extracted by the carrier recovery block and synchronizes the local IF carrier oscillator. The received

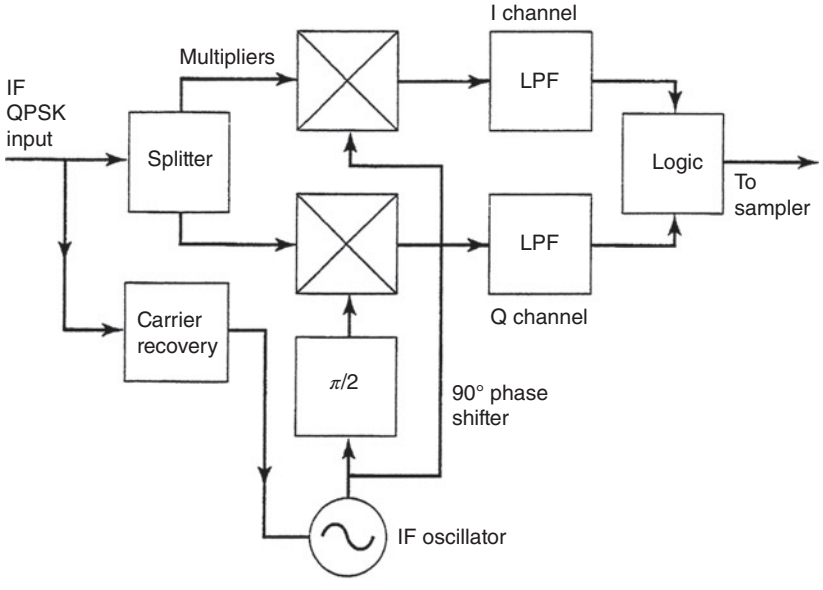


Figure 6.7 Block diagram of a QPSK demodulator implemented in hardware.

QPSK signal is applied to two multipliers driven in phase quadrature from the IF oscillator to create I and Q channels. The output of the in-phase multiplier and LPF is the u_i bit stream and the output of the quadrature multiplier and LPF is the u_q bit stream. The demodulator must be preceded by an SRRC filter to avoid the generation of inter-symbol interference.

The received QPSK symbols are periods of the IF waveform with one of four possible phases. The waveform is applied to both the I and Q channel multipliers, but only one of the two multipliers will have an output in any given period. The IF oscillator drives the I channel multiplier with a waveform $\cos \omega_c t$ and the Q channel multiplier with waveform $\sin \omega_c t$. If the received QPSK waveform is an in-phase symbol, $u_i\ V \cos \omega_c t$ where u_i has a value +1 or −1, the I channel output after the LPF will be $\frac{1}{2}u_i$ as indicated in Eq. (6.7), with $\cos \omega_c t$ substituted for $\sin \omega_c t$. The output of the Q channel LPF will be zero, because multiplying the signal $u_q V \sin \omega_c t$ by the IF oscillator in-phase waveform $\cos \omega_c t$ yields only a double frequency term from the product of $\cos \omega_c t \sin \omega_c t$. The same analysis applies to the I channel, which has zero output when the input is $u_q V \sin \omega_c t$. Hence the logic block in Figure 6.7 is simply an adder when QPSK signals are demodulated.

When QAM signals are received, they are applied to a QPSK demodulator, which outputs I and Q pulses with multiple amplitudes. The logic block must then include two analog to digital converters (ADCs) that determine the bit pattern for each received symbol. For example, the 16-QAM signal in Figure 6.3 consists of I and Q channel symbols with voltage levels of $\pm 1V$ and $\pm 3V$. The demodulator for the 16-QAM is identical to a QPSK demodulator but with different logic for combining the I and Q channel outputs. Two bit ADCs are needed that convert inputs of ±1 and ±3 V to bit sequences 00, 01, 10, and 11. For example, if a 16-QAM signal for the bit sequence 0001 shown in Figure 6.3 is received, the I channel sample will be $+3V$ and the Q channel sample will be $+V$. The logic circuit is a mapping device that uses the combinations of I and Q channel voltages shown in Figure 6.3 and will output the bit sequence 0001 in this example.

6.2.5 QPSK Variants

We noted that QPSK may be visualized as the sum of two BPSK signals whose carriers are in phase quadrature. In conventional QPSK the bits u_i and u_q that modulate these carriers both make step changes at the same time. When both changes result in phase reversals, there is significant spectral spreading of the QPSK signal requiring very tight filtering by the transmitter's SRRC filter. If the bit changes are staggered so that u_i makes step changes at the beginning of each symbol period and u_q makes step changes at the midpoint of each symbol period, the result is called *offset quadrature phase shift keying* (OQPSK). This reduces the spectral spreading and eases the SRRC filter design problem.

The baseband waveforms used with QPSK and 8-PSK and variants are all non-return to zero functions that allow instantaneous transitions from +V to −V volts. The output of the modulator can change by 180° for certain bit patterns, for example, transitions from 00 to 11 and 01 to 10 in the QPSK constellation diagram in Figure 6.2. This results in a sinc function spectrum that has high sidelobes in the frequency plane, as illustrated in Figure 6.8. The SRRC filter, and any additional BPFs in the transmitter must heavily attenuate the sidelobes of the transmitted spectrum to avoid interference spreading into adjacent channels. This is particularly important when frequency division multiple access (FDMA) is used in a SCPC satellite communication system. Mobile satellite telephones (satphone) can use frequency division multiple access-single channel per

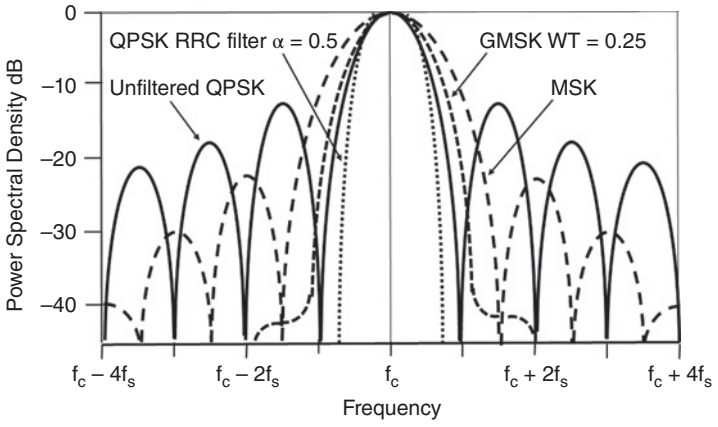


Figure 6.8 Comparison of spectra for unfiltered QPSK, QPSK with $\alpha = 0.5$ SRRC filtering, MSK, and Gaussian MSK with $WT = 0.25$. Unfiltered QPSK has the widest spectrum and QPSK with ideal SRRC filtering has the narrowest spectrum. MSK has a broader spectrum than QPSK but lower sidelobes. GMSK with $WT = 0.25$ has a broader spectrum than SRRC filtered QPSK, but does not require filtering in the transmitter making it useful for SCPC FDMA systems.

carrier (FDMA-SCPC) to allow a large number of users to access a single transponder. The process of creating a connection between a transmitter and a receiver in a satphone network allocates a transmit frequency to the individual satphone, with a specific channel bandwidth. Many channels are stacked in frequency across the bandwidth of the satellite transponder, with small guard bands between the channels. Guard bands are typically 10–15% of the channel bandwidth. The BPFs in the transmitter must attenuate the transmitted RF signal by at least 30 dB at the edge of the adjacent channel, requiring very accurate filter transfer functions. Spectral spreading can be reduced if sudden steps in phase are avoided.

In MSK modulators the baseband signal waveform applied to a linear phase modulator is a series of half sinusoidal transitions between the allowed baseband values of +V and −V, resulting in a linear phase increment over the symbol period. A linear increase or decrease in the phase of a waveform is an up or down shift in frequency, giving MSK the alternative name of fast frequency shift keying (FFSK). The frequencies that are generated by an MSK modulator are $f_c + 1/4T$ and $f_c - 1/4T$, with 99% of transmitted power contained within a bandwidth 1.2/T, where T is the symbol period. With QPSK the bandwidth that contains 99% of transmitted power is much wider at 8/T (Rappaport 2002). The bandwidth of an MSK transmission can be further reduced by changing the baseband pulse shape from a half sinusoid to a Gaussian function, giving Gaussian minimum shift keying (GMSK). In GMSK the baseband pulses are filtered with a Gaussian shaping filter with the parameter W denoting the 3 dB bandwidth of the Gaussian shaped pulse. T is the pulse period, and $WT = 0.25$ is an optimized form of GMSK that is valuable in single channel per carrier-frequency division multiple access (SCPC-FDMA) systems because the transmitter does not require a SRRC filter.

GMSK signals can be demodulated coherently, as with BPSK and QPSK, or noncoherently using a FM detector (discriminator). There is a small penalty in performance from increased ISI and higher BER, but this is quoted to be less than 0.5 dB (Haykin and Moher 2005). Figure 6.8 shows a comparison between the spectra for unfiltered QPSK,

Table 6.2 Mapping for QPSK and 8-PSK modulations shown in Figures 6.1 and 6.2

Bit pattern	*I* channel output	*Q* channel output
QPSK		
00	+0.707 V	+0.707 V
01	−0.707 V	+0.707 V
10	+0.707 V	−0.707 V
11	−0.707 V	−0.707 V
8-PSK		
000	+0.707 V	+0.707 V
001	0	+V
010	0	−V
011	−0.707 V	−0.707 V
100	+V	0
101	−0.707 V	+0.707 V
110	+0.707 V	−0.707 V
111	−V	0

QPSK with $\alpha = 0.5$ SRRC filtering, MSK, and GMSK with $WT = 0.25$. For additional information on MSK and GMSK the reader should consult (Rappaport 2002; Haykin and Moher 2005).

6.2.6 Mapping

The process of converting incoming bits in a transmitter into phase and amplitude states in a modulator is known as *mapping*. The simplest form of mapping is BPSK, where a logical 1 represented by a voltage level of $+V$ volts is mapped to a phase angle of 0°, and a logical 0 represented by a voltage level of $-V$ volts is mapped to a phase angle of 180°. Higher order modulators use the addition of the outputs of an in-phase carrier and a quadrature carrier to create QPSK, m-PSK, m-QAM, and m-APSK. Table 6.2 shows the mapping required for the QPSK and 8-PSK modulations shown in Figures 6.1 and 6.2.

6.3 Multiple Access

The ability of a satellite to carry many signals at the same time is known as *multiple access*. Multiple access allows the communication capacity of the satellite to be shared among a large number of earth stations, and to accommodate the different mixes of communication traffic that are transmitted through a satellite transponder.

The basic form of multiple access employed by most communications satellites is the use of many *transponders,* as discussed in Chapter 3. A large GEO satellite can have a communication bandwidth many times the allocated RF bandwidth; for example, 4000 MHz of capacity can be used within an allocated RF bandwidth of 1000 MHz.

Through *frequency re-use* with multiple antenna beams and *orthogonal polarizations*, the spectrum can be re-used many times over – as many as 18 times in the case of some large GEO satellites. The frequency spectrum used by the satellite is divided into smaller bandwidths, which are allocated to transponders, allowing separate communication links to be established via the satellite on the basis of transmit frequency. Transponder bandwidths from 20 to 200 MHz have been employed on GEO communications satellites, with a trend toward larger bandwidths over time. The individual transponders may carry one signal – a high speed digital stream made up of a number of television programs, for example, or hundreds of signals, as with mobile satellite telephone systems.

Smaller low earth orbit (LEO) satellites may have only one transponder used for a specific service, or multiple transponders connected to multiple beams. When the satellite has a particular application, such as earth surveillance, the information it collects is transmitted on a downlink that is usually sized to match the rate at which data is collected. If that is not possible, the LEO satellites are designed to collect and store data as they orbit the earth and then download the contents of the memory when in range of a receiving earth station. There are no transponders as in a communications satellite, but there is always an uplink to the satellite for control purposes. These systems are discussed in Chapter 8. This chapter is concerned mainly with larger GEO satellites that have many transponders and the techniques that are used to manage traffic between earth stations. The use of multiple transponders to divide up a frequency band is not generally considered as multiple access, although the reason for their use is to make it easier for earth stations to share the available frequency spectrum efficiently.

The signals that earth stations transmit to a satellite may differ widely in their character – video, data, voice – but they can be sent through the same satellite using multiple access and *multiplexing* techniques. Multiplexing is the process of combining a number of signals into a single signal, so that it can be processed by a single amplifier or transmitted over a single radio channel. Multiplexing can be done at baseband or at an IF. The corresponding technique that recovers the individual signal is called *demultiplexing*. Multiplexing is a key feature of all commercial long distance communication systems, and is part of the multiple access capability of all satellite communications systems.

In the early days of satellite communications, traffic was mainly analog video and voice signals. Analog signals were combined using frequency division multiplexing (FDM), a technique that had been used for terrestrial trunk telephone links for many years. In FDM telephony, analog signals are shifted in frequency at baseband so that each signal has its own center frequency, and then added into a single signal that covers a wide bandwidth. FDM telephony is now obsolete; details can be found in some texts on communication systems (Couch 2007, pp. 396–397; Pratt et al. 2003 Appendix B) and an internet search for *frequency division multiplexing* will yield many web sites that discuss the technique. A brief discussion is included as a sidebar in this chapter.

The designer of a satellite communication system must make decisions about the form of multiple access to be used. The multiple access technique will influence the capacity and flexibility of the satellite communication system, its cost, and its ability to earn revenue. The basic problem in any multiple access system is how to permit a changing group of earth stations to share a satellite in such a way that satellite communication capacity is maximized, bandwidth is used efficiently, flexibility is maintained, and cost to the user is minimized while revenue to the operator is maximized. The multiple access system should also allow for changing patterns of traffic over the 10 or 15 years of the expected lifetime of the satellite, especially as in-orbit refueling of GEO satellites will likely be

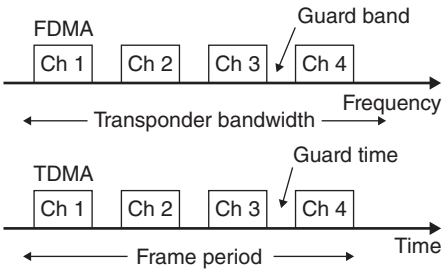


Figure 6.9 FDMA and TDMA. The blocks represent signals, which can consist of a single channel or multiple channels combined by TDMA. In FDMA, the signals are identified in a receiver by their position in the frequency spectrum (bandwidth) of the transponder. In TDMA, the receiver must synchronize to the transmitted frames and then identify individual signals by their position within the frame.

possible in the near future. Usually, all of these requirements cannot be satisfied at the same time and some may have to be traded off against others. Generally, the trend in large GEO satellites has been to provide wide bandwidth high power transponders that can carry any mixture of RF signals.

There are three basic multiple access techniques. In *frequency division multiple access* (FDMA) all users share the satellite at the same time, but each uplink earth station transmits at a unique allocated frequency. This approach to sharing the frequency spectrum is familiar to us all, as it is the way that radio broadcasting has always shared the air waves. Each radio station is allocated a frequency and a bandwidth, and transmits its signals within that part of the frequency spectrum. FDMA can be used with analog or digital signals. In *time division multiple access* (TDMA) each user is allocated a unique time slot at the satellite so that signals pass through the transponder sequentially. Because TDMA causes delays in transmission, it is used only with digital signals. FDMA and TDMA are illustrated in Figure 6.9. The signals in Figure 6.9 have equal bandwidth or occupy equal time periods; in practice, different bandwidth signals can share a transponder in FDMA and signals with different durations can share a TDMA frame.

In *code division multiple access* (CDMA) all users transmit to the satellite on the same frequency and at the same time, so the signals are overlaid on one another as illustrated in Figure 6.10. The earth stations transmit coded spread spectrum (SS) signals that can be separated at the receiving earth station by correlation with the transmitted code. For example, in the global positioning system (GPS) each individual GPS satellite transmits a different coded spread spectrum signal. The signals are nearly orthogonal, allowing a GPS receiver to extract the spread spectrum signal for one satellite in the presence of similar spread spectrum signals from other visible GPS satellites. CDMA is inherently a digital technique.

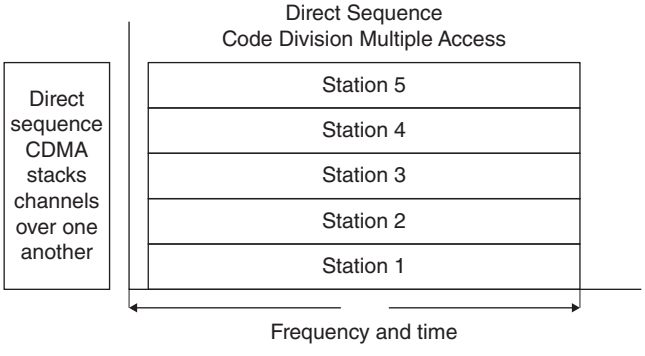


Figure 6.10 Direct sequence code division multiple access. Individual signals are recognized in the receiver by correlation with the specific CDMA code used to generate the signal.

In each of the multiple access techniques, some unique property of the signal – frequency, time, or code – is used to label the transmission such that the wanted signal can be recovered at the receiving terminal in the presence of all other signals. CDMA is much less efficient than TDMA and FDMA in terms of bits per hertz of transponder bandwidth, so its use is restricted to applications in which the unique features of CDMA are required.

In addition to the three basic techniques for multiple access, satellites that have multiple beams covering a region of the earth surface, the continental United States, for example, share the capacity of the satellite by having one or more transponders connected to each beam. Some LEO communications satellites also employ multiple beams to increase capacity. Examples of multiple beam satellites are discussed in this chapter and also in Chapters 9, 10, and 11.

Multiplexing applies to signals that are combined together into a single stream at one location, whereas multiple access refers to signals from a number of different geographical locations that pass through the same satellite transponder. The terminology of FDM is applied to any system that combines signals in the frequency domain.

An earth station can use *time division multiplexing* (TDM) to create a high speed digital data stream from many digital channels delivered to that earth station, and then modulate the data stream onto an RF carrier and transmit the carrier to the satellite. At the satellite, the carrier can share a transponder using TDMA or FDMA with other carriers from earth stations anywhere within the satellite's coverage zone. The resulting signal is called TDM-TDMA or TDM-FDMA. Note the distinction between TDM and TDMA: signals at one earth station are combined by multiplexing (TDM), and then share a satellite transponder with signals from other earth stations by multiple access (TDMA or FDMA). Direct to home satellite TV systems use TDM to deliver sequential packets or frames that contain digital video and audio from one or many TV programs. The bit stream made up of repeating frames comes from a single source, and is therefore an example of multiplexing in the time domain (TDM).

In all three of the classical multiple access techniques, some resource is shared. If the proportion allocated to each earth station is fixed in advance, the system is called *fixed access* (FA) or *preassigned access* (PA). If the resource is allocated as needed depending on changing traffic conditions, the multiple access technique is called *demand assignment* multiple access (DAMA). Demand assignment blurs some of the distinctions between FDMA and TDMA, since stations in a FDMA-DAMA system transmit only when they have traffic. Demand assignments with FDMA is sometimes used in very small aperture terminal (VSAT) systems, where earth stations may have traffic to send only intermittently. Fixed assignment is wasteful of transponder capacity, so demand assignment is used. Similarly, a group of earth stations may access part of the bandwidth of a transponder using TDMA, while other TDMA groups of earth stations share different sections of the transponder bandwidth. This approach has been used in both VSAT and mobile satellite systems (see Chapter 9). Demand assignment can also be used with CDMA to reduce the number of signals in the transponder at any one time. The Globalstar LEO mobile satellite system uses CDMA with demand assignment (Globalstar 2018).

Systems that combine both FDMA and TDMA techniques are sometimes called *hybrid multiple access* schemes or *multi-frequency time division multiple access* (MF-TDMA). In the sections that follow, we will first discuss FDMA, TDMA, and CDMA as fixed assignment schemes, and then cover DAMA and hybrid multiple access.

6.4 Frequency Division Multiple Access (FDMA)

The main advantage of FDMA is that filters can be used to separate signals. Filter technology was well understood when satellite communications began, and microwave filters were used in earth stations to select the signal from a given transponder. In a fixed assignment system, each transmitting earth station was allocated a frequency and bandwidth for each group of signals it wished to send.

Frequency Division Multiplexing in Analog Telephone Systems

FDMA was the first multiple access technique used in satellite communication systems. When satellite communications began in the 1960s, most of the traffic carried by satellites was telephony. All signals were analog, and analog multiplexing was used at earth stations to combine large numbers of telephone channels into a single baseband signal that could be modulated onto a single RF carrier. The technology had been developed since the 1920s for trunk telephone links using coaxial cables and microwave links. Individual telephone channels can be shifted in frequency from baseband to a higher frequency so that they can be stacked into a group of channels using *frequency division multiplexing* (FDM). The process begins by limiting individual telephone channels to the frequency range 300–3400 Hz, and then frequency shifting 12 channels to the frequency range 60–108 kHz with 4 kHz spacing between channels by generating single sideband suppressed carrier signals with 12 carrier frequencies spaced 4 kHz apart. The 12 channels occupying 60–108 kHz are known as a *basic group*. Five basic groups can be frequency shifted to the range 60–300 kHz to make a 60 channel *super group* occupying a baseband bandwidth of 240 kHz. Super groups can be stacked in the baseband to make up single signals that consist of 300, 600, 900, or 1800 multiplexed telephone channels. The analog multiplexing process described here is very efficient in its use of baseband bandwidth, allowing 60 voice channels to fit into a bandwidth of 240 kHz.

By comparison, combining 60 standard pulse code modulation (PCM) voice channels at 64 kHz using TDM results in a raw bit rate of 3.84 Mbps, and requires additional *overhead bits* to form a packet. Transmission of a 3.84 Mbps bit stream using half rate FEC coding and QPSK modulation requires a bandwidth of 4.608 MHz with SRRC filters with $\alpha = 0.2$. However, wide-band FM had to be used to transmit the FDM baseband so that the voice channels could be recovered with adequate quality with a receiver CNR around 15 dB, requiring an RF bandwidth of 4.0 MHz. Thus voice transmission using FDM/FM and PCM/TDM/QPSK required comparable RF bandwidths. The development of voice compression techniques allows the bit rate of a digital voice signal to be reduced substantially from the 64 kbps of a basic PCM voice signal, resulting in much lower RF bandwidths than could be achieved with analog systems.

The analog FDM technique is now obsolete in the United States and many other countries, but was the primary method of multiplexing telephone channels for transmission over terrestrial cable or microwave links for about 50 years. Early satellite systems used FDM to multiplex up to 1800 telephone channels into a wide baseband occupying up to 8 MHz, which was modulated onto an RF carrier using FM. The FDM-FM RF carrier was transmitted to the satellite, where it shared a transponder with other carriers using FDMA. The technique is known as FDM-FM-FDMA, and was the preferred method for the transmission of telephone channels over Intelsat satellites for more than 20 years.

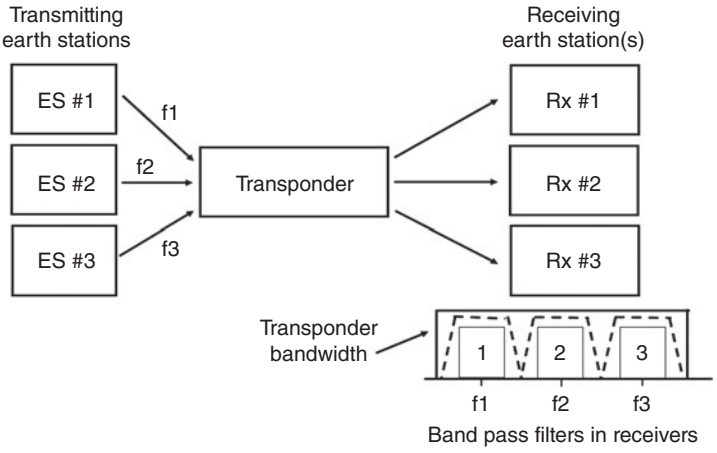


Figure 6.11 Illustration of FDMA. Three transmitting stations send signals at different carrier frequencies to a single transponder on a GEO satellite. All earth stations within the satellite footprint can receive all three signals at downlink frequencies *f*1, *f*2, and *f*3. Band pass filters in the earth station receivers select the wanted signals.

6.4.1 Implementing FDMA

Figure 6.11 shows a transponder operating with FDMA. Three transmitting earth stations send signals at different uplink frequencies to a single transponder on a GEO satellite. The transponder amplifies the received signals and retransmits them on the downlink at frequencies *f*1, *f*2, and *f*3. All earth stations within the satellite's coverage zone receive all three signals.

The three receivers shown in Figure 6.11 could be at one earth station or at three separate earth stations; in either case, BPFs centered at the frequencies *f*1, *f*2, and *f*3 are used to select the wanted transmission from within the bandwidth of the transponder. The BPFs are usually in the intermediate (IF) section of the receiver to simplify their design. For example, suppose the earth stations operate in Ku-band, the transponder has a bandwidth of 36 MHz and the downlink frequencies lie between 11.500 and 11.536 GHz. Let each of the three transmissions have a bandwidth of 10 MHz. To extract one of the three signals at a downlink frequency of 11.511 GHz requires a BPF with a fractional bandwidth of 0.085%. Filters are characterized by their Q factor, the ratio of center frequency to bandwidth. Q for this microwave filter is 1150, which is difficult to achieve. Now consider the same signal extracted with a BPF in an IF bandwidth of 36 MHz centered at 140 MHz. The IF band extends from 122 to 158 MHz and the wanted 10 MHz signal is centered at 128 MHz. The Q factor for the IF filter is now 12.8 and the filter can easily be implemented in hardware or digitally as an FIR filter. See Chapter 5 for details of digital filtering.

Figure 6.12 shows a typical fixed assignment FDMA plan for two C-band transponders. The triangles represent RF carriers with the transmitting earth station country and RF bandwidth shown inside the triangle. The signals could be video, data, or voice. Frequencies shown are for the downlink from the satellite; the triangles are not spectral diagrams and may also be shown as rectangles. The triangles represent the location of each signal within an allocated bandwidth such as that of a transponder. Transponder #1

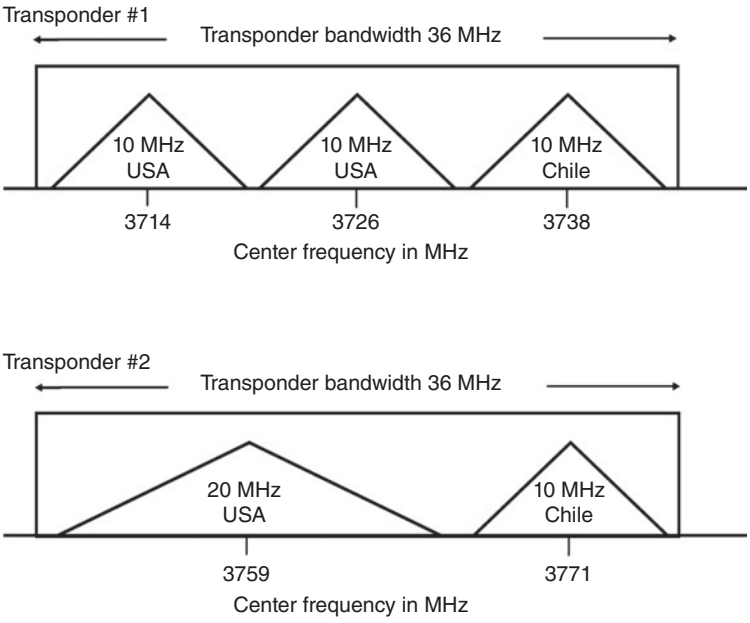


Figure 6.12 Frequency plan for two C-band transponders using fixed assignment FDMA. The triangles represent the bandwidth occupied by the signals, not power spectral density. The bandwidth and locations within the triangles relate to the transmitting earth stations and the frequencies are for the downlink transmissions from the transponders.

in Figure 6.12 receives three signals from different uplink earth stations; in this example, two are in the United Sates and one is in Chile. Each of the signals has a bandwidth of 10 MHz. The uplink signals from the two earth stations in the United States are transmitted on carrier frequencies of 5939 and 5951 MHz, and the uplink signal from the earth station in Chile is transmitted with a carrier frequency of 5963 MHz. The transponder down converts each received signal by 2225 MHz giving the downlink carrier frequencies of 3714, 3726, and 3738 MHz. All earth stations within the antenna beam connected to transponder #1 can receive all of the signals transmitted by the transponder, and each receiving earth station can extract any signals that are destined for that particular earth station.

Transponder #2 in Figure 6.12 carries two signals with different bandwidths. The 20 MHz wide signal originates from an earth station in the United States at a carrier frequency of 5984 and the 10 MHz bandwidth signal originates from an earth station in Chile at a carrier frequency of 5996 MHz. Transponder #2 down converts these signals by 2225 MHz and transmits them at carrier frequencies of 3759 and 3771 MHz. Both of these signals can be received by the same earth stations that receive signals from transponder #1. Typically, large C-band earth station receivers have front ends with a bandwidth of 500 or 1000 MHz to allow reception of all C-band carriers. Down conversion to an IF of 140 MHz, for example, allows IF filters with a bandwidth of 36 MHz to separate the signals from the two transponders. Further filtering and down conversion is needed to extract the individual carriers from each transponder, as illustrated in Figure 6.11.

The use of microwave filters to separate transponders makes the fixed assignment approach to FDMA very inflexible. Changing the frequency assignment or bandwidth of any one transmitting earth station requires retuning of the filters at several receiving earth stations. The fixed assignment FDM-FM-FDMA scheme illustrated in Figure 6.12 also makes inefficient use of transponder bandwidth and satellite capacity.

As an example, consider an earth station in the west of the United States using a Pacific Ocean GEO satellite to send telephone channels to earth stations in Korea, Japan, and Chile. The time difference between North America and the Pacific Rim countries means that the channels will be busy for only a few hours per day, and at a different time of day than the United States–Chile links. With fixed assignment, the frequencies and satellite capacity cannot be reallocated between routes, so much of the satellite capacity remains idle. Estimates of average loading of Intelsat satellites using fixed assignment were typically around 15%. It is not possible to achieve 100% loading of satellites used for international traffic, or even for domestic traffic in many cases. Demand assignment and single channel per carrier techniques allow higher loadings and therefore give satellite operators increased revenue. Fixed assignment systems are rarely used now with new satellite systems; demand assignment is preferred. The development of agile frequency synthesizers was a key factor in the introduction of demand assignment FDMA.

6.4.2 FDMA Receiver

Every earth station that operates in a FDMA network must have a separate IF receiver for each of the carriers that it wishes to receive. SCPC systems can have a very large number of carriers in one transponder; as a result, FDMA earth stations tend to have a very large number of IF receivers and demultiplexers which select individual carriers using narrowband IF filters.

Figure 6.13 shows how the IF bandwidth of a receiving earth station could be configured to receive 25 digital data channels, each with an occupied bandwidth of 1.94 MHz from a 54 MHz wide Ku-band transponder. The IF band is centered at 70 MHz requiring the BPFs that extract the individual signals to have a Q factor of 36. The 200 kHz frequency spaces between the channels are called *guard bands*. Guard bands are essential in FDMA systems to allow the filters in the receiver to select individual channels

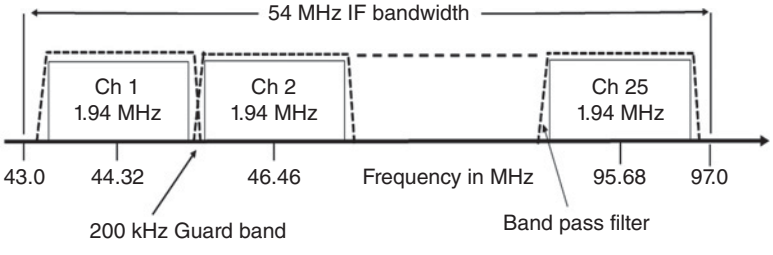


Figure 6.13 FDMA with 25 channels in a receiver IF with center frequency 70 MHz. The occupied bandwidth of each signal is 1.94 MHz and channels are spaced by guard bands of 200 kHz. Dotted lines around channels represent band pass filters that are centered at the indicated frequencies. Each channel is a T1 data signal at 1.544 Mbps with half rate FEC modulated by QPSK and $\alpha = 0.25$ SRRC filtering.

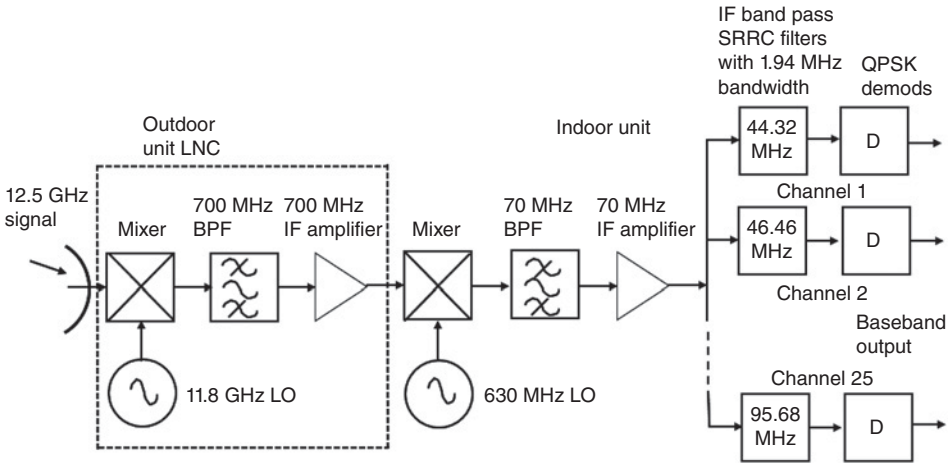


Figure 6.14 Intermediate frequency section of an FDMA receiver for 25 data channels. The first down conversion of the 11.5 GHz RF signal takes place in the low noise block converter (LNB) of the outdoor unit. The 700 MHz signal from the outdoor unit is sent to the indoor unit where a second down conversion to 70 MHz takes place. Individual SRRC receive filters set to the second IF frequency of each channel are required, followed by a QPSK demodulator and baseband processing. The spectrum of the signal at the 70 MHz IF amplifier output is shown in Figure 6.13. The entire 70 MHz section of the receiver can be implemented digitally.

without excessive interference from adjacent channels. All filters have a *roll off* characteristic, which describes how rapidly a filter can change from near zero attenuation in its pass band to high attenuation in the stop band. Typically, guard bands of 10–15% of the channel bandwidth are needed to minimize adjacent channel interference. The occupied bandwidth of a 1.94 MHz channel corresponds to a T1 data signal at 1.544 Mbps with half rate FEC coding and QPSK modulation transmitted through an SRRC filter with $\alpha = 0.25$. The channels at the highest and lowest frequencies are set a small frequency distance away from the edge of the transponder bandwidth, 350 kHz in Figure 6.13, because the phase and amplitude characteristics of the transponder pass band tend to be changing rapidly in these regions.

Figure 6.14 shows the IF portion of an earth station receiver that extracts the 25 T1 channels shown in Figure 6.13. The outdoor unit sends a wideband signal centered at 700 MHz to the indoor unit where the IF sections of the receiver are located. The Ku-band downlink frequency for the transponder is 11.5 GHz, requiring a local oscillator (LO) in the outdoor unit at 10.8 GHz to down convert the received signal to 700 MHz. A second down conversion is required with a local oscillator at 630 MHz to translate the signal to a center frequency of 70 MHz. An SRRC filter is required at the frequency of each individual channel and is followed by a QPSK demodulator at the channel IF frequency. Baseband processing to recover the individual T1 bit streams is required for each of the 25 IF channels in the receiver illustrated in Figure 6.14. The earth station may not be the destination for all of the 25 channels carried by the transponder, but the receiver is typically designed to receive all of the channels and discards those that are destined for other earth stations. The baseband T1 bit stream of any of the output channels can consist of many time division multiplexed signals, for example, 24 PCM

telephone channels at 64 kbps, 48 adaptive pulse code modulation (APCM) channels at 32 kbps, or up to 160 compressed voice channels at 9.6 kbps.

6.4.3 FDMA Transmitter

The IF section of the FDMA receiver discussed in Section 6.4.2 delivers 25 channels, each at a different IF frequency. The signals may have been transmitted by a single uplink earth station, or by up to 25 earth stations in a VSAT network. A transmitter for three T1 channels is shown in Figure 6.15, from the baseband input to the 700 MHz output that drives the RF section of the earth station. The transmitter is basically the complement of the receiver in Figure 6.14 and operates in demand assignment. The RF frequencies of the transmitted signals are assigned by a controller that tells the uplink earth station which three of the 25 RF channels to use. If a change to a different transponder is required, both the transmitting and the receiving earth stations must change their RF local oscillators in Figures 6.14 and 6.15 to new RF frequencies.

When demand assignment is provided in the transmitter and receiver, 25 IF local oscillator frequencies and 25 SRRC filters centered at the IF frequencies are required. A frequency synthesizer is needed to generate the LO frequencies. This requires a great deal of hardware, so many FDMA-DAMA links use digital signal processing (DSP) to generate the SRRC waveforms at the required IF frequencies under software control. In the transmitter in Figure 6.15, only three channels are present, so only three SRRC waveforms need to be generated. This can be done in a single (ASIC) or field programmable gate array (FPGA) instead of providing the 25 sets of local oscillators and SRRC filters required by a hardware transmitter.

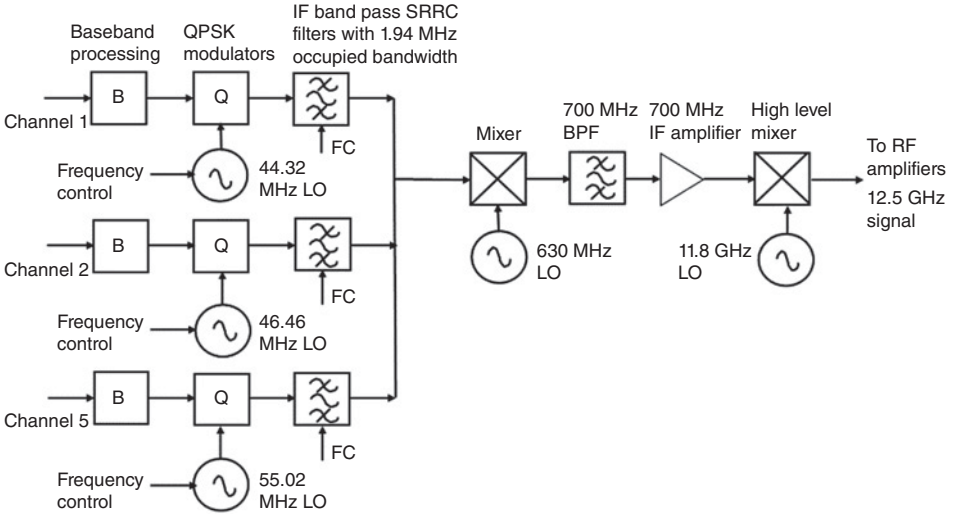


Figure 6.15 12.5 GHz FDMA transmitter for three T1 channels. The transmitter sends its signals to the receiver in Figure 6.14. The baseband processing units include the FEC stages and add headers to the T1 packets. The SRRC filters include NRZ equalization. The inputs marked *frequency control* and FC allow the IF local oscillator and SRRC filter frequency of each channel to be changed when the link uses FDMA-DAMA (demand assignment). The 700 MHz amplifier and band pass filter cover the full 54 MHz bandwidth of the transponder. Output power delivered to the RF amplifiers is typically set to 10 mW.

FDMA is widely used as a method of sharing the bandwidth of satellite transponders. In large GEO satellites with multiple downlink beam antennas, a transponder is connected to each beam and can carry a single RF carrier. If the satellite has 72 downlink beams, there may be 72 transponders with single polarization and 144 transponders if each beam has two polarizations. Alternatively, one transponder may be connected to several beams via RF filters that select the frequency band to be transmitted by each beam. The builders and operators of satellites have historically shown a strong preference for wideband transponders that can carry any type of traffic – the *bent pipe* transponder that can carry video, data, or voice as the marketplace demands. Bent pipe refers to a transponder that amplifies a signal received from the uplink and retransmits it on the downlink at a different frequency and at a higher power. By contrast, an *onboard processing* satellite has transponders that demodulate signals received from the uplink, process the signals at baseband, and then remodulate the signal onto a downlink RF carrier. Bent pipe transponders on commercial GEO satellites usually have wide bandwidths, with bandwidths of 24, 36, 54, 72, and up to 200 MHz commonly employed. When an earth station has a carrier that occupies less than the transponder bandwidth, FDMA can be used to allow that carrier to share the transponder with other carriers.

When an earth station sends one signal on a carrier, the FDMA access technique is called *single channel per carrier* (SCPC). Thus a system in which a large number of small earth stations, such as mobile telephones, that access a single transponder using FDMA is called a single channel per carrier frequency division multiple access (SCPC-FDMA) system. Hybrid multiple access schemes can use TDM of baseband channels, which are then modulated onto a single carrier. A number of earth stations can share a transponder using FDMA, giving a system known as TDM-SCPC-FDMA. Note that the sequence of abbreviations is baseband multiplexing technique first, then multiple access technique next. TDM-SCPC-FDMA is often used by VSAT networks in which the earth stations transmit many digital signals.

FDMA has a disadvantage in satellite communications systems when the satellite transponder has a non-linear characteristic. Most satellite transponders use HPAs, which are driven close to *saturation*, causing non-linear operation. A transponder using a *traveling wave tube amplifier* (TWTA) is more prone to non-linearity than one with a *solid state high power amplifier* (SSHPA). Equalization at the transmitting station, in the form of *predistortion* of the transmitted signal can be employed to linearize the transponder when fixed assignment is used. Linearization of solid state and traveling wave tube high power amplifiers (TWT HPAs) in the satellite transponder is also possible. Non-linearity of the transponder HPA causes a reduction in the overall $(\text{CNR})_o$ at the receiving earth station when FDMA is used because *intermodulation* (IM) products are generated in the transponder. Some of the IM products will be within the transponder bandwidth and will cause interference. The IM products are treated as though they were thermal noise, adding to the total noise in the receiver of the receiving earth station. Intermodulation is discussed in the following section.

6.4.4 Intermodulation

Intermodulation (IM) products are generated whenever more than one signal is carried by a non-linear device. Sometimes filtering can be used to remove the IM products, but if they are within the bandwidth of the transponder they cannot be filtered out. The saturation characteristic of a transponder can be modeled by a cubic curve to illustrate the

generation of *third order* intermodulation. Third order IM is important because third order IM products often have frequencies close to the signals that generate the intermodulation, and are therefore likely to be within the transponder bandwidth.

To illustrate the generation of third order intermodulation products, we will model the non-linear characteristic of the transponder HPA with a cubic voltage relationship and apply two unmodulated carriers with equal magnitudes at frequencies f_1 and f_2 at the input of the amplifier

$$V_{\text{out}} = AV_{\text{in}} + b(V_{\text{in}})^3 \text{ volts} \tag{6.14}$$

where $A \gg b$.

The amplifier input signal is

$$V_{\text{in}} = V_1 \cos(\omega_1 t) + V_2 \cos(\omega_2 t) \text{ volts} \tag{6.15}$$

The amplifier output signal is

$$V_{\text{out}} = \underbrace{AV_1 \cos(\omega_1 t) + AV_2 \cos(\omega_2 t)}_{\text{linear term}} + \underbrace{b[V_1 \cos(\omega_1 t) + V_2 \cos(\omega_2 t)]^3}_{\text{cubic term}} \tag{6.16}$$

The linear term simply amplifies the input signal by a voltage gain A. The cubic term, which will be denoted as $V_{3\text{out}}$ can be expanded as

$$V_{3\text{out}} = b[V_1 \cos(\omega_1 t) + V_2 \cos(\omega_2 t)]^3$$

$$V_{3\text{out}} = b\big[{V_1}^3 \cos^3(\omega_1 t) + {V_2}^3 \cos^3(\omega_2 t) + 3{V_1}^2 \cos^2(\omega_1 t) \times V_2 \cos(\omega_2 t) + 3{V_2}^2 \cos^2(\omega_2 t) \times V_1 \cos(\omega_1 t)\big] \tag{6.17}$$

The first two terms in Eq. (6.17) contain frequencies $f_1, f_2, 3f_1$, and $3f_2$. The triple frequency components can be removed from the amplifier output with a band pass filter. The second two terms generate the third order IM frequency components.

We can expand the cosine squared terms using the trig identity

$$\cos^2 x = \frac{1}{2}[\cos(2x) + 1]$$

Hence the IM terms of interest become

$$V_{\text{IM}} = \frac{3}{2} b\, {V_1}^2\, V_2 [\cos(\omega_2 t)\, \cos((2\omega_1 t) + 1)] + \frac{3}{2} b\, {V_2}^2\, V_1 [\cos(\omega_1 t)\, \cos((2\omega_2 t) + 1)]$$

$$V_{\text{IM}} = \frac{3}{2} b\, {V_1}^2\, V_2 [\cos(\omega_2 t) \cos(2\omega_1 t) + \cos(\omega_2 t)] + \frac{3}{2} b\, {V_2}^2\, V_1 [\cos(\omega_1 t) \cos(2\omega_2 t) + \cos(\omega_1 t)] \tag{6.18}$$

The terms at frequencies f_1 and f_2 in Eq. (6.18) add to the wanted output of the amplifier; the third order intermodulation products are generated by the $f_1 \times 2f_2$ and $f_2 \times 2f_1$ terms. Using another trig identity

$$\cos x \cos y = \frac{1}{2} \cos(x + y) + \frac{1}{2} \cos(x - y)$$

the output of the amplifier contains IM frequency components given by V'_{IM} where

$$V'_{\text{IM}} = \frac{3}{4} b {V_1}^2 V_2 [\cos(2\omega_1 + \omega_2)t + \cos(2\omega_1 - \omega_2)t] + \frac{3}{4} b {V_2}^2 V_1 [\cos(2\omega_2 + \omega_1)t + \cos(2\omega_2 - \omega_1)t] \tag{6.19}$$

We can filter out the sum terms in Eq. (6.19), but the difference terms, with frequencies $2f_1 - f_2$ and $2f_2 - f_1$ may fall within the transponder bandwidth. These two terms are known as the *third order intermodulation products* of the HPA, because they are the only ones likely to be present at the output of a transponder that incorporates a narrow BPF at its output. Thus the third order intermodulation products that are of concern are given by V_{3IM} where

$$V_{3IM} = \frac{3}{4} b {V_1}^2 V_2 \cos(2\omega_1 - \omega_2)t + \frac{3}{4} b {V_2}^2 V_1 \cos(2\omega_2 - \omega_1)t \tag{6.20}$$

The magnitude of the IM products depends on the parameter b, which describes the non-linearity of the transponder, and the magnitude of the signals. The wanted signals at the transponder output, at frequencies f_1 and f_2, have magnitudes AV_1 and AV_2. The wanted output from the amplifier is

$$V_{out} = AV_1 \cos(\omega_1 t) + AV_2 \cos(\omega_2 t) \text{ volts} \tag{6.21}$$

The total power of the wanted output from the HPA, referenced to a 1 Ω load, is therefore

$$P_{out} = \frac{1}{2} A^2 {V_1}^2 + \frac{1}{2} A^2 {V_2}^2 = (P_1 + P_2) \text{ W} \tag{6.22}$$

where P_1 and P_2 are the power levels of the wanted signals. The power of the IM products at the output of the HPA is

$$P_{IM} = \frac{1}{2} \times \frac{9}{16} b^2 {V_1}^6 + \frac{1}{2} \times \frac{9}{16} b^2 {V_2}^6 = \frac{9}{32} b^2 \left({P_1}^3 + {P_2}^3\right) \text{ W} \tag{6.23}$$

It can be seen from Eqs. (6.22) and (6.23) that IM products increase in proportion to the cubes of the signal powers, with power levels that depend on the ratio $(b/A)^2$. The greater the non-linearity of the amplifier (larger b/A ratio), the larger the IM products.

Example 6.1 Intermodulation

Consider the case of a 36 MHz bandwidth C-band transponder, which has an output spectrum for downlink signals in the frequency range 3705–3741 MHz. The transponder carries two unmodulated carriers at 3718 and 3728 MHz with equal magnitudes at the input to the HPA. Using Eq. (6.20), the output of the HPA will contain additional frequency components at frequencies

$$f_{31} = (2 \times 3718 - 3728) = 3708 \text{ MHz}$$

$$f_{32} = (2 \times 3728 - 3718) = 3738 \text{ MHz}$$

Both of the IM frequencies are within the transponder bandwidth and will therefore be present in an earth station receiver that is set to the frequency of this transponder. The magnitude of the IM products will depend on the ratio b/A, a measure of the non-linearity of the HPA, and on the actual level of the two signals in the transponder.

Now consider the case where the two signals carry modulation, which spreads the signal energy into a bandwidth of 8 MHz around each carrier. Carrier #1 has frequencies 3714–3722 MHz and carrier #2 has frequencies 3726–3734 MHz. Denoting the band of frequencies occupied by the signals as f_{nlo} to f_{nhi}, the intermodulation products cover the frequency bands

$$(2f_{1lo} - f_{2hi}) \text{ to } (2f_{1hi} - f_{2lo}) \text{ and } (2f_{2lo} - f_{1hi}) \text{ to } (2f_{2hi} - f_{1lo})$$

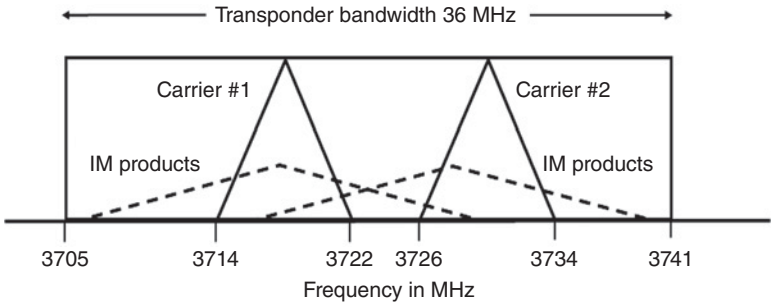


Figure 6.16 Illustration of intermodulation with two C-band carriers in a non-linear transponder HPA. Each carrier is 8 MHz wide and generates third order intermodulation products over a 24 MHz band that interfere with both signals. The triangles are not power spectral density plots; they simply represent the portion of the spectrum occupied by a signal or IM product.

The IM products are spread over bandwidths $(2B_1 + B_2)$ and $(2B_2 + B_1)$ where B_1 and B_2 are the bandwidths of the two signals in the transponder. Hence the third order IM products for this example cover these frequencies:

3706–3730 MHz and 3716–3740 MHz with bandwidths of 24 MHz.

The location of the 8 MHz wide signals and 24 MHz wide IM products is illustrated in Figure 6.16. The intermodulation products now interfere with both signals, and also cover the empty frequency space in the transponder.

Third order IM products grow rapidly as the output of the transponder increases toward saturation. Equation (6.23) shows that IM power increases as the cube of signal power; in decibel units, every 10 dB increase in signal power causes a 30 dB increase in IM product power. Consequently, the easiest way to reduce IM problems is to reduce the level of the signals in the HPA. The output power of an operating transponder is related to its saturated output power by *output backoff*. Backoff is measured in decibel units, so a transponder with a 100 W rated (saturated) output power operating with an output power of 50 W has output backoff of 20 dBW − 17 dBW = 3 dB. Intermodulation products are reduced by 9 dB when 3 dB backoff is used, so any non-linear transponder carrying more than one signal will usually have some backoff applied. Since a transponder is an amplifier, the output power level is controlled by the input power, and there is a saturated input power level corresponding to the saturated output level. When the transponder is operated with output backoff, the power level at its input is reduced by the *input backoff*. Because the transponder characteristics are not linear, input backoff is always larger than output backoff. Figure 6.17 illustrates the operating point and input and output backoff for a transponder with a non-linear TWTA. The non-linearity of the transponder causes the input and output backoff values to be unequal. In the example shown in Fig. 6.17, the transponder saturates at an input power of −68 dBm (−98 dBW). The transponder is operated at an input power of −72.0 dBm, giving an input backoff of 4 dB. The corresponding output backoff is 2.0 dB, giving an output power of 18 dBW (63 W), 37 W below the saturated output power of 100 W.

Transponder non-linearity is illustrated in Figure 6.17 for a transponder employing a TWTA as the HPA stage. A compensating equalizer can be incorporated ahead of

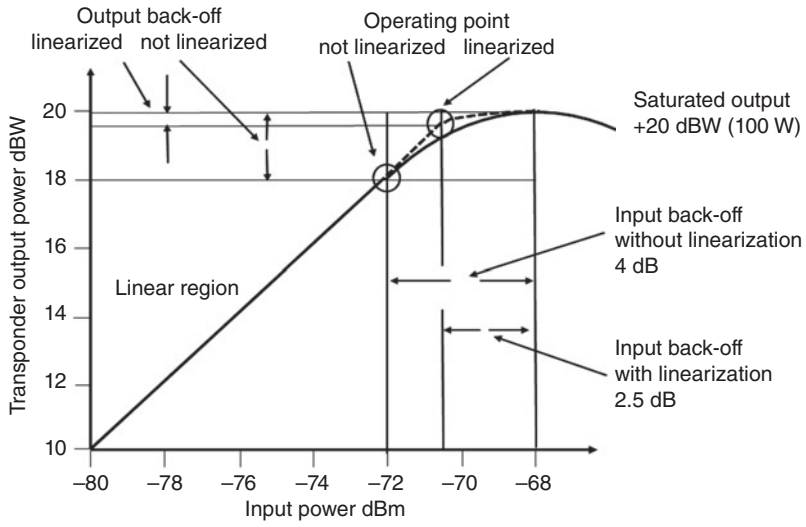


Figure 6.17 Illustration of TWTA HPA transponder non-linear characteristic. The saturated output power of the transponder is 100 W. Two operating points are shown. Without linearization, 2 dB output backoff is required to achieve quasi-linear operation and maximum output power is 63 W. With linearization applied, output back off can be reduced to 0.5 dB increasing the maximum output power to 90 W. Operating the transponder with 2 dB output back off and linearization applied will lower the 3IM products substantially compared to non-linearized operation.

the TWTA to partially linearize its characteristic giving the dashed line shown in Figure 6.17. Quasi-linear operation at the operating point indicated in Figure 6.17 without linearization requires 2 dB of output backoff with 4 dB of input backoff, providing a linear transponder gain of 90 dB and an output power of 18 dBW (63 W). Inclusion of an equalizer in the transponder amplifier chain allows quasi linear operation to be extended to the operating point indicated with maximum output power level of 19.5 dBW (89 W). The corresponding input power is −70.5 dBm. When operated with multiple FDMA channels and 2 dB output backoff, the third order intermodulation products would be substantially lower when equalization is applied than without equalization.

In the example above, both carriers had equal power. If the powers are unequal, the weaker signal may be swamped by intermod products from the stronger carrier. This can be seen from Eq. (6.20); the IM products that tend to affect carrier #1 have voltages proportional to the square of the voltage of carrier #2.

Operation of a non-linear transponder with multiple carriers requires careful balancing of the power levels of each carrier so that intermodulation products are evenly spread across the transponder's bandwidth. Judicious spacing of the carriers can be used to place the highest intermods in gaps between carriers. The process is known as *loading* the transponder. Sophisticated computer programs are used by satellite operators to optimize the backoff level of a transponder such that intermodulation is minimized while output power is maximized. When a very large number of carriers access a transponder using FDMA, as might happen with a network of VSAT stations or a transponder used with mobile satellite telephones the transponder must operate in a *quasi-linear* region of its characteristics. Quasi-linear means almost linear, either by equalization or by the application of a large output backoff.

Earth station HPAs can also cause intermodulation if they carry multiple carriers and operate close to saturation. In large earth stations where multiple carriers are more likely to be transmitted, the HPA is often rated at a much higher level than the expected transmit power. This allows substantial backoff to be used, keeping the amplifier in its linear region.

In the above analysis of third order intermodulation, only two carriers were considered. If there are three (or more) carriers present in a non-linear transponder, intermodulation products at frequencies such as $f_1 + f_2 - f_3$ can be generated that are likely to be within the transponder bandwidth. When many carriers are present, as with a transponder carrying narrowband SCPC signals, there will be a very large number of IM products, making quasi-linear operation essential. Larger gaps may be needed between carriers in the transponder bandwidth to avoid intermodulation products, which reduces the capacity of the transponder.

6.4.5 Calculation of CNR With Intermodulation

Intermodulation between carriers in a non-linear transponder adds unwanted products into the transponder bandwidth that are treated as though the interference were Gaussian noise.

For wideband carriers, the behavior of the IM products will be noiselike; with narrowband carriers, the assumption may not be accurate, but is applied because of the difficulty of determining the exact nature of the IM products.

The output backoff of a transponder reduces the output power level of all carriers, which therefore reduces the CNR in the transponder. The transponder CNR appears as $(\text{CNR})_{\text{up}}$ in the calculation of the overall $(\text{CNR})_{\text{o}}$ in the earth station receiver. IM noise in the transponder is defined by another CNR, $(\text{CNR})_{\text{IM}}$, which enters the overall $(\text{CNR})_{\text{o}}$ through the reciprocal formula (using linear CNR power ratios) first encountered as Eq. (4.49) in Chapter 4.

$$(CNR)_{\text{o}} = \frac{1}{1/(CNR)_{\text{up}} + 1/(CNR)_{\text{dn}} + 1/(CNR)_{\text{IM}}} \tag{6.24}$$

There is an optimum output backoff for any non-linear transponder operating in FDMA mode. Figure 6.18 illustrates the effect of the HPA operating point on each CNR in Eq. (6.24) when the operating point is set by the power transmitted by the uplink earth station.

The uplink $(\text{CNR})_{\text{up}}$ increases linearly as the transponder input power is increased, leading to a corresponding non-linear increase in transponder output power, indicated by the dashed line in Figure 6.18. As the non-linear region of the transponder is reached, the downlink $(\text{CNR})_{\text{dn}}$ increases less rapidly than $(\text{CNR})_{\text{up}}$ because the non-linear transponder is going into saturation. Intermodulation products start to appear as the non-linear region is approached, increasing rapidly as saturation is reached. With a third order model for non-linearity, the intermodulation products increase in power at three times the rate at which the input power to the transponder is increased, causing $(\text{CNR})_{\text{IM}}$ to decrease rapidly as saturation is approached, eventually dominating CNR overall. When all three CNRs are combined through Eq. (6.24), the overall $(\text{CNR})_{\text{o}}$ in the receiving earth station receiver has a maximum value at an input power level of −73.5 dBm in the example in Figure 6.18. This is the optimum operating point for this transponder, with an output back off of approximately 4 dB. The optimum operating

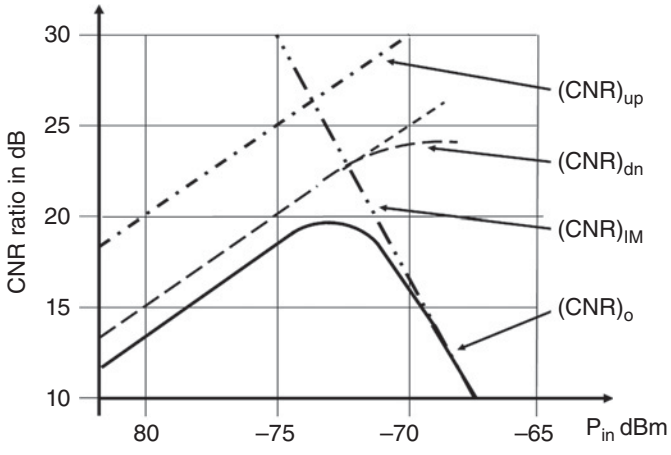


Figure 6.18 Illustration of optimization of overall CNR in a typical satellite transponder. $(CNR)_{up}$ is uplink CNR, $(CNR)_{dn}$ is downlink CNR, $(CNR)_{IM}$ is intermodulation CNR, and $(CNR)_o$ is overall CNR. Overall CNR is the CNR value measured in the earth station receiver. The combination of CNRs for the uplink, downlink, and intermodulation leads to an overall CNR that is maximized at a transponder input power of −73.5 dBm in this example. As the transponder HPA starts to saturate (shown by the dashed line for downlink CNR) the intermodulation CNR falls rapidly and eventually dominates the overall CNR.

point may be many decibels below the saturated output level of the transponder when a large number of carriers are present (Maral and Bousquet 2002, pp. 249–253).

VSAT networks and mobile satellite telephones often use SCPC FDMA to share transponder bandwidth. Because the carriers are narrowband, in the 10–128 kHz range typically, a 36 or 54 MHz transponder may carry many hundreds of carriers simultaneously. The balance between the power levels of the carriers may not be maintained, especially in a system with mobile transmitters that can be subject to fading. The transponder must operate in a linear mode for such systems to be feasible, either by the use of a linear transponder with equalization to make it linear, or by applying large output backoff to force operation of the transponder into its linear region.

6.4.6 Power Sharing in FDMA

Intermodulation between multiple carriers in a satellite transponder is minimized when each signal in the transponder has the same power spectral density (PSD). In a GEO satellite system where the earth stations are all the same distance from the satellite, this can be achieved by making the power transmitted by each earth station proportional to the bandwidth of the transmitted signal. This approach is illustrated in Example 6.1 where a number of fixed earth stations are sharing a transponder. When earth stations are mobile or transmit to the satellite intermittently, as in electronic news gathering for example, a different approach may be needed.

When an earth station wants to transmits its signal through a transponder that is operated in FDMA and already has other signals present, a spectrum analyzer is used in a loop back test to set the transmitter power level correctly. A loop back test is where the transmitting earth station receives its own signal. The power from the transmitter is increased until the spectrum analyzer display shows that the new signal has the same

PSD as other signals already in the transponder. Alternatively, the spectrum analyzer can be located at a central control station that manages the satellite. A voice link is established between the mobile unit and the operator at the central station who instructs the mobile unit operator to slowly raise transmitter power until the required transponder PSD is achieved.

FDMA may not be the best choice for LEO satellites where the distance from the earth stations to the satellite varies a great deal resulting in wide variation in path loss. Attenuation through trees can also reduce the power reaching the satellite when the earth station is mobile. Equalizing the PSD of many channels in this situation is difficult, so TDMA may be a better choice.

Example 6.2

Three identical large earth stations with 500 W saturated output power transmitters access a 36 MHz bandwidth transponder of a GEO satellite using FDMA. The earth stations are all at the same distance from the satellite. The transponder saturated output power is 100 W and it is operated with 3 dB output backoff when FDMA is used. The gain of the transponder is 105 dB in its linear range. The bandwidths of the earth station signals are

Station A: 15 MHz
Station B: 10 MHz
Station C: 5 MHz

Find the power level at the output of the transponder, and at the input to the transponder, in dBW, for each earth station signal, assuming that the transponder is operating in its linear region with 3 dB output backoff. Each earth station must transmit 250 W to achieve an output power of 25 W from the transponder. Find the transmit power for each earth station when the transponder is operated with FDMA to make the PSD of each signal equal.

Answer
The output power of the transponder must be shared between the three signals in proportion to their bandwidths. The output backoff of 3 dB means that the output power from the transponder is P_t where

$$P_t = 10\log10_{10}\ (100/2) = 20 - 3\ \text{dBW} = 17\ \text{dBW} \rightarrow 50\ \text{W}$$

The total bandwidth used is 15 + 10 + 5 = 30 MHz. The output power must be shared in proportion to bandwidth used, so the transponder output power allocated to each earth station's signal is

$$\text{Station A}: B = 15\ MHz\ P_t = \frac{15}{30} \times 50\ W = 25.0\ W \rightarrow 14\ \text{dBW}$$

$$\text{Station B}: B = 10\ MHz\ P_t = \frac{10}{30} \times 50\ W = 16.67\ W \rightarrow 12.2\ \text{dBW}$$

$$\text{Station C}: B = 5\ MHz\ P_t = \frac{5}{30} \times 50\ W = 8.33\ W \rightarrow 9.2\ \text{dBW}$$

The transponder gain is 105 dB, in its linear range, so for linear operation the transponder input power for each earth station signal is

$$\text{Station A} : P_{\text{in}} = 14.0 - 105 = -91.0 \text{ dBW} = -61.0 \text{ dBm}$$

$$\text{Station B} : P_{\text{in}} = 12.2 - 105 = -92.8 \text{ dBW} = -62.8 \text{ dBm}$$

$$\text{Station C} : P_{\text{in}} = 9.2 - 105 = -95.8 \text{ dBW} = -65.8 \text{ dBm}$$

The effective isotropically radiated power (EIRP) at each earth station must be set to give the correct power at the input to the transponder. A single earth station must transmit 250 W = 24 dBW to achieve a transponder output of 25 W = 14 dBW. For the transponder output power levels of each signal calculated above, the earth station transmitter powers are

$$\text{Station A} : P_{\text{t}} = 24.0 \text{ dBW} \rightarrow 250 \text{ W}$$

$$\text{Station B} : P_{\text{t}} = 24.0 - 1.8 = 22.8 \text{ dBW} \rightarrow 190 \text{ W}$$

$$\text{Station C} : P_{\text{t}} = 24.0 - 4.8 = 19.2 \text{ dBW} \rightarrow 83 \text{ W}$$

Note that in the above calculations transponder and earth station output power is given to three significant figures in watts and the nearest tenth of a decibel, and received signal power is quoted in dBm rather than dBW. Quoting power levels to more decimal places is unrealistic because the accuracy of link budget calculations is never better than 0.1 dB given the vagaries of atmospheric loss and the accuracy to which antenna gain can be determined in practice. Common practice in radio communication systems is to quote transmitter power in watts or dBW unless the power is less than 1 W, in which case milliwatts and dBm are quoted. Received powers are commonly given in dBm rather than dBW.

Example 6.3 Channel Capacity With Demand Assignment FDMA

A large number of satellite telephones can access a single transponder on a LEO satellite using FDMA-DAMA. In this example, L-band frequencies are used for the uplinks and downlinks. Data transmitted from the satellite on initial access by the satellite telephone is used to set the transmit frequency and output power of the satellite telephone. The telephones transmit compressed digital voice signals using QPSK modulation and half rate forward error correction coding with an occupied bandwidth of 8 kHz and an output power level between 0.05 and 0.3 W, such that the power level at the input to the transponder is always −144 dBW = −114 dBm for any uplink signal. The resulting CNR in clear air conditions for any one signal in the transponder is 10 dB. The transponder has a bandwidth of 2.0 MHz, a gain of 134 dB, and a maximum permitted output power of 12.6 W. The center frequencies of the telephone transmitters are spaced 10 kHz apart to provide a 2 kHz guard band between each signal.

Determine the maximum number of satellite telephones that can simultaneously access the transponder. Is the transponder power or bandwidth limited? If the transponder is power limited, what change could be made to increase the number of signals the transponder carries? What effect would the change have on overall $(\text{CNR})_{\text{o}}$ for the link?

Answer

If the transponder is bandwidth limited, the maximum number of signals, N_{max}, that it could carry is the available bandwidth divided by the signal bandwidth plus the guard band width

$$N_{max} = 2000\ \text{kHz} \div 10\ \text{kHz} = 200\ \text{channels}$$

If the value of N_{max} is a fraction, it must be rounded down to the next lowest integer because we cannot send fractional signals. The power level of each signal at the input to the transponder is −114 dBm. The gain of the transponder is 134 dB, so the output power for each signal is

$$P_t = -114 + 134 = 20\ \text{dBm} = 0.1\ \text{W}$$

If we have 200 signals, each at a power level of 0.1 W, the total power at the output of the transponder is 20 W. This exceeds the maximum permitted output power of the transponder, which was set at 12.6 W. Hence the maximum number of satellite telephones that can simultaneously access the transponder is 125, and the transponder is power limited.

We can increase the number of signals in the transponder to 200, which is the maximum possible number of telephones that can share the transponder at the same time because of the bandwidth limit, by reducing the input power level by 10 $\log_{10}$ (200/125) = 2.0 dB. Then the output power from the transponder, per signal, is

$$P_t = -116 + 134 = 18\ \text{dBm} = 0.063\ \text{W or } 63\ \text{mW}$$

We can now transmit 200 signals from the transponder with a total output power level of 200 × 0.063 = 12.6 W, which meets the power limitation for the transponder. The CNR in the transponder will be reduced by 2.0 dB because the input signal is 2.0 dB weaker. Hence $(CNR)_{up}$ = 10.0–2.0 = 8.0 dB. The transponder now transmits 2.0 dB less power per signal, which will reduce the $(CNR)_{dn}$ at the receiving earth station by 2.0 dB. Hence the overall $(CNR)_o$ for the link will be reduced by 2.0 dB when the number of satellite telephones sharing the transponder is increased from 125 to 200. The lower limit for BER in a digital voice channel is 10^{-4}, which requires an overall CNR of 11.4 dB with QPSK modulation. Half rate FEC with LDPC encoding can achieve a coding gain of 9 dB, so the minimum BER criterion requires a minimum $(CNR)_o$ in the earth station receiver of 2.4 dB. This allows a margin for path attenuation caused by trees and buildings of 5.6 dB. We can assume that the link between the satellite and the earth station has a $(CNR)_{dn}$ that is much higher than 8.0 dB, so the uplink to the transponder is the critical factor in determining whether the link is viable.

Note that the power transmitted by the telephone handset can be increased by command from the receiving earth station if the received signal becomes too weak. This allows a further uplink margin of 6.8 dB when the transmit power is increased to 0.3 W. The upper limit on transmit power is set by regulations that limit the maximum absorption of microwave power by the user's head (Allnutt 2013).

6.4.7 Third Order Intermodulation Intercept Point and Compression

The non-linearity of an RF amplifier is typically characterized by two numbers: the 1 dB compression point and the third order intermodulation (3IM) point. The 3IM point is established by a *two tone test* in which two signals with equal magnitude at closely

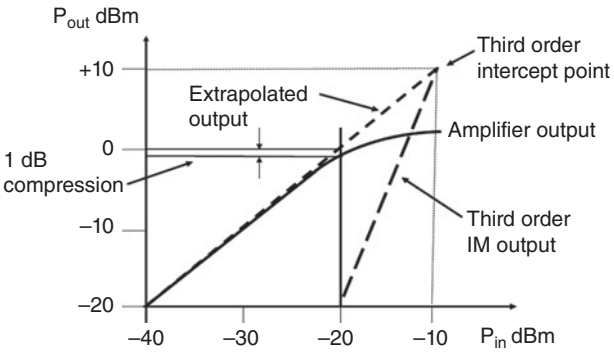


Figure 6.19 Illustration of third order intermodulation in a low power RF amplifier. The saturated output power of the amplifier is +2 dBm (1.6 mW) when the input power level is −10 dBm (0.1 mW). The linear gain of the amplifier is 20 dB, with the 1 dB compression point at an output power of −1 dBm. The third order intermodulation point (3IM point) is at an extrapolated output power of +10 dBm and an input power of −10 dBm.

spaced frequencies f_1 and f_2 are applied to the amplifier input and the output power at frequencies $f_1, f_2, 2f_1 - f_2$, and $2f_2 - f_1$ frequencies is measured.

Knowledge of these two values allows a system designer to set the operating point of the amplifier. Figure 6.19 shows an example of an RF amplifier that has a saturated output power of 2 dBm. The straight short dashed line is an extension of the amplifier's linear range beyond the point where saturation starts to occur. The solid curved line is the actual output power of the amplifier, which saturates at +2 dBm (1.6 mW) with an input power of −10 dBm. If the amplifier input is driven beyond the saturation point the output power starts to fall. The long dashed line is the third order intermodulation power generated at the amplifier output when the input consists of two identical carriers with equal power. As was seen in Section 6.4.5, IM products increase in proportion to the cubes of the signal powers. Hence for each 1 dB increase in input power the 3IM power increases by 3 dB, and the slope of the long dashed 3IM output power line in Figure 6.19 has a slope three times higher than the linear amplifier solid line. The point where the 3IM dashed line and the extrapolated amplifier linear gain line intersect is the third order intermodulation (3IM) point. The 1 dB compression point, shown by the dotted line in Figure 6.19, is where the amplifier's actual output power is 1 dB below the straight line extrapolation of the linear operating region. This occurs for an input power of −20 dBm and an output power of −1 dBm.

We can construct a graph like Figure 6.19 for any transponder or RF amplifier if we have the following information: The gain of the transponder in its linear region, the input and output saturation powers, the 1 dB compression point and the 3IM point. The graph can then be used to determine the 3IM level in the amplifier for two equal power input signals and the actual amplifier output at any input power.

Example 6.4 Calculation of Intermodulation in an RF Amplifier

The amplifier illustrated in Figure 6.19 is used in a satellite transponder to drive the high power output stage. The uplink CNR in the transponder is 20 dB. Figure 6.19 shows the characteristics of an RF amplifier with a 3IM point at an extrapolated output power of +10 dBm and an input power of −10 dBm. The 3IM characteristic in Figure 6.19 is created by plotting a line with a slope of three downward from the 3IM point.

Using the amplifier characteristics shown in Figure 6.19, find the output power from the amplifier when the input power is −20 dBm, and the gain of the amplifier. What is the C/I ratio for third order intermodulation when two equal amplitude signals at different frequencies are applied to the input? If the uplink CNR at the input of the

RF amplifier is 20 dB when only one signal is present, what is the overall CNR at the amplifier output when two equal signals at −20 dBm are applied to its input? Ignore the CNR of the amplifier itself as it is much greater than 20 dB. What is the overall CNR if the input power level is reduced to −15 dBm?

Answer
From Figure 6.19 the amplifier is being operated at its 1 dB compression point where the output power of the amplifier is −1.0 dBm with an input of −20 dBm. Hence the gain of the amplifier is 19 dB. The intermodulation power at the amplifier output is −20 dBm for an input power of −20 dBm, which is 19 dB below the amplifier output power of −1.0 dBm, giving a C/I ratio of 19 dB. Combining the uplink CNR in the amplifier of 20 dB and the C/I ratio yields an overall $(CNR)_o$ of 16 dB.

If we reduce the input power level to −15 dBm, the amplifier operates in its near-linear region with an output power of −5 dBm. The 5 dB reduction in input power produces a 15 dB reduction in intermodulation power, to an output level of −35 dBm. The C/I ratio is now 30 dB, and the overall $(CNR)_o$ becomes 19.6 dB. This is a much more satisfactory operating point for a low power amplifier in a satellite transponder.

6.5 Time Division Multiple Access (TDMA)

In TDMA a number of earth stations take turns transmitting *bursts* of RF signals through a transponder. The bit rate of a burst is determined by the bandwidth of the RF signals and the modulation. The RF bandwidth can be equal to the full transponder bandwidth that typically will create a high bit rate, or in a MF-TDMA system can be a fraction of the transponder bandwidth with a lower bit rate.

Since all practical TDMA systems are digital, TDMA has all the advantages over FDMA that digital signals have over analog. TDMA systems, because the signals are digital and can be divided by time, are easily reconfigured for changing traffic demands, are resistant to noise and interference, and can readily handle mixed video, data, and voice traffic. One major advantage of TDMA when using the entire bandwidth of a transponder is that only one signal is present in the transponder at one time, thus overcoming some of the problems caused by non-linear transponders operating with FDMA. However, using all of the transponder bandwidth requires every earth station to transmit at a high bit rate, which requires high transmitter power, making the basic form of TDMA not well suited to narrowband signals from small earth stations. TDMA can be used to assemble multiple bit streams into a single higher speed digital signal that has an RF bandwidth much less than the transponder bandwidth. Several such MF-TDMA signals can then share a transponder using FDMA. MF-TDMA is well suited to internet access systems using GEO and LEO satellites, and systems with satellite telephones and mobile video links. These systems are discussed in Chapter 11.

It is important to distinguish between TDM and TDMA. The difference between TDM and TDMA is that TDM is a baseband technique used at one location (e.g., a transmitting earth station) to multiplex several digital bit streams into a single higher speed digital signal. Groups of bits are taken from each of the bit streams and formed into baseband packets or frames that also contain synchronization and identification bits. At a receiving earth station, the high speed bit stream must first be recovered using the techniques discussed in Chapter 5, which requires demodulation of the RF carrier,

generation of a bit clock, sampling of the received waveform and recovery of the bits. The synchronization bits or words in the packets or frames must then be found so that the high speed bit stream can be split into its original lower speed signals. The clock frequency for the bit stream is fixed, and the frame length is usually constant. Packet lengths can vary however, which is the main difference between frames and packets. The entire process requires considerable storage of bits so that the original signals can be rebuilt, leading to delays in transmission. In a GEO satellite system, the largest delay is always the transmission time to the satellite and back to earth, typically 240 ms. The transmission delay is unavoidable, but any additional delays should be minimized. LEO systems have much lower delay because of the shorter path length between the satellite and earth stations.

Direct to home satellite TV systems use TDM to deliver multiple TV channels to their customers. The gateway station (or hub) collects video and audio signals from many programming sources and assembles the signals into frames creating a high speed TDM bit stream that is transmitted to a GEO satellite. Compressed digital television signals conforming to the DVB-S and DVB-S2 standards have bit rates between 1.2 and 10 Mbps depending on the program material and definition. Frames are built up from a number of signals with sufficient identification to allow individual video plus audio signals to be reassembled by the home satellite receiver. A similar approach is used for satellite radio transmissions, with much lower bit rates. Delays in satellite broadcasts caused by propagation time through a GEO satellite and buffering at the transmitter and receiver are not significant because the transmission is one way. For more details of satellite television systems see Chapter 10, Direct Broadcast Satellite Television (DBS-TV) and Radio.

6.5.1 TDMA for Fixed Networks of Earth Stations

TDMA is an RF multiple access technique that allows a single transponder to be shared in time between RF carriers from different earth stations. In a TDMA system, the RF carrier from each earth station sharing a transponder is sent as a burst at a specific time. At the satellite, bursts from different earth stations arrive sequentially, so the transponder carries a near continuous signal made up of a sequence of short bursts coming from different earth stations. The principle of TDMA is illustrated in Figure 6.20.

The burst transmission is assembled at a transmitting earth station so that it will correctly fit into the TDMA frame at the satellite. The frame typically has a length between 125 μs and 20 ms, and the burst from the earth station must be transmitted at the correct time to arrive at the satellite in the correct position within the TDMA frame. This requires synchronization of all the earth stations in a TDMA network, adding considerable complexity to the equipment at the transmitting station compared to FDMA. Each station must know exactly when to transmit, typically within one or 2 μs, so that the RF bursts arriving at the satellite from different earth stations do not overlap. (A time overlap of two RF signals is called a *collision* and results in data in both signals being lost. Collisions must not be allowed to occur in a TDMA system.)

A receiving earth station must synchronize its receiver to each of the sequential bursts in the TDMA signal and recover the transmission from each uplink earth station. The transmissions are then broken down to extract the data bits, which are stored and reassembled into their original bit streams. The individual transmissions from different uplink earth stations are usually sent using QPSK or higher order modulation, and may have small differences in carrier and clock frequencies, and different carrier phases. The

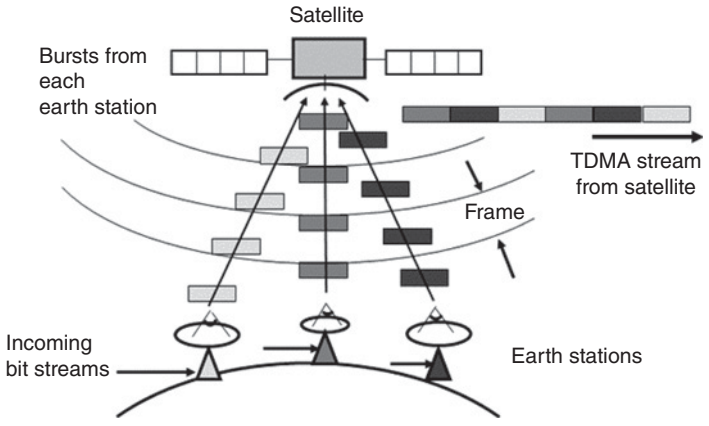


Figure 6.20 Concept of TDMA. Earth stations in a TDMA network transmit bursts at specific times such that the bursts arrive at the satellite in sequence. A TDMA stream is formed in the satellite transponder that is transmitted to receiving earth stations.

receiving earth station must synchronize its PSK demodulator to each burst of signal within a few microseconds, and then synchronize its bit clock in the next few microseconds so that a bit stream can be recovered. In high speed TDMA systems, operating at 120 Mbps for example, these are demanding requirements. MF-TDMA with its lower bit rates is an attractive alternative.

6.5.2 Bits, Symbols, and Channels

A potential source of confusion in the discussion of TDMA systems is that QPSK and 8-PSK modulations are frequently used by transmitting earth stations, and data rates can then be described either by bit rate or symbol rate. Both bit rates and symbol rates need to be used in the discussion of digital radio transmission and TDMA systems, so the reader must be clear on the distinction between a bit and a symbol.

A bit is the fundamental unit in digital transmission. Data are generated by terminals (e.g., a personal computer) as bits, or by conversion of an analog speech or video signal to digital form as a serial bit stream. The bit stream is described by its bit rate, in bits per second, bps, thousands of bits per second, kbps, or millions of bits per second, Mbps. (Note that the k and M prefixes are in units of 10^3 and 10^6, not the binary digital version of 1024 and 1 048 576.) The bit stream must be modulated onto an RF carrier for transmission to the satellite. PSK is invariably used as the modulation technique. In QPSK, two bits at a time are converted into one of four phase states of the RF carrier (see Section 6.1 for details of QPSK).

The state of the RF carrier is called a *symbol*, and the symbol rate is in units of bauds, or symbols per second. For BPSK, bit rate and symbol rate are the same, for QPSK the symbol rate is one half the bit rate, and for 8-PSK the symbol rate is one third of the bit rate. The importance of symbol rate in any digital radio system is that it is the symbol rate, not the bit rate, which determines the bandwidth of the RF signal, and consequently the bandwidth of the filters in the transmitter and receiver.

6.5.3 TDMA Frame Structure

A TDMA frame contains the signals transmitted by all of the earth stations in a TDMA network, or all of the earth stations in one MF-TDMA group. A frame typically has a fixed length, and is built up from the burst transmissions of each earth station, with *guard times* between each burst. The frame exists only in the satellite transponder and on the downlinks from the satellite to the receiving earth stations. The frame structure can differ greatly between different satellite communication systems depending on whether the satellites are GEO or LEO, whether the data has a high bit rate or a low bit rate, and whether the system has fixed or mobile earth stations. The following discussion relates primarily to earth stations that have fixed locations, communicating via satellites in GEO, and data rates that are relatively high. Chapter 8 on low throughput systems and LEO satellites discusses how TDMA is applied in those systems and Chapter 11 on internet access describes TDMA for that application.

Figure 6.21 shows a simplified diagram of a generic TDMA frame for four transmitting earth stations. Each frame contains synchronization and other data essential to the operation of the network, as well as data. Each earth station's transmission is followed by a guard time to avoid possible overlap of the following transmission. In GEO satellite systems, frame lengths of 125 μs up to 20 ms have been used, although 2 ms has been widely used by stations using Intelsat satellites. Earth stations must be able to join the network, add their bursts to the TDMA frame in the correct time sequence, and leave the network without disrupting its operation. They must also be able to track changes in the timing of the frame caused by motion of the satellite toward or away from the earth station. GEO satellites are never in a perfectly circular orbit above the earth's equator. The orbit always has some ellipticity and inclination, resulting in variation of the distance from an earth station to the satellite. Each earth station must also be able to extract the data bits and other information from burst transmissions of other earth stations in the TDMA network. The transmitted bursts must contain synchronization and identification information that help receiving earth stations to extract the traffic portions of the frame without error.

These goals are achieved by dividing TDMA burst transmissions into two parts: a *preamble* or *header* that contains a synchronization waveform, identification bits, and

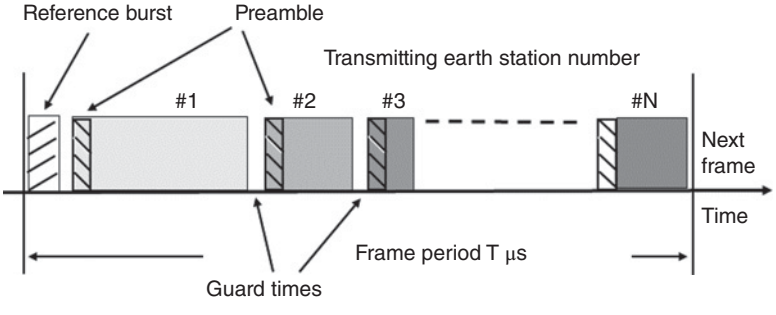


Figure 6.21 A TDMA frame with *N* earth stations in the network. The reference burst is used to mark the start of the frame and may be omitted. Short guard times are required between bursts to allow for small variations in the time of arrival of bursts and motion of the satellite. The shaded block at the beginning of each earth station burst is a preamble.

control bits, and a *traffic* portion containing data bits. Synchronization of a TDMA receiver is achieved with the portion of the frame that contains carrier and bit clock synchronization waveforms. In some systems, a separate *reference burst* may be transmitted by one of the stations, designated as the master station. A reference burst is a preamble followed by no traffic bits. The control bits in a preamble contain information for each earth station in the network to assist the station in timing its transmissions correctly. Traffic bits are the revenue producing portion of each frame, and the preamble and reference bursts represent *overhead*. The smaller the overhead, the more efficient the TDMA system, but the greater the difficulty of acquiring and maintaining network synchronization.

The preamble of each station's burst transmission requires a fixed transmission time. A longer frame contains proportionally less preamble time than a short frame, so more revenue producing data bits can be carried in a long frame. Early TDMA systems were designed around 125 μs frames, to match the sample rate of digital speech in telephone systems, in exactly the same way that T1 24 channel systems operate. A digital telephone channel generates one 8-bit digital word every 125 μs (8 kHz sampling rate), so a 125 μs frame transmits one word from each speech channel. However, it is more efficient to lengthen the frame to 2 ms or longer so that the proportion of overhead to message transmission time is reduced. It must be remembered that a longer frame requires multiple 8-bit words when transmitting digital speech. For example, in a time period of 2 ms, a digital terrestrial channel will deliver sixteen 8-bit words to a transmitting earth station, so a 2 ms TDMA frame requires sixteen 8-bit words (128 bits) from each terrestrial channel to be sent in each transmitted burst. GEO satellites are not widely used for telephone traffic now, with use restricted to places that are not served by optical fiber cables such as small islands in a large ocean.

Figure 6.22 shows a generic TDMA burst from one earth station. All bursts start with a preamble or header. In Figure 6.22 CBTR stands for carrier and bit timing recovery, often 176 symbols in duration, formed of a period of unmodulated carrier to synchronize the locally generated carrier that drives the demodulator in the receiver, and a sequence of modulated symbols that are used to synchronize the receiver bit clock. Once the demodulator is synchronized, the demodulator can output bits and the bits are used to synchronize the receiver bit clock. The next symbols in the burst are a *unique word* (UW), typically 16–64 bits that are used to identify the transmitting earth station and to determine whether the demodulator locked up correctly. A transmitting station identifier (address) may be added if all transmitting stations use the same unique word. The next block in the burst is for control, marked CNTL in Figure 6.22, and can take many forms. Information in the control block includes instructions for the receiver such as the modulation and FEC applied to the preamble and traffic segments, the length of the traffic burst, and warnings of any changes that will occur in the next frame. There may be a forward error correction (FEC) segment at the end of the preamble that can be used by both the transmitting and receiving stations to ascertain whether the preamble was received correctly. Errors in the preamble can result in the traffic section of the burst being corrupted, requiring a retransmission of the entire frame. For example, in the DBS-S2 standard for satellite television very powerful forward error correction coding is applied to header information and a different FEC rate can be selected for traffic bits (ETSI 2009). See Chapter 10 for details.

QPSK carrier recovery can result in ambiguity if the carrier recovery circuit locks up in the incorrect phase, which is possible in most QPSK demodulators. When this happens,

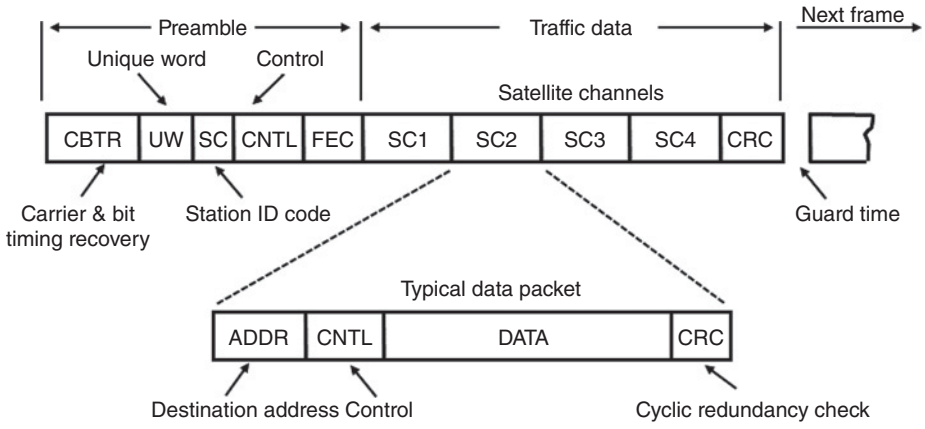


Figure 6.22 Typical structure of a transmitted TDMA burst in a network of large earth stations. The burst starts with a preamble that contains synchronization symbols (CBTR), a unique word (UW), a station identification word (SC), a control segment (CNTL), and possibly a forward error correction segment (FEC) for the preamble. The preamble is followed by traffic data formed from a sequence of packets addressed to different receiving earth stations. Each packet forms a satellite channel (SC*n*) with a header that contains address (ADDR) and control information (CNTL). Both the frame and the packet may end with a checksum (CKS) or cyclic redundancy check (CRC), used by earth stations as a final check that the entire frame and all packets were received correctly.

one or both of the I and Q bit streams is inverted. A known bit sequence is required in the received signal for ambiguity resolution, called a *unique word*. The pattern of ones and zeroes in the unique word allows the receiver to check for phase ambiguity and to invert the appropriate bit stream (I, Q, or both) if ambiguity is found. The unique word correlator functions in exactly the same way as a baseband correlator in a direct sequence spread spectrum (DSSS) receiver. See section 6.14 for a description of the process.

When a new RF burst is received, the carrier recovery circuit locks up the local carrier PLL, and the bit clock then synchronizes to the bit rate of the received signal. Bits then begin to flow into the correlator, which detects one of the four possible forms of the UW and sets logical inverters that invert the appropriate bits, if necessary. The resulting bit stream after the end of the UW is then output correctly and can be used by the receiver. The end of the UW sequence marks a known point in the TDMA burst. This time is critical, because all subsequent bits from the demodulator will be demultiplexed based on a count that begins when the UW is detected. If the UW is detected at the wrong time, the recovered data in the entire burst will be scrambled and the burst is lost. The UW and the correlator circuits must therefore be designed to ensure that the UW is detected correctly in every burst with a very low probability of a timing error. An incorrectly detected UW is known as a *miss*, and the probability that a miss occurs can be calculated from the bit error probability (BER) of the recovered bit stream and the length of the UW (Maral and Bousquet 2002, pp. 296–297). A *false alarm* can occur if a unique word sequence happens to occur within the traffic data when an earth station is trying to achieve synchronization of a TDMA burst. Once the time position of a UW within the TDMA frame is determined, a *window* can be placed over the UW so that the correlator is operated only during a period slightly longer than the UW duration. This will greatly reduce the chances of a false alarm. Use of a long unique word reduces

the likelihood of false alarms. For further details of unique word detection and TDMA burst design in satellite TDMA systems, the reader is referred to one of the references (Maral & Bousquet 2002; Tri Ta Ha 1990).

6.5.4 Calculating Earth Station Throughput With TDMA

Throughput is defined as the rate at which traffic bits are received at an earth station. If no preambles, headers, or guard times were used in transmitted TDMA frames, throughput at each earth station in a TDMA network would be equal to the transmitted bit rate divided by the number of earth stations. This is the maximum possible rate at which traffic data can be received by a TDMA earth station. Because time for traffic transmissions is lost to overhead, throughput is always less than the maximum possible rate. We can determine the throughput for one of N receiving earth stations in a TDMA frame shared equally by the N earth stations with a transmission bit rate of R_b bps. The maximum possible bit rate at any one earth station is R_b/N, which is a useful value for checking that throughput has been calculated correctly. The duration of the TDMA frame is T_{frame} in seconds, the guard time and preamble times are t_g and t_{pre}, in seconds, and the time, T_d, available to each station burst for transmission of traffic bits is

$$T_d = [T_{frame} - N(t_g + t_{pre})]/N \text{ seconds} \tag{6.25}$$

The number of frames transmitted each second is M where

$$M = \frac{1}{T_{frame}} \tag{6.26}$$

In one second, the total number of bits, C_b, transmitted by each earth station is

$$C_b = T_d \times M \times R_b \text{ bits} \tag{6.27}$$

The traffic data transmitted by each earth station consists of packets destined for one or more receiving earth stations. Let the data rate for the packets be R_{tc} bps; this is the rate at which data arrives at the earth station and must be equal to the average data rate in the TDMA frame for that station. Then the number of packets that can be carried by each earth station is given by n where

$$n = \frac{C_b}{R_{tc}} \tag{6.28}$$

Example 6.5 TDMA in a Fixed Station Network

A TDMA network of five earth stations shares a single transponder equally. The frame duration is 2.0 ms, the overhead time per station is 20 μs, and guard bands of 5 μs are used between bursts. Transmission bursts are QPSK at 30 Msps.

Calculate the number of 256 kbps channels that each TDMA earth station can transmit, assuming that each channel is encoded with rate ¾ FEC coding. What is the efficiency of the TDMA system expressed as

$$\text{Efficiency} = \frac{100\% \times \text{Number of channels in a frame}}{\text{Maximum number of channels that could be sent in the frame}}$$

The maximum number of channels that could be sent is when the transmission is a continuous stream of data bits with no time lost in overhead or guard times.

If the frame length is increased to 20 ms, what is the new TDMA system efficiency?

Answer
With ¾ rate forward error correction the transmitted data rate

$$R_{tc} = 2 \times 256 \times \frac{3}{4} = 384 \text{ kbps}$$

Using Eq. (6.25) we can find the data burst duration for each earth station, T_d, in microseconds

$$T_d = [2,000 - 5 \times (5 + 20)]/5 = 375 \ \mu s$$

The frame duration is 2 ms, so the number of frames per second, M, from Eq. (6.26) is 500. A burst transmission rate of 30 Msps is 30 million symbols per second, and QPSK symbols carry two bits. Hence the transmitted bit rate in each burst is

$$R_b = 2 \times 30 = 60 \text{ Mbps}$$

The capacity of each earth station in bits per second is C_b where, from Eq. (6.27)

$$C_b = 375 \times 10^{-6} \times 60 \times 10^6 \times 500 = 11.25 \text{ Mbps}$$

Since these bits are sent as 500 bursts each second, the number of traffic bits P per burst is

$$P = 11.25 \times 10^6/500 = 22,500 \text{ kb}$$

The number of channels that can be carried by one earth station is found from Eq. (6.28)

$$n = \frac{C_b}{R_{tc}} = \frac{11.25}{0.384} = 29.297 \text{ channels}$$

We have to discard the fractional channel since we can send only whole channels. The 0.297 fraction of an encoded 256 kbps channel represents 114 048 bit per second or 228 bits per burst that cannot be transmitted by each earth station as part of the data stream. Each transmitting station can send an additional 228 *stuffing bits* to complete its burst to maintain a constant burst duration. The stuffing bits form an extra short packet that is discarded by the receiving earth stations.

It is always a good idea to check the answer to any problem against a reference value.

If the earth stations transmitted without any guard times or preambles, the maximum number of encoded channels that could be sent by each of the five earth stations using a transmission burst rate of 60 Mbps would be

$$n_{max} = r_b/(N \times r_{tc}) = 60 \times 10^6/(5 \times 3.84 \times 10^5) = 31.25 \text{ channels}$$

The calculated capacity per station of 29 channels is lower than the maximum number of 31 channels; the difference is the number of channels lost by the need to include guard times and overhead in the TDMA frame structure.

The efficiency of the TDMA system is

$$\textit{Efficiency} = 100 \times 29/31 = 93.55\%$$

If we increase the frame duration to 20 ms, the time available for a traffic burst increases to 3975 μs. With a transmission rate of 60 Mbps, each traffic burst contains 238 500 traffic bits.

There are 50 bursts per second with a 20 ms frame, so each transmitting station sends 11.925 Mb of traffic data each second. The new number of channels that can be sent is n where

$$n = \frac{11.925}{0.384} = 31.055 \text{ channels}$$

The capacity of the TDMA system has improved by two channels per receiving station, and is now $31/31 \times 100\%$ or 100%.

The efficiency of the TDMA system can be increased if burst transmissions are allowed to have variable length. This only works if another transmitting earth station does not have enough data to fill its 3975 μs time allocation and is willing to give the spare time to a station with more data that it wants to send. This illustrates one of the problems with fixed access multiple access systems, and applies to both FDMA and TDMA. In FDMA, spectrum goes unused when a station doesn't have traffic to send; in TDMA, time slots go unused when an earth station doesn't have sufficient traffic for a full burst. The solution is DAMA in which frequencies or time slots are requested by a station when it has traffic to send.

There are many different formats for preambles in TDMA systems, depending on the design of the particular system. Figure 6.22 illustrates a TDMA preamble structure designed for a large fixed network with high speed bit streams. A network of mobile earth terminals using TDMA would have different requirements and would require a different preamble structure. However, the earlier segments that control CBTR, phase ambiguity removal, and station identification must always be present.

A large earth station carrying high speed data must link into a terrestrial data network to deliver received bits to customers, and to receive incoming data for transmission over the satellite link. The satellite link connects two earth stations, which may be in different countries thousands of kilometers apart, each interconnecting to its own high speed terrestrial network. The individual terrestrial networks may not be synchronized, and will therefore inevitably run at slightly different rates. This makes the interconnection process difficult, because a data stream delivering bits at 1.000 01 Mbps cannot be connected directly to another data stream running at 0.999 99 Mbps. Twenty bits would have to be discarded every second if a direct connection were made. A mechanism must be developed that allows for a difference in bit rates at the two ends of the link. The usual solution is to run the bit clock of the satellite link slightly faster than the fastest of the terrestrial link clocks, and to allow additional bit slots for stuffing bits. Stuffing bits are inserted when there are no data bits available from the source because of the difference in bit rates. At the receiving end of the link, the stuffing bits are removed and the received data stream is retimed to match the outgoing terrestrial data channels. The CNTL block of the burst header is where information about stuffing bits is relayed to the receiving earth station.

6.5.5 Guard Times

Guard times must be provided between bursts from each earth station so that collisions are avoided. Earth stations must transmit their bursts at precisely the correct instant so that each burst arrives at the satellite in the correct position within the TDMA frame. This requires burst transmission timing to microsecond accuracy and tracking of the position of the burst within the TDMA frame by the transmitting earth station. Long guard times make it easier for the earth stations to avoid collisions, but waste time that

could be used to send traffic data. Typical guard times in high speed satellite networks appear to be in the range 1–5 μs. The transmission time between an earth station and a GEO satellite is approximately 120 ms. If a 2 ms frame time is used, there are 60 bursts in transit between the earth station and the satellite at any time. The bursts must arrive at the correct time to mesh between the bursts that arrive from other earth stations. If the satellite range from the earth station increases by 300 m, due to eccentricity or inclination of a GEO satellite's orbit, or E-W drift in the orbital plane, the transmission delay increases by 1 μs. Thus earth stations must monitor the guard times before and after their bursts to ensure that transmission timing is correct. In LEO satellite networks that use TDMA, range to the satellite is changing continuously and larger guard times may be needed.

6.6 Synchronization in TDMA Networks

Earth stations operating in a TDMA network must transmit their RF bursts at precisely controlled times such that bursts from each of the earth stations arrive at the satellite in the correct sequence. This poses two problems: how to start up a new earth station that is joining the TDMA network, and how to maintain the correct burst timing. If the satellite is in LEO, or if it is a GEO satellite with a rapidly changing range, each earth station will perceive a different carrier frequency and frame rate, and even a different frame length. It is usual for the bit rate of transmitted bursts to be an integer multiple of the frame rate, which means that different earth stations must transmit at slightly different bit rates.

Maintaining synchronization with the TDMA frame is easier than initial synchronization when an earth station joins a TDMA network. One station is typically designated as the master station, and may generate a reference burst to mark the start of the frame, or successive reference bursts to indicate where each earth station should positions its burst (Maral and Bousquet 2002). In large networks there may be two earth stations that transmit reference bursts for redundancy. Each of the stations within the network has a time slot within the frame, and must maintain its transmissions within that time slot. There are guard times at each end of each station's burst, which define the accuracy that the burst timing must achieve. If the guard times are 2 μs, each earth station in the network must keep its bursts timed to within 1 μs.

This is usually done by monitoring the TDMA frame at the transmitting station and adjusting the burst timing to keep the transmitted burst in the correct time slot in the frame. The start of the reference burst, or the start of the master station's preamble, marks the *start of transmit frame*, SOTF, which is the master timing mark for all transmissions. All earth stations in the TDMA network synchronize their clock timing with the SOTF marker. When an earth station monitors its own transmissions to maintain the correct burst timing, this is called *satellite loop-back synchronization*. The TDMA frame is established at the satellite, so an earth station receiving the frame must subtract the transmission delay from the satellite to the earth station to obtain the SOTF timing at the satellite. It must then transmit its bursts ahead of the SOTF by the same delay time so that the bursts arrive at the correct instant at the satellite. Knowledge of the range of the satellite from the earth station is crucial in calculating delay times. The range can be calculated from the orbital elements of the satellite, which can be determined by a control station that repeatedly measures range to the satellite. Motion of a typical GEO

satellite results in a maximum change in transmission time of 1 μs after approximately 30 seconds. Transmitting earth stations must therefore update their transmit time every second to ensure that guard times are not eroded.

There are several ways that earth stations can enter a TDMA network. In fixed networks, the precise time at which an earth station should transmit can be calculated. Provided the calculation is accurate, the earth station can transmit a preamble burst, which has no traffic, timed to fall in the center of its time slot. When the frames containing the preamble burst arrive back at the earth station, the actual position of the burst can be checked and corrections to the timing made if necessary. The station can then transmit traffic bursts with the correct timing.

With the advent of Global Navigational Satellite Systems (GNSS) such as GPS, accurate timing of TDMA transmissions in a fixed network with GEO satellites can be achieved by locking the master clock at every earth station to GPS time, or another GNSS time. The location of the satellite is known by its controlling earth station, which sends this information to TDMA master stations that then distribute the data to all uplink stations in the TDMA network. The uplink stations know their own location and calculate the delay time for a burst transmission to reach the satellite from that station. Updates of the satellite location are distributed frequently so that earth stations can maintain accurate knowledge of their transmission time slot. The master station also informs uplink stations of the start time of each frame and the position and duration of each uplink station's burst within the frame. This provides the transmitting station with all the data it needs to transmit its bursts at the correct time to ensure that the burst arrives at the satellite at exactly the right time. See Chapter 12 for details of the GPS and other GNSS systems.

In TDMA networks that lack sophisticated timing control, an earth station wishing to join the network can transmit a CDMA sequence at a low level, at an arbitrary time. The CDMA sequence will inevitably collide with another earth station's traffic burst, causing minor interference. The receiving earth station uses a correlator to determine the time at which the CDMA sequence started using exactly the same process as is used to find a unique word. However, in this case the CDMA sequence is overwritten by interference from the traffic burst with which it collided. Given a suitably long sequence, coding gain can overcome the interference and the start time of the CDMA sequence can be found. Alternatively, the transmitting station can use a shorter sequence and step the timing of the CDMA burst until it falls in the empty slot allocated to that station. The position of the pulse within the TDMA frame gives the transmitting earth station the timing information needed to transmit its bursts at the correct time.

If the signal transmitted by the satellite cannot be monitored by the transmitting earth station, *cooperative synchronization* can be used instead. This situation arises when a satellite has multiple beams, or when satellite switched TDMA is used. A TDMA burst received in one beam can be retransmitted by the satellite in another beam that does not cover the transmitting station. A master station is required that can monitor the timing of each of the earth station's bursts as they arrive at the satellite and send out instructions to the earth stations when changes in timing are needed. In the Intelsat TDMA system, the master station determines a delay time, D_N, for each earth station that gives the time between the start of a receive frame and the start of a transmit frame at that earth station. The correct transmit time is then determined by the position of earth station's burst within the transmit frame. If transmitting earth stations fall out of sync, the master station will send a do not transmit (DNTX) code to the station to tell it

to stop transmitting because serious loss of data will occur to other users of the network when a station sends its bursts at the incorrect time.

In satellite switched and multiple beam satellite systems, the cooperating control station must provide information to a new earth station that wishes to join the network. The same techniques described above can be used, but an earth station within the receiving beam must determine the timing of test transmissions and send that information to the transmitting station.

6.7 Transmitter Power in TDMA Networks

TDMA works well in fixed networks carrying high speed data streams. Transponders can be more heavily loaded because less backoff is needed with TDMA; only one RF signal is present in the transponder at any time and there is no third order intermodulation, so backoff is needed only to keep PM-AM conversion at an acceptable level. More back off is required when 16-APSK modulation is used to avoid a decrease in the voltage spacing between higher voltage levels caused by the compression characteristics of the transponder. Alternatively, predistortion of the uplink signal can be used to effectively linearize the transponder characteristic. Burst lengths can be made variable to accommodate stations that have different bit rates.

Compared to FDMA with a number of earth stations sharing the transponder bandwidth, higher uplink transmitter power is required with TDMA because every station must transmit bursts at the same high bit rate and high bit rate signals occupy a wide bandwidth. Maintaining an adequate CNR in the transponder forces the uplink earth station to use a high power transmitter. For small earth stations such as VSATs and satellite phones, this is a major disadvantage compared to SCPC-FDMA.

Example 6.6 TDMA in a VSAT Network

As an example, consider a typical VSAT earth station in the United States that is part of a TDMA network using a 54 MHz bandwidth transponder on a domestic Ku-band GEO satellite. The VSAT earth station has a 1 m antenna that transmits a single 64 kbps signal at 14 GHz. The TDMA network uses QPSK modulation and all transmitters have a symbol rate of 30 Msps. We will set $(\text{CNR})_{\text{up}}$ at 20 dB, and then calculate the required uplink transmit power. The following system parameters are used:

Earth station antenna gain is 41.5 dB, satellite antenna gain (on axis) is 32.0 dB, edge of beam loss is 3 dB, path loss at 14 GHz is 207.1 dB, receiver noise bandwidth is 30 MHz, transponder noise temperature is 500 K, and atmospheric and other losses are 1.0 dB.

The uplink power and noise budgets are given in Table 6.3. The noise bandwidth used in this calculation is the noise bandwidth of the earth station receiver, numerically equal to the symbol rate of 30 Mbps.

We require

$$(CNR)_{up} = \frac{P_r}{k\, T_{\text{s}}\, B_{\text{n}}} = 20 \text{ dB}$$

Hence

$$P_{\text{t}} - 137.6 + 126.8 = 20 \text{ dBW}$$

and P_{t} is 30.8 dBW or 1200 W.

Table 6.3 Link budget for uplink in Example 6.6

Earth station transmit power (P_t)	P_t dBW
Earth station antenna transmit gain at 14 GHz (G_t)	41.5 dB
Satellite antenna receive gain at 14 GHz (G_r)	32.0 dB
Edge of beam loss	3.0 dB
Other losses (L_{misc})	1.0 dB
Path loss at 14 GHz	207.1 dB
Power at transponder input	P_t – 137.6 dBW
Boltzmann's constant (k)	−228.6 dBW/K/Hz
Noise bandwidth 30 MHz (B_n)	74.8 dBHz
Transponder noise temperature 500 K (T_s)	27.0 dBK
Transponder input noise power (N)	−126.8 dBW

Now consider the same earth station transmitting the same 64 kbps signal in a SCPC-FDMA VSAT network using QPSK with a symbol rate of 32 ksps and a receiver noise bandwidth of 32 kHz. The uplink power budget is unchanged, but the noise power in the transponder, measured in a bandwidth of 32 kHz is −156.5 dBW.

Answer
To achieve $(CNR)_{up}$ of 20 dB in the transponder now requires an uplink transmitter power of

$$P_t = 20 + 137.6 - 156.5 = 1.1 \text{ dBW or } 1.3 \text{ W}.$$

The above example illustrates a key problem with TDMA for any small earth station: uplink transmit power. No one is going to equip a 1 m VSAT station with a 1200 W transmitter. Apart from the excessive cost, Federal Communications Commission (FCC) regulations in the United States do not allow small VSAT stations to transmit more than 4 W to limit interference to adjacent satellites.

If we change the multiple access technique for just two earth stations, so that each transmits a burst of QPSK signal at 64 ksps for half the time, the uplink transmitter power requirement is doubled to 4.1 dBW or 2.6 W. MF-TDMA is needed whenever small earth stations are using TDMA to access a wide bandwidth transponder. The early Iridium LEO system was designed to use a hybrid TDMA-FDMA multiple access scheme at L-band to combine a small number of digital telephone transmissions into a 100 kbps QPSK signal. Similar techniques are used in VSAT networks.

Example 6.7 TDMA in a Fixed Earth Station Network

In Example 6.2, three identical large earth stations shared a single 36 MHz bandwidth transponder using FDMA. The three earth stations transmit signals with powers and bandwidths given in Table 6.4.

The transponder total power output was 20 dBW with 3 dB output backoff and 105 dB transponder gain. The three earth station accesses to the transponder are changed to TDMA, with a frame length of 2.0 ms, a preamble time of 20 μs, and a guard time of 10 μs. There is no reference burst in the TDMA frame; station A provides the reference

Table 6.4 Bandwidths and power levels for three FDMA earth stations

Station	Signal bandwidth (MHz)	P_t (W)	P_t (dBW)
Station A	15	125	21.0
Station B	10	83	19.2
Station C	5	42	16.2

Table 6.5 Bit rates for three TDMA earth stations

Station	Station A	Station B	Station C
Bit rate R_b (Mbps)	15.0	10.0	5.0

time marker for the other two transmissions. The signals are transmitted using QPSK, and within the earth stations the bit rates of the signals are given in Table 6.5.

Calculate the burst duration and symbol rate for each earth station, and the earth station transmitter output power required if the transponder output backoff is set at 1.0 dB and the gain of the transponder with this output backoff is 104 dB. Compare the uplink CNR in the transponder for FDMA and TDMA operation given that station A's transmission has a $(CNR)_{up}$ of 34 dB when the transponder is operated in FDMA.

Answer

The transponder must carry a total bit rate of $15 + 10 + 5 = 30$ Mbps within the 2.0 ms frames, with 500 frames per second. Thus each frame carries $30\ \text{Mbps} \times 0.002\ \text{s} = 60$ kb. The three preamble and guard times take up $3 \times (20 + 10) = 90$ μs in each frame, leaving $2000 - 90 = 1910$ μs for transmission of data. Hence the burst bit rate is

$$R_{b\ burst} = 60\ \text{kb}/1910\ \mu\text{s} = 31.414\ \text{Mbps}$$

Since we are using QPSK for the transmissions, the burst symbol rate on the link is

$$R_{s\ burst} = 31.414\ \text{Mbps}/2 = 15.707\ \text{Msps}$$

Each of the stations must transmit at the same burst rate of 15.707 Msps. The burst lengths can be calculated from the time available in each frame for data transmission and the number of bits each station must send in a 2 ms TDMA frame. The time available for data transmission is 1910 μs, which must be shared in proportion to the number of bits N_b each station sends in a frame. The number of bits in a frame is 60 000, transmitted at a bit rate of 31 414 Mbps, so the sharing of bits and times within a frame is given in Table 6.6, rounded to the nearest higher one tenth of a microsecond. N_b is the number

Table 6.6 Bits per burst and time for data transmission per burst for three TDMA earth stations

Station A	$N_b = 30\,000$ bits	$T_A = 955.0\ \mu\text{s}$
Station B	$N_b = 20\,000$ bits	$T_B = 636.7\ \mu\text{s}$
Station C	$N_b = 10\,000$ bits	$T_C = 318.3\ \mu\text{s}$

of bits transmitted by each station within a frame and T is the duration of the traffic data portion of each station's burst.

We can easily check to see if these results are correct. The total time for which data is transmitted in any 2 m fame is the sum of the times in the Table 6.5. Adding the data burst times yields 1910 μs, as expected.

Each station must transmit at the same symbol rate of 15.707 Msps, regardless of the number of bits sent per frame. In the previous FDMA example, a transponder total output power of 20 W (13 dBW) was achieved with a total earth station power of 250 W (24 dBW) and a transponder gain of 105 dB. With TDMA, we are using a 1 dB transponder output backoff, and a transponder gain of 104 dB, so the transponder output power is now 15 dBW, an increase of 2 dB, and we have lost 1 dB of gain in the transponder. This requires an earth station output power, from each earth station, of

$$P_{t\ up} = 24 + 2 + 1 = 27\ \text{W} = 500\ \text{W}$$

TDMA requires a substantial increase in earth station transmitter power, relative to FDMA, for a low capacity earth station that joins high capacity stations in a TDMA system. In this example, a station that was transmitting 42 W to support a bit rate of 5 Mbps when the transponder was operated in FDMA must now transmit 500 W when the transponder is operated in TDMA.

The uplink $(\text{CNR})_{\text{up}}$ for station A's 15 MHz signal was 34.0 dB when the transponder was operated in FDMA. With QPSK and a burst rate of 15.707 Msps, the noise bandwidth of the earth station receiver, assuming ideal SRRC filters, will be 15.7 MHz. The output power of station A has been increased from 21 to 24 dBW, so the input power level at the transponder will also have increased by 3 dB. Hence the uplink $(\text{CNR})_{\text{up}}$ ratio in the transponder for station A signals is

$$(\text{CNR})_{\text{up}} = 34.0 + 10\log_{10}\ (15.0/15.7) + 3 = 36.8\ \text{dB}$$

Since all earth stations in a TDMA network must transmit at the same power level and with the same burst rate, and all the signals have the same noise bandwidth, the $(\text{CNR})_{\text{up}}$ ratio for each of the signals in the transponder is identical, at 36.8 dB. This is 2.8 dB higher than for the FDMA operation, but at the expense of a large increase in the uplink power from the three earth station transmitters.

6.7.1 Multiple Beam Antennas and Satellite Switched TDMA

One advantage that TDMA has when used with a satellite that has a multiple beam downlink antenna and an *onboard processing* (OBP) transponder is the option to employ satellite switched TDMA. Instead of using a single antenna beam to maintain continuous communication with its entire coverage zone, the satellite has a number of narrow antenna beams that can be used to cover the zone and to concentrate transmitted power on those regions that have the greatest volume of traffic. A narrow antenna beam has a higher gain than a broad beam, which increases the satellite EIRP and therefore increases the capacity of the downlink. Uplink signals received by the satellite are demodulated to recover the bit streams, which are structured as a sequence of packets addressed to different receiving earth stations. The satellite creates TDMA frames of data that contain packets addressed to specific earth stations within each downlink beam, and switches its transmit power and bandwidth to the direction of the receiving earth station as the packets are transmitted. Note that control of the TDMA network timing could now be

on board the satellite, rather than at a master earth station. The satellite operates in much the same way as a data router in an internet network.

In the above example, the VSAT earth station could be configured to transmit data to a baseband processing satellite using SCPC-FDMA to permit the use of a small antenna and low power transmitter. The satellite could then use satellite switched TDMA to send that data to multiple earth stations, creating a mesh VSAT network. It is difficult to create a mesh VSAT network using SCPC-FDMA. Baseband processors are considered in more detail in Section 6.10. VSAT networks are discussed in Chapter 8.

6.8 Star and Mesh Networks

The selection of a multiple access technique depends heavily of the way a satellite communication network is organized. The two basic approaches for GEO satellites are known as *star* and *mesh* networks. These are illustrated in Figures 6.23 and 6.24. The star network is the simpler configuration, and is how DBS-TV and VSAT networks with small earth stations are configured. A *Gateway* or *Hub* earth station generates all *outbound* or *forward* traffic as one or more high speed TDM bit streams consisting of packets addressed to individual receiving earth stations. The gateway station sends its uplink signals to one or more transponders on the satellite and the satellite retransmits the bit streams to all the receiving terminals in the network. In a DBS-TV system, there is only outbound traffic, and the majority of the data goes to every DBS-TV receiving terminal. However, when the DBS-TV satellite has multiple beams, local programming can be directed to different regions on earth via separate downlink beams.

In a VSAT network, each terminal looks for packets from the gateway station that contain its specific address and pulls them out of the received bit stream, while discarding all the other traffic packets. Control packets must also be extracted and checked for information about network operation. The individual terminals can transmit data back to the gateway station via the *inbound* or *reverse* link, typically using a different transponder from the outbound signals. The VSAT network is usually highly asymmetric, with the outbound link having a much higher bit rate than the inbound link, and the

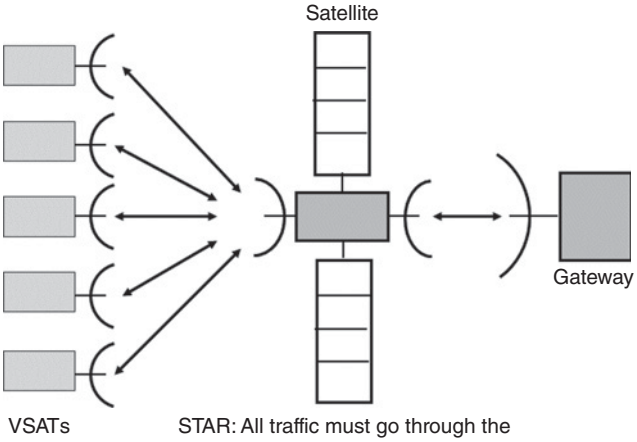


Figure 6.23 A STAR network. Outbound traffic is sent from the gateway to VSAT stations via the satellite. Inbound traffic is sent from the VSAT stations to the gateway via the satellite. VSATs cannot connect directly to one another.

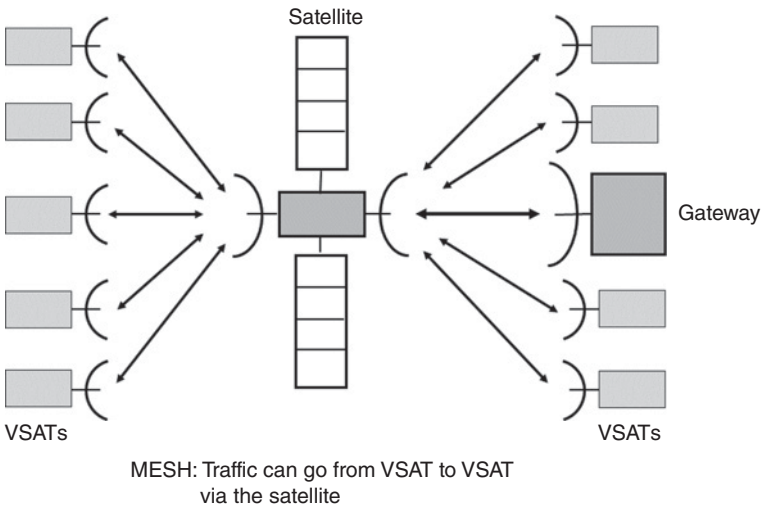


Figure 6.24 A MESH network. All stations can connect to one another, or to the gateway.

gateway station having a much more powerful transmitter than the VSAT terminals. The gateway station controls the entire network, and in a TDMA system controls the timing of bursts from the VSAT terminals. Multiple access in a VSAT star network can be SCPC-DAMA or MF-TDMA-DAMA, where DAMA indicates demand assignment. More detail on VSAT networks can be found in Chapter 8.

In a mesh network, all terminals have the same status, and control of the network is by mutual agreement between the terminals. A single frequency slot in FDMA or a single time slot in TDMA is allocated as a *request channel*. A station that wants to transmit packets looks for a vacant time slot within the TDMA frame or frequency slot in a SCPC FDMA system and sends a control packet that states the terminal's intention to use that slot for data transmission. All the other terminals make a note of the request and allow the transmitting terminal to transmit packets within set limits on the volume of data that may be sent. When the transmitting terminal has sent all of its data, a control packet is transmitted releasing the time or frequency slot. In SCPC FDMA systems, the request channel is operated in random access (RA) – see Section 6.12.

6.9 Onboard Processing

The discussion of multiple access so far has assumed the use of a bent pipe transponder, which simply amplifies a signal received from earth and retransmits it back to earth at a different frequency. The advantage of a bent pipe transponder is flexibility. The transponder can be used for any combination of signals that will fit within its bandwidth. Consider a link between a small transmitting earth station and a large hub station via a bent pipe GEO satellite transponder. There will usually be a small rain fade margin on the uplink from the transmitting station because of its low EIRP. When rain affects the uplink, the CNR in the transponder will fall. The overall CNR in the gateway station receiver cannot be greater than the CNR ratio in the transponder, so the BER at the gateway station will increase quickly as rain affects the uplink. The solution is to use

forward error correction coding on the link, which lowers the data throughput but is actually needed for less than 5% of the time. Adaptive coding and modulation (ACM) can be used to lower the CNR required on the uplink from the small transmitting earth station, with the penalty that the data rate slows down. However, it is usually preferable to maintain the connection at a lower rate than to allow an outage to occur. This approach is common in satellite internet access systems. See Chapter 11 for details.

The problem of uplink attenuation in rain is most severe for Ka-band and V-band uplinks with small margins. Outages are likely to be frequent unless a large rain fade margin is included in the uplink power budget. Onboard processing or *a baseband processing transponder* can overcome this problem by separating the uplink and downlink signals and their CNRs. The baseband processing transponder can also have different modulation schemes on the uplink and downlink to improve spectral efficiency, and can dynamically apply forward error control to only those links affected by rain attenuation. Some LEO satellites providing mobile telephone service use onboard processing, and also some Ka-band satellites providing internet access to individual users.

6.9.1 Baseband Processing Transponders

A baseband processing transponder is illustrated in Figure 6.25, consisting of a receiver, a baseband processing unit, and a transmitter. The received signal from the uplink is converted to an IF and demodulated to recover the baseband signal, which is then processed and reassembled. The baseband signal is modulated onto a carrier at a downlink frequency and transmitted back to earth. Baseband processing allows control information to be extracted from the uplink signal so that individual packets or frames can be routed to different downlink beams, and different modulation and coding can be applied on the downlinks. It is possible to establish virtual circuits between the gateway station and each individual receiving terminal, making every link fully adaptive in terms of data rates, modulation, FEC coding, and transmitted power. With multiple beam satellites this approach led to GEO satellites launched after 2010 with capacities that were a tenfold increase over earlier generations (ViaSat 2017). See Chapter 11 for further details on how baseband processing and multiple beams are used in GEO satellites for internet

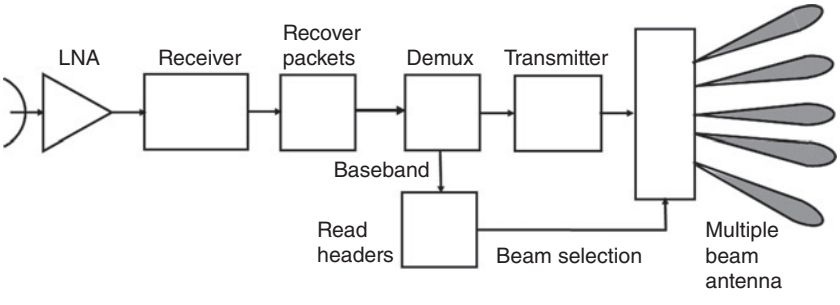


Figure 6.25 Schematic diagram of a baseband processing transponder with a multiple beam antenna. The receiving antenna, low noise amplifier (LNA) and receiver are similar to a bent pipe transponder, but the signals are demodulated and individual packets recovered at baseband. The headers are extracted from the packets and the address of the receiving earth station is used to send beam selection instructions to the multiple beam downlink antenna. Baseband packets are passed to the transmitter for transmission by the selected beam.

access. The multiple beam antenna can consist of a reflector with multiple feeds, or a phased array. With a phased array antenna one or more beams can be switched to serve different zones on earth based on the beam selection information in the frame or packet header.

With onboard processing on the satellite, the CNRs of the uplink and downlink are not tied together through the reciprocal CNR formula, Eq. (6.24). BERs on the uplink and down link add, which results in a lower BER at the receiving earth station than when noise powers add in the reciprocal formula. If the CNR on the uplink is low, because of an uplink rain fade for example, bit errors will be present in the recovered data in the transponder. If the CNR on the downlink is high, as is usually the case for star networks working with a large gateway station, no additional bit errors will occur on the downlink and the BER will depend only on the uplink CNR.

Separation of the uplink and downlink signals allows different modulation methods to be used, as well as flexible error correction codes. In star networks, the CNR of the uplink and downlink between the satellite and the gateway station is usually high because of the large antenna gain and high transmit power of the gateway station. The high CNR can be traded for a high level modulation such as 16-APSK, or 32-APSK, which reduces the bandwidth required for the uplink and increases the spectral efficiency of the communication system. 16-APSK sends four bits per symbol, and requires only half the bandwidth of QPSK. 32-APSK sends five bits per symbol.

As an example, consider a system in which half rate forward error correction and QPSK is used on the uplink from a small earth terminal. For a message data rate of R_d bits per second, the transmitted symbol rate R_s will be equal to R_d symbols per second. An RF bandwidth of R_s $(1 + \alpha_{up})$ Hz will be required on the uplink, where α_{up} is the roll off parameter of the SRRC filter in the transmitter. On the downlink to the gateway station, where the CNR is high, 16-APSK can be used with minimal forward error correction, for example rate 9/10. The RF bandwidth required for the downlink will be $1.11 \times R_s$ $(1 + \alpha_{dn})/4$, where α_{dn} is the roll off parameter of the SRRC filter in the transponder transmitter section. If we assume the same roll off parameter α in both the uplink and the downlink transmitters and receivers, the downlink requires just over one fourth of the uplink bandwidth to send the same number of bits. The fourfold reduction in downlink bandwidth represents a considerable improvement in the spectral efficiency of the satellite system. In practice, the downlink data will be encoded with low rate FEC such as 7/8 or 9/10 rate, to ensure that all errors are corrected.

The concept of onboard processing was developed in the 1980s (Betaharon et al. 1987). Several large GEO satellites with OBP were launched in the late 1990s and early 2000s, some with a switch that allowed the signals in the satellite to bypass the OBP section and permit the transponders to revert to bent pipe mode should the OBP system not work correctly. This was the case for two of the three Spaceway® GEO satellites, originally envisaged for internet access at Ka-band. Boeing retrofitted the first two Spaceway satellites for *bent pipe* communications so that they could be used for high definition (HD) DBS-TV transmission and disabled the regenerative onboard processing of the original system that was to be used for broadband satellite communications (Spaceway 2017; Spaceway 3 2015).

The ViaSat I and II satellites launched in 2011 and 2017 used 54 wide-band bent pipe K_a-band transponders and 72 multiple antenna beams to achieve a 10-fold increase in satellite capacity over the previous generations of large GEO satellites (ViaSat 2017).

Further development of high capacity LEO and GEO satellites for internet access is anticipated (Perrot 2015).

6.9.2 Multiple Beam Satellites

The combination of multiple satellite beams and TDMA can provide a large increase in satellite capacity when OBP is employed. When the full bandwidth of a transponder is used in TDMA, at any given instant the signal is a stream of packets transmitted to a single user terminal. If the satellite has a single beam with wide coverage, most of the transmitted energy goes to users who do not need that signal. Ideally, a very narrow beam covering only the location of the intended user would be much more efficient. Limitations on the satellite antenna dimensions make this impossible, but large GEO satellites with 72–200 transponders connected to spot beams with 3 dB beamwidths of 0.5°–0.25° can increase the satellite antenna gain from typically 34 dB for coverage of North America to 56 dB for a 0.25° spot beam. A further 3 dB increase in antenna gain can be achieved if the spot beams are steerable because the center of the spot beam can be directed to the location of the intended receiving station for the duration of the transmission period. GEO satellites utilizing spot beams and TDMA such as ViaSat I and II with 72 spot beams, and SES 17 with 200 spot beams have throughputs that are 10–20 times higher than satellites with continental coverage beams (ViaSat 2017; SES 17 2017). ViaSat III claims to have 30 000 spot beams, but these are actually multiple steerable spot beams capable of pointing in 30 000 different directions (ViaSat 3 2018).

The many LEO satellite constellations proposed for internet access in the 2020 time frame use a similar approach.

Communications with commercial aircraft crossing oceans is one application for Ka-band satellites with steerable spot beams, since in cruise flight these aircraft are above clouds and rain and do not suffer the same propagation problems of terrestrial Ka-band terminals. Two-way communication is needed between the aircraft and the gateway station so that the location of the aircraft (derived from GPS signals) can be sent to the gateway station to set the satellite beam pointing. High data rate two-way communication with aircraft has been an objective for many years so that business passengers can have fast access to the internet, allowing them to continue working while in flight (Aviation week 2018). This is a lower cost alternative than trying to make aircraft go faster to reduce journey times, which requires supersonic flight and much higher fares (Concorde 2017). (The Concorde supersonic aircraft operated service across the Atlantic Ocean from 1976 through 2003, with fares that were up to 30 times higher than the lowest available on those routes, but the US$1.3B cost to manufacture 14 revenue earning aircraft and six test models was paid by the governments of France and the UK. Concorde never made a profit.)

Baseband processing is essential in satellites using satellite switched TDMA, because data packets must be routed to different antenna beams based on the address of the destination earth station. The data in such systems is always sent in packets that contain a header and a traffic section. The header contains the address of the originating station and the address of the destination earth station. When satellite switched TDMA is used, the transponder must extract the destination information and use it to select the correct downlink beam for that packet. The satellite is operating much like a router in a terrestrial data transmission system. Switched beam operation of an uplink from a small

earth station is more difficult to achieve because it requires synchronization of the earth station transmit time with the satellite beam pointing sequence, in much the same way that a TDMA uplink operates. However, the uplink can operate in a small bandwidth, which overcomes the chief disadvantage of classical TDMA – the requirement for high burst rate transmissions and high transmit power.

A satellite with multiple antenna beams can greatly increase the throughput of its transponders. Consider for example a GEO satellite providing internet access to individual users in the United States. The uplink and downlink beams at the satellite must provide coverage over an area approximately 5.5° by 2.5°, as seen from the satellite. Antenna gain and beamwidth are related by the approximate relationship

$$G = 30{,}000/\text{Product of beamwidths in degrees} \tag{6.29}$$

This limits the maximum achievable satellite antenna gain to approximately 33.4 dB.

A satellite with multiple beam capability can have much narrower beams with higher gain than a satellite with a single fixed beam. The limitation on gain is the diameter of the antenna, which must fit inside the launch vehicle shroud. For launchers available in 2017, this limit was about 5 m with the largest launchers and typically 3.5 m in less expensive launchers. Most multiple beam GEO satellites have antennas in the 2–3.5 m range.

At 20 GHz, the main downlink frequency for Ka-band, an antenna with a circular aperture of diameter 3.5 m and aperture efficiency of 65% has a gain

$$G = \eta_A(\pi D/\lambda)^2 = 55.4\,\text{dB} \tag{6.30}$$

Beamwidth is approximately 75 λ/D degrees, giving a beamwidth of 0.32°. The corresponding satellite uplink antenna for 30 GHz with a beamwidth of 0.32° and a gain of 55.4 dB has a diameter of 2.86 m. The switched beam satellite has an antenna gain 21.6 dB higher than the single beam satellite, which can be traded directly for reduction in uplink or downlink transmit power, and uplink or downlink data rate. However, the satellite must generate at least 170 beams to cover all of the United States with 0.32° beams, with a consequent increase in satellite antenna complexity. Multiple beam antennas are a feature of most Ka-band internet access satellites (ViaSat 2017; ViaSat 3 2018).

Coverage of the United States with multiple beams is not always provided uniformly.

Differences in population densities and the frequency of heavy rainfall make it advantageous to provide more system capacity to metropolitan areas, and also to provide higher link margins to areas with more frequent heavy rainfall, such as Florida and the south eastern states in the United States. In the current generation of large GEO satellites (2018) a phased array feed system combined with a reflector is used to provide multiple beams. The growth of the terrestrial optical fiber network will eventually fulfill the need for high speed access to the internet in urban and suburban areas. Where direct access to an internet service provider (ISP) is available via optical fiber, the transmission rate is likely to be higher than can be achieved with a satellite system and the cost to the user could be lower. As the fiber network spreads through metropolitan areas, internet access satellites can concentrate their service on less well populated and rural areas that are rarely served by high speed terrestrial internet access. This is particularly true for less well developed counties with lower gross domestic product, where satellite access to the internet is a viable alternative to the development of optical fiber networks. The 03B network of medium earth orbit (MEO) satellites provides internet access for populations in countries between latitude 45°S and 45°N. The name 03B stands for *the other*

three billion, the population of countries where other forms of internet access are rarely available (O3B 2017).

6.10 Demand Assignment Multiple Access (DAMA)

Demand assignment can be used in any satellite communication link where traffic from an earth station is intermittent. An example is a LEO satellite system providing links to mobile telephones. Telephone voice users typically communicate at random times, with call duration ranging from less than one minute to several minutes. As a percentage of total time, the use of an individual telephone is likely to be less than 1%. If each user were allocated a fixed channel, the utilization of the entire system might be as low as 1%, especially at night when demand for telephone channels is small. Demand assignment allows a satellite channel to be allocated to a user on demand, rather than continuously, which greatly increases the number of simultaneous users who can be served by the system. The two-way telephone channel may be a pair of frequency slots in a SCPC-FDMA-DAMA system, a pair of time slots in a TDM or TDMA system, or any combination or FDMA, TDM, and TDMA. Most SCPC-FDMA systems use demand assignment to ensure that the available bandwidth in a transponder is used as fully as possible. VSAT networks also need to employ demand assignment because individual terminals do not necessarily have sufficient data to transmit continuously.

In the early days of satellite communication, the equipment required to allocate channels on demand, either in frequency or time, was large and expensive. The growth of cellular telephone systems has led to the development of low cost, highly integrated controllers and frequency synthesizers that make demand assignment feasible. Cellular telephone systems use demand assignment and techniques similar to those used by satellite systems in the allocation of channels to users. The major difference between a cellular system and a satellite system is that in a cellular system the controller is at a base station that is close to the user and is connected by a single hop radio link. In a satellite communication system, there is always a two hop link via the satellite to a controller at the gateway earth station and there are much longer transmission delays in GEO links. In international satellite systems, the controllers are not placed on the satellites largely because of the difficulties in determining which links are in use, and who will be charged for the connection. As a result, all connections pass through a controlling earth station that can determine whether to permit the requested connection to be made, and who should be charged. In international satellite communication systems issues such as *landing rights* require the owner of the system to ensure that communication can take place only between users in preauthorized countries and zones. The presence of the signals from all destinations at a central earth station in a particular country also allows security agencies the option of monitoring any traffic deemed to be contrary to the national interest (Everett 1992).

Demand assignment systems require two different types of channel: a *common signaling channel* (CSC) and a communication channel. A user wishing to enter the communication network first calls the controlling earth station using the CSC, and the controller then allocates a pair of channels to that user. The CSC is usually operated in random access mode (see Section 6.11) because the demand for use of the CSC is relatively low, messages are short, and the CSC is therefore lightly loaded, a requirement for any random access link. Packet transmission techniques are used in demand assignment

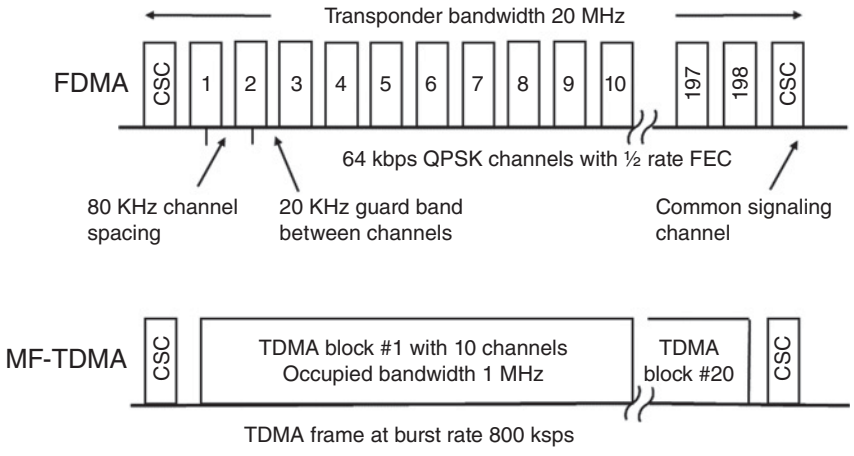


Figure 6.26 FDMA and MF-TDMA loading of a 20 MHz transponder in a VSAT star network. There are 200 channels available in each case, with two channels devoted to common signaling for requesting and releasing individual channels. The individual channels carry 64 kbps data, modulated by QPSK with half rate FEC, giving a symbol rate of 64 ksps. Occupied bandwidth for an FDMA channel with $\alpha = 0.25$ SRRC filters is 80 and a 20 kHz guard band separates channels. The TDMA frame has a duration of 20 ms and contains 1280 traffic symbols per frame from each of 10 channels. Frame burst rate is 640 ksps and occupied bandwidth is 800 kHz with $\alpha = 0.25$ SRRC filters. Headers and guard times occupy 4 ms of each frame. A 20 kHz guard band is allowed between each MF-TDMA channel.

systems because of the need for addresses to determine the source and destination of signals. Section 6.12 discusses the design of packets for use in satellite communication systems.

Bent pipe transponders are often used in demand assignment mode, allowing any configuration of FDMA of MF-TDMA channels to be adopted. There seem to be few standards for demand assignment systems in the satellite communication industry, with each network using a different proprietary configuration. Figure 6.26 shows an example of a frequency plan for the inbound channels of a VSAT star network using FDMA with single channel per carrier demand assignment (FDMA-SCPC-DAMA) or MF-TDMA.

6.10.1 FDMA-SCPC Operation

When operated in FDMA-SCPC, the individual inbound RF channels from the VSATs to the gateway station in Figure 6.26 are 80 kHz wide, to accommodate a 64 kbps bit stream with QPSK modulation, half rate FEC and SRRC filters with $\alpha = 0.25$. A guard band of 20 kHz is allowed between each RF channel, so the RF channel spacing is 100 kHz. A bandwidth of 20 MHz in the transponder can accommodate 200 of these channels, but it is unlikely that all are in use at the same time. Two channels are allocated as CSCs. Many VSAT systems are power limited, preventing the full use of the transponder bandwidth, and the statistics of demand assignment systems ensure that the likelihood of all the channels being used at one time is small. Considerable backoff is required in a bent pipe transponder with large numbers of FDMA channels, as discussed earlier in this chapter. The gateway station receiver has 200 IF receivers with 80 kHz noise bandwidth and 100 kHz frequency spacing, corresponding to the 200 FDMA VSAT channels. When a

VSAT station sends a request for a connection, the gateway station responds by allocating a transmit frequency to the station, and identifies each transmitting station by its allocated frequency. Outbound data is assumed to be delivered using one or more continuous TDM data streams.

6.10.2 MF-TDMA Operation

Ten VSAT earth stations are allocated to a TDMA group that occupies 1 MHz of the transponder bandwidth. The frame duration is 20 ms with a burst rate for each station of 800 ksps. Each VSAT burst consists of a packet with a header and traffic symbols, followed by a cyclic redundancy check (CRC) of four symbols and a guard time of 80 μs. The gateway station has 20 IF receivers with 800 kHz noise bandwidth and 1 MHz frequency spacing, corresponding to the 200 MF-TDMA VSAT channels. When MF-TDMA is used, the system has the same capacity as the SCPC-FDMA system, but the VSAT stations must have transmitters with approximately 10 times the EIRP of FDMA stations to achieve the same overall CNR, because the noise bandwidth of the gateway receiver channels is 10 times larger than for an FDMA receiver. This usually means that the VSAT stations require larger antennas, possibly 2 m diameter, rather than 1 m diameter, providing a 6 dB increase in antenna gain. A 4 dB increase in transmit power meets the 10 dB additional EIRP requirement.

6.10.2.1 Outbound Link

The outbound link transmits a continuous bit stream so that receivers can maintain carrier phase and bit clock synchronization. The data is organized into a sequence of packets, addressed to the receiving stations, and organized into a frame. One frame contains one packet for each receiving earth station, similar to the packet illustrated in Figure 6.22. Typical packets are formed with a header that contains the address of the VSAT and control information, a traffic segment, and a CRC at the end. A CRC is similar to a checksum but can detect multiple errors in a packet. If there is no data to send to a particular VSAT, the packet will have only a header, or may be omitted from the outbound TDM transmission in a demand assignment system.

If the network illustrated in Figure 6.26 is symmetric, the outbound link must deliver 64 kbps data to each VSAT station. With 198 VSAT stations, a 20 ms frame delivers 1280 data bits to each VSAT to which must be added header bits and a CRC. If we allocate 80 bits for a header and 20 bits for the CRC, the gateway station must transmit 1380 bits per frame to each station giving a transmission requirement of 13.662 Mbps. With half rate FEC encoding and QPSK modulation, the symbol rate is 13.662 Msps, and using SRRC filters with $\alpha = 0.25$ the occupied bandwidth of the transmission is 17.078 MHz. This may well be too wide a bandwidth for small VSAT terminals, resulting in unacceptably low overall CNR. The outbound transmission can be divided into a number of FDMA groups. For example, with four FDMA groups, each group transmits at 3.662 Msps, occupying a bandwidth of 4.578 MHz, and links to 50 receiving VSAT stations.

In VSAT systems, the inbound and outbound channels may be symmetric, offering the same data rate in opposite direction. In a symmetric system the outbound TDM channel must transmit at the same bit rate as all the VSAT added together. Internet access systems are often asymmetric, because requests for information can be short but the resulting replies may be lengthy. The packet length of the TDM signal in the

outbound direction may be fixed, which suits a symmetrical network, or variable, which better suits an internet channel capable of downloading large files or video from the internet.

6.10.2.2 Common Signaling Channel

The CSCs shown in Figure 6.26 are located at the ends of the transponder occupied bandwidth. When a VSAT earth station wants to access the satellite, it transmits a *control packet* to the satellite on the CSC frequency and waits for a reply. The control packet is received by the gateway earth station and decoded. The control packet contains the address of the station requesting the connection, any other relevant data (such as a character, CP, to indicate that this is a control packet with no traffic data) and a CRC that is used in the receiver to check for errors in the packet. The control station may record both origination and destination station addresses and measure the duration of the connection in order to generate billing data. In a true demand assignment system, the control station allocates the VSAT an uplink frequency or a time slot of specified duration in the outbound TDM frame. If the gateway station has a large volume of data to send

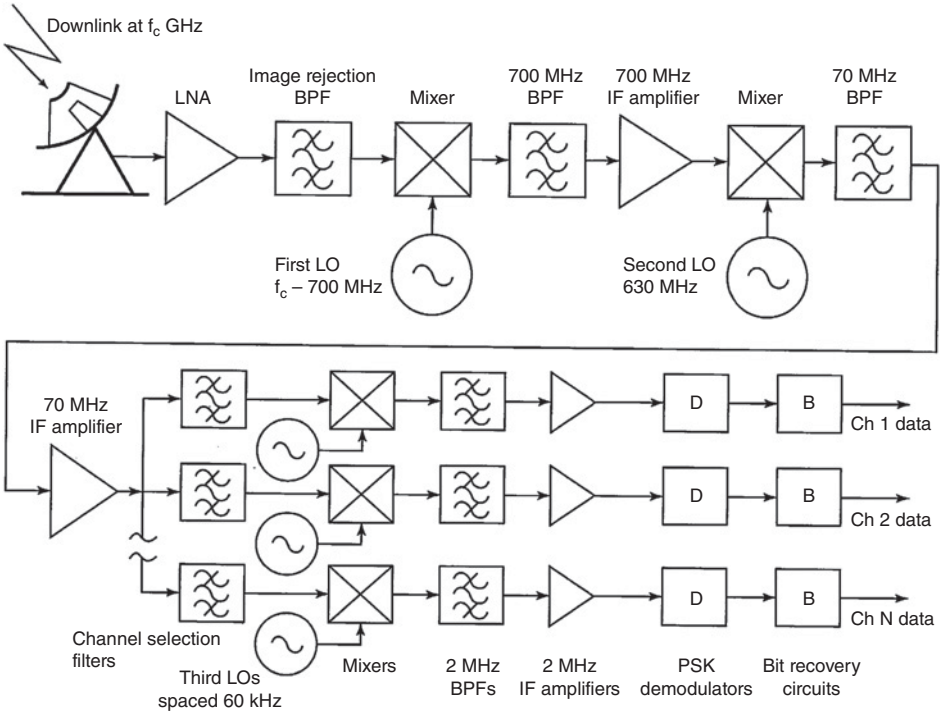


Figure 6.27 Schematic diagram of a gateway station for FDMA VSAT signals using hardware in the second and third IF sections. The third local oscillators are the outputs from a frequency synthesizer and the 2 MHz BPFs are SRRC filters. Digital signal processing can replace all the blocks after the 70 MHz IF amplifier by sampling the 70 MHz signal into I and Q channels and then implementing the second down conversion digitally, implementing the SRRC filters as FIR digital filters, and the QPSK demodulators in software. In an MF-TDMA version of the gateway receiver, the structure of the receiver is unchanged except that the 70 and 2 MHz channels have a wider bandwidth and time division demultiplexing is required after the demodulators.

to a particular VSAT station, it can allocate a longer time slot in the TDM frame to that station. This is important in internet access systems where a large file of video or other multimedia data may have to be sent. The timeslots usually come in multiples of a fixed minimum duration so that clock rates and buffer sizes are compatible. If the system becomes busy and many stations are requesting large files, throughput to any one station will slow down toward the standard minimum rate, exactly as in a terrestrial internet server.

A block diagram of a gateway receiver for the signals shown in Figure 6.26 is illustrated in Figure 6.27. The receiver amplifies and down converts the received signal to an IF of 700 MHz and then to a second IF at 70 MHz. In the hardware FDMA receiver illustrated in Figure 6.27, individual FDMA-SCPC channels within the band 60–80 MHz are down converted to a standard IF frequency of 2 MHz using local oscillators with frequencies 58–78 MHz in steps of 100 kHz. There are a total of 200 such 2 MHz IF receivers to cover all the frequency slots. A microwave frequency synthesizer is needed to generate the 200 local oscillator frequencies.

A better alternatively to building 200 hardware IF receivers is a DSP receiver where the second IF signal is split into two channels and sampled by fast ADCs driven in phase quadrature to create I and Q channels. Sampling the second IF signal at 200 MHz is required, and SRRC filtering with FIR filters centered on the 200 channel frequencies is used to extract the 200 channels. Digital QPSK demodulation of the I and Q channels for each received frequency is followed by the usual baseband processing to create an output of 200 digital signals. One or more FPGAs or ASICs can be used to replace the 200 IF receivers of the hardware version. A digital receiver for the TDMA version of the VSAT signals in Figure 6.26 has the same form as the FDMA receiver, but for 20 RF channels, each with an occupied bandwidth of 1 MHz. The output of the TDMA receiver is 20 TDMA signals with packets from 10 VSAT transmitters. The 10 TDMA signals are then separated by time division techniques to deliver 200 data channels.

6.11 Random Access (RA)

Random access is a widely used satellite multiple access technique where the traffic density from individual users is low. For example, VSAT terminals and satellite mobile telephones often require communication capacity infrequently. These users can share transponder space without any central control or allocation of time or frequency, provided the average activity level is sufficiently low. In a true random access network, a user transmits packets whenever they are available. The packet has a destination address, and a source address. All stations receive the packet and the station with the correct address stores the data contained in the packet and sends an acknowledgement back to the transmitting station. All other stations ignore the packet, unless it is designated as a broadcast packet with information for all stations, in which case no acknowledgment is sent.

In satellite communication systems, the network is more usually a star configuration, with a single gateway and many small earth stations or portable terminals. Inbound packets are received by the gateway earth station and forwarded to their destinations. Early work on random access techniques for radio channels was done at the University of Hawaii, led by Norman Abramson where the system was called Aloha (Hawaiian for hello) and was known by the generic term *packet radio* (Alohanet 2017). As more

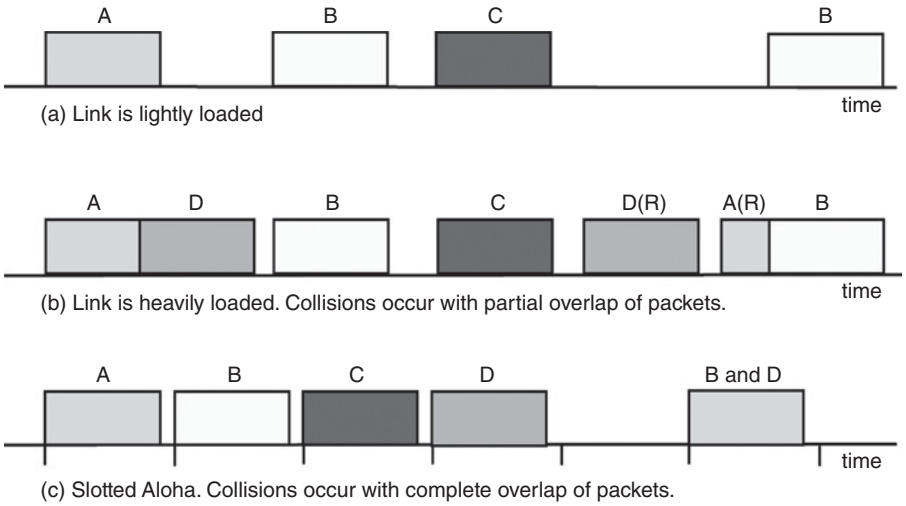


Figure 6.28 Illustration of Aloha and slotted Aloha. Letters indicate the transmitting station. In (a) the link is lightly loaded and no collisions occur. In (b) the link is heavily loaded and partial overlap of packets from stations A and D occurs. This requires retransmission of both A and D packets. The D packet is successfully transmitted as D(R), but a collision occurs between the A(R) repeat packet and a B packet, which will require retransmission of both packets. In (c) slotted Aloha is used and packets must arrive in regular time slots. A collision occurs between packets B and D requiring retransmission of both packets.

transmitters attempt to share the same transponder, collision between packets will occur. Transmitting stations either monitor their own transmissions and can therefore determine when a transmitted packet has been corrupted by a collision, or wait to receive an acknowledgement from the receiving station. If no acknowledgement arrives within a set time, a repeat transmission of the packet is made. A retransmission of the corrupted packet is required whenever a collision is detected; the transmitter waits for a random time and then retransmits the packet. The capacity of the random access system is maximized at 18% when an average of 2.7 retransmissions of each packet is required (Maral and Bousquet 2002, pp. 322–326). Efficiency of the random access system can be improved to 36% using slotted Aloha if transmitters are allowed to transmit only in specific time slots so that partial collisions do not occur. Figure 6.28 illustrates the collision problem.

Although there is a saving in transmission time because no call set up is required, the low throughput and poor spectral efficiency has restricted random access use in satellite systems to cases where traffic bursts are short and highly intermittent. In general, it is used on single SCPC-FDMA channels, rather than on whole transponders. The CSC, described in the previous section, is an example of a SCPC-FDMA random access channel within a transponder that can successfully use random access because it is lightly loaded.

6.12 Packet Radio Systems and Protocols

Data transmission between computers or terminals requires agreed methods by which connections are established and data is transferred. When we make a telephone call,

there are conventions and etiquette, which define how a telephone connection is established and when each person speaks. For example, you decide to call your friend John Doe. You lift the telephone handset and hear dial tone. The telephone system is telling you that it is ready for you to dial a number. You dial the telephone number of your friend and wait to hear a ringing tone. The telephone system is telling you that it is trying to attract the attention of your friend. If your friend answers, you expect to hear "Hello, this is John Doe." You might reply "Hello, this is Bill Smith. How are you?" If the person who answers just says "Hello," you cannot be sure that you have reached John Doe, and might have to say "Hello, this is Bill Smith, is that John Doe?" If you dialed a wrong number and have connected to the wrong person, you might say "Sorry, I must have dialed a wrong number." Then you would put the phone down and try again.

If you weren't sure whether the answerer said "John Doe" or "John Roe" because of noise on the line, you would ask for a repeat transmission.

The process of creating a telephone connection to a friend makes use of signals provided by the telephone system and a set of procedures based in human etiquette and intelligence. Humans can readily determine whether they have reached the correct person, and how to proceed if the call does not go through correctly. A data transmission system lacks the intelligence of a human, and cannot readily adapt to changing responses in the way that humans can. Data transmissions make use of packets and protocols to ensure that automatic connection and transfer of data can be achieved reliably without human intervention. One dictionary definition of *protocol* is:

"The code of ceremonial forms and courtesies of precedence, etc., accepted as proper and correct in official dealings, as between heads of state or diplomatic officials" (Webster 1980). It is this sense of the word protocol that is used to describe the rules by which two data terminals can automatically connect to each other through a communication system and then transfer data.

The creation of protocols for data transmission is a very large subject, with many books and papers devoted to the design and performance of different methods. In this text, only the briefest summary of the subject is included. The widespread ownership of personal computers and growth of the internet helped spur development of efficient and powerful protocols like transmission control protocol/internet protocol (TCP/IP). The International Standards Organization (ISO) has created a seven layer model for machine to machine communication known as the open systems interconnect (OSI), which separates the functions of different parts of the system. The ISO-OSI model is discussed in Chapter 8 and illustrated in Figure 8.9. Although the model is widely quoted as describing the structure of data communication systems, it rarely seems possible to identify seven separate layers within any given system. The lowest layer, the physical layer, is the one with which this text is concerned – the transport of bits from one point to another. Regardless of the method of transportation, the ISO model assumes that bits are carried across the physical layer, in both directions, possibly arriving with some errors. The remaining six layers of the model are embedded in the hardware and software of the terminals at each end of the link. The second layer of the model provides error detection and correction, either in hardware or software, and the remaining layers are responsible for organizing the data transfer, from making connections to billing for the service.

Terrestrial data communication has evolved through a series of protocols, beginning with a standard by IBM, known as high level data link (HDL), through X.25 to ATM (asynchronous transfer mode) and TCP-IP. ATM uses a 53 byte packet and was designed to be transmitted over fiber optic line networks using the DS-3 44.736 Mbps transmission rate according to an IEE standard (P802.6). Digital cellular radio systems,

although using three different standards, are compatible, and any cellular telephone designed for TDMA can work with any provider's TDMA system. With the exception of DBS-TV, satellite communication systems have not evolved a set of standards, and many systems use proprietary protocols. For example, the Iridium and Globalstar LEO satellite communication systems use digital transmission with completely different protocols. Handsets designed for use with the Iridium system cannot communicate with Globalstar satellites. Typically, the long delay time inherent in transmission via a GEO satellite creates problems when interfacing to a terrestrial protocol such as TCP/IP designed for much shorter delays. Special interfaces are needed at the earth stations that allow the protocol to be adapted for satellite use.

One protocol and packet design that has been widely accepted for use in amateur satellite systems is AX.25. The AX.25 protocol is based on X.25, a protocol developed for terrestrial data communications, and is used by amateur radio operators in terrestrial data communication networks. The protocol was adapted for use in amateur radio LEO satellites with vhf and uhf transponders operating in a store and forward mode. Several of these satellites were built and orbited, providing a method for amateur radio operators to send messages by satellite (Davidoff 1998).

Figure 6.29 shows the structure of the AX.25 packet. All packets begin and end with a unique word, called a *flag*, 01111110, which is not allowed to appear in any other part of the packet. The flag marks the end of one packet and the start of the next packet, so that the receiving data terminal can extract the packet contents correctly. The general format of the packet contents is a header, followed by message bits, followed by a CRC. The header contains addresses, in the form of amateur radio call signs, for the sender and intended recipient, and control information that helps the receiving station identify the contents of the packets. The control bits, for example, tell the receiving terminal how long the packet is, and define whether this is a broadcast packet, intended to be viewed by all receiving stations, or a packet for a specific recipient. Control bits also specify the type of packet – some packets contain no message bits and are sent to convey system information. The CRC allows the receiver to check whether the packet was received correctly, and to call for a retransmission if an error is detected. The interested reader

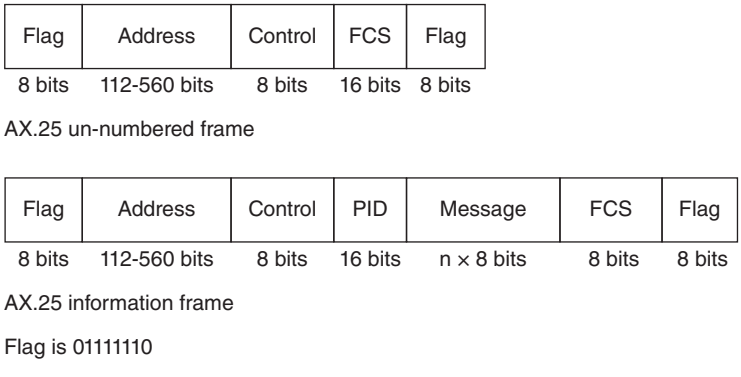


Figure 6.29 Packet structure used in the AX.25 amateur protocol. The unnumbered frame is used only for control purposes and contains no message data. Flag is located at the start of each frame and also forms the end of frame marker when the next frame is sent. Flag contains a sequence of six ones, which is not allowed in the message field. The value of *n* can be anywhere between 1 and 256. PID is packet ID number. FCS is frame check sequence.

should refer to the *Amateur Satellite Handbook* for further details of the amateur radio satellite system (Davidoff 1998).

All data transmission system must have some form of protocol, and data is almost always sent in packet form. Thus whenever multiple access techniques are discussed and digital data are transmitted, it can be assumed that some form of packet transmission is used.

6.13 Code Division Multiple Access (CDMA)

CDMA is a system in which a number of users can occupy all of the transponder bandwidth all of the time. CDMA signals are encoded such that information from an individual transmitter can be recovered by a receiving station that knows the code being used, in the presence of all the other CDMA signals in the same bandwidth. This provides a decentralized satellite network, as only the pairs of earth stations that are communicating need to coordinate their transmissions. Subject to transponder power limitations and the practical constraints of the codes in use, stations with traffic can access a transponder on demand without coordinating their frequency (as in FDMA) or their time of transmission (as in TDMA) with any central authority. Each transmitting station is allocated a CDMA code; any receiving station that wants to receive data from that earth station must use the correct code. CDMA codes are typically 16 bits to many thousands of bits in length, and the bits of a CDMA code are called *chips* to distinguish them from the message bits of a data transmission. The data bits of the original message modulate the CDMA chip sequence, and the chip rate is always much greater than the data rate. This greatly increases the speed of the digital transmission, widening its spectrum in proportion to the length of the chip sequence. As a result, CDMA is also known as *spread spectrum*. *Direct sequence spread spectrum* (DSSS) is the only type currently used in civilian satellite communication; *frequency hopping* spread spectrum (FH-SS) is used in the Bluetooth system for multiple access in short range local area wireless networks.

CDMA was originally developed for military communication systems, where its purpose was to spread the energy of a data transmission across a wide bandwidth to make detection of the signal more difficult (called *low probability of intercept)*. Spreading the energy in a signal across a wide bandwidth can make the noise power spectral density (NPSD) in the receiver larger than the PSD of the received signal. The signal is then said to be *buried in the noise*, a common feature of DSSS signals, and the signal is much harder to detect than a signal with a PSD greater than the receiver's NPSD. The correlation process that recovers the original data bits from a DSSS spread spectrum signal is also resistant to *jamming*, the deliberate transmission of a radio signal at the same frequency to blot out someone else's transmission. Both of these attributes are valuable in tactical military communication systems.

CDMA has become popular in cellular telephone systems where it is used to enhance cell capacity. However, it has not been widely adopted by satellite communication systems because it usually proves to be less efficient, in terms of capacity, than FDMA and TDMA. The Globalstar LEO satellite system was designed to use CDMA for multiple access by satellite telephones; one advantage of CDMA in this application is *soft handoff* in which the same signal is received from two satellites during the period that one satellite is about to disappear below the horizon and another satellite has just appeared

above the horizon. This technique increases the CNR in the receiver when the satellites are at their maximum range and the signals are weakest. More details of the Globalstar system can be found at (Globalstar 2017).

The GPS navigation system uses DSSS CDMA for the transmission of signals that permit precise location of a receiver in three dimensions. Up to 14 GPS satellites may be visible to a receiver close to the earth's surface at any one time. CDMA is used to share a single RF channel in the receiver between all of the GPS satellite transmissions. Chapter 12 gives details of the GPS signal structure and describes the process of data recovery from the DSSS satellite signals.

6.13.1 Spread Spectrum Transmission and Reception

This discussion of CDMA for satellite communications will be restricted to direct sequence systems, since that is the only form of spread spectrum that has been used by commercial satellite systems to date. The *spreading codes* used in DSSS CDMA systems are designed to have good autocorrelation properties and low cross correlation. Various codes have been developed specifically for this purpose, such as Gold and Kasami codes (Pseudorandom noise 2018; Pseudo-random noise codes 2013). The following discussion is based on the C/A spread spectrum codes used in civil receivers of GPS position location signals, which are all Gold codes. More details of the CDMA system of GPS satellites and receivers can be found in Chapter 12.

GPS satellites transmit *pseudo-random sequence* (PRN) codes, also known as *pseudo noise codes*. All GPS satellites transmit a C/A (course acquisition) code at the same carrier frequency, 1575.42 MHz, called L1, using BPSK modulation. The C/A code has a clock rate of 1.023 MHz and the C/A code sequence has 1023 chips, so the PRN sequence lasts exactly 1.0 ms. A second spread spectrum signal, the P code is also transmitted by GPS satellites. Its use is restricted to authorized users, primarily military. The C/A code is transmitted as BPSK modulation of the L1 frequency RF signal, and is also modulated by a 50 bps navigation signal. The navigation signal contains information essential to the calculation of the location of a GPS receiver.

Figure 6.30 shows the way in which the C/A code is generated on board a GPS satellite. There are two 10 bit shift registers known as G1 and G2 that are clocked at 1.023 MHz. Each shift register generates a 1023 chip PRN code sequence using feedback loops not shown in Figure 6.30. The position of the S1 and S2 output taps of the G2 shift register determine which version of the C/A code is generated. The outputs of the G1 and G2 shift registers are added (modulo 2) to create a 1023 bit long Gold code sequence, which is the C/A code for one satellite. There are a total of 37 C/A codes available to GPS satellites, determined by the S1 and S2 settings. Every GPS receiver contains an identical C/A code generator.

In a GPS satellite, the C/A code is modulated with 50 bps navigation data. The C/A code sequence lasts exactly 1.000 ms, so there are 20 repetitions of the C/A code within each bit of the navigation message. When the navigation message bit changes from a 0 to a 1, or a 1 to a 0, the next 20 C/A sequences are inverted. The C/A code is modulated onto the L1 carrier using BPSK and sent to the satellite L1 transmitter.

The C/A code for a particular satellite is created with an algorithm that includes the identification number of the GPS satellite, creating a unique code with a signal number

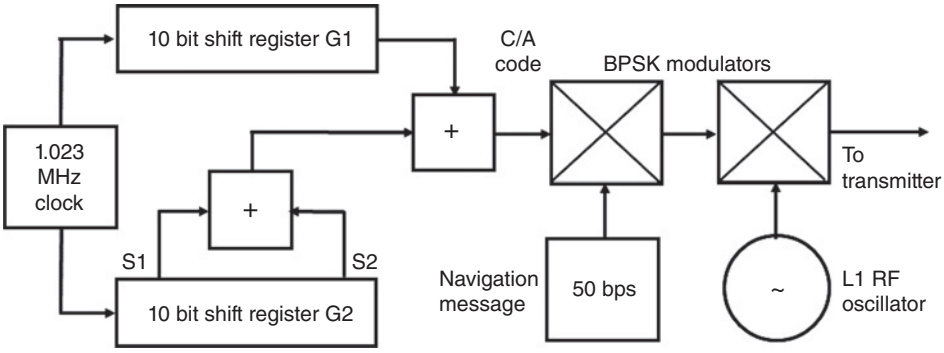


Figure 6.30 Simplified diagram of the C/A code generator used on GPS satellites. The outputs of two 10 bit shift registers G1 and G2 are added to form a Gold code. The settings of the taps S1 and S2 on the G2 shift register determine which of the 37 C/A codes is generated. The C/A code is modulated with a 50 bps navigation message and transmitted as BPSK modulation of the L1 carrier.

that is the same as the GPS satellite number (*space vehicle, SV number*). The algorithm for generating a C/A code for SV number i is

$$C_{\mathrm{i}}(t) = G1(t) \oplus G2\,(t + n_{\mathrm{i}} T_{\mathrm{c}}) \tag{6.31}$$

where n_{i} is a unique value for each C/A code sequence and T_{c} is the C/A code chip period. The $\oplus$ symbol is the exclusive OR function. Refer to Chapter 12 for more details of the C/A codes.

6.13.2 GPS C/A Code Receiver

GPS signals are very weak, as in many spread spectrum systems, with noise power typically exceeding signal power by 19 dB in a receiver with an omnidirectional antenna. That makes it impossible to tune a conventional radio receiver to a GPS satellite. In a conventional GPS C/A code receiver with an omnidirectional antenna the signal can only be observed and utilized after correlation between the received signal and an identical locally generated C/A code has been obtained. The bandwidth of the BPSK C/A code signal is 2.046 MHz between the first nulls of the RF spectrum, and this value is used in GPS system design. The correlation process in the receiver removes the BPSK C/A code modulation (at an IF frequency) and recovers an IF carrier modulated by the 50 bps navigation message. The recovered carrier has a bandwidth of 100 Hz between first nulls of its spectrum, allowing the signal to pass through a 1000 Hz BPF. Narrowing the bandwidth of the signal from 2 MHz to 1000 Hz improves the SNR by a factor of 2000 or 33 dB, so the correlator output has a nominal SNR of 14 dB. This is sufficient to allow demodulation of the 50 bps BPSK navigation message, which can be filtered by a 50 Hz baseband LPF to increase the SNR to nominally 27 dB. If the wrong C/A code is used in the correlator of the receiver, its output is a BPSK modulated signal with a bandwidth of 2.046 MHz, which gives near zero output from the 1000 Hz filter. Given no information about which satellites are visible (a cold start) a search must be conducted to find the correct C/A code and its starting position. Direct conversion to baseband and a baseband correlator is also used in some GPS receivers.

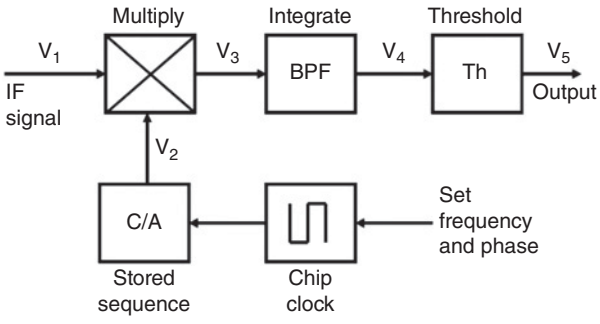


Figure 6.31 Single channel correlator for C/A code acquisition. The BPF has a bandwidth of 1 kHz during the acquisition process and is narrowed to 50 Hz once the signal is acquired. The C/A code sequence generator is stepped through all the GPS C/A codes and code timing until lock is established.

6.13.3 C/A Code Correlator

Figure 6.31 is a simplified block diagram of a single channel correlator. The correlator has a C/A code generator identical to the C/A code generator carried by every GPS satellite. The IF signal $V1$ is BPSK modulated by a 1.023 Mcps C/A chip sequence and the signal is buried well below receiver noise. The signal $V1$ is multiplied by the output $V2$ of a C/A code generator that produces the selected code sequence. When the locally generated C/A code matches the received C/A code and is correctly synchronized, the output of the multiplier, $V3$, is a sine wave at the IF frequency with the C/A code removed. The signal still has BPSK modulation by the navigation message at 50 Hz. The threshold detector can be as simple as an envelope detector that produces a constant DC output $V5$ once the correlator is locked to the received signal. To acquire a C/A code signal from a GPS satellite, a particular C/A code is selected and the locally generated C/A code is stepped in one chip increments through the entire 1023 chip sequence until the voltage $V5$ indicates that the C/A code has been detected, at which point the stepping of the code is stopped and the channel changes to a locked state. If the wrong C/A code was selected, the receiver must try different codes until lock is established.

The correlation process is illustrated in Figures 6.32 and 6.33 with a nominal IF frequency of 2.0 MHz. In Figure 6.32, the locally generated C/A code $V2$ exactly matches the BPSK modulated C/A code present in the IF signal $V1$. If the signal $V1$ were viewed on an oscilloscope it would appear to be white noise because the signal voltage $V1$ has an rms value one tenth of the rms noise voltage. The output of the multiplier $V3$ is the product of $V1$ and $V2$. Each time the IF BPSK signal $V1$ has a phase reversal, the locally generated C/A code signal $V2$ changes from −V to +V volts, or +V to −V volts, exactly in step with the IF signal modulation. The result is that the multiplier output is a sine wave at the IF frequency without C/A code modulation, for the 1.00 ms duration of the C/A code sequence. The process repeats each millisecond for 20 ms, or multiples of 20 ms, until the navigation message changes the polarity of the sine wave. The spectra of the $V1$ IF signal and the $V3$ multiplier output are illustrated in Figure 6.32 for the case when the correct C/A code has been selected and the start of the locally generated C/A chip sequence is correctly aligned with the start of the receiver GPS C/A code signal. When an output from the correlator is detected, the receiver locks to the specific C/A code, tracks the received signal in frequency and time, and demodulates the 50 Hz BPSK signal to recover the navigation message.

Figure 6.33 illustrates the situation when the locally generated C/A code $V2$ is not synchronized to the received signal $V1$. The signal $V2$ may be the correct C/A code and only one chip away from the correct timing, or it could be the wrong C/A code, but in

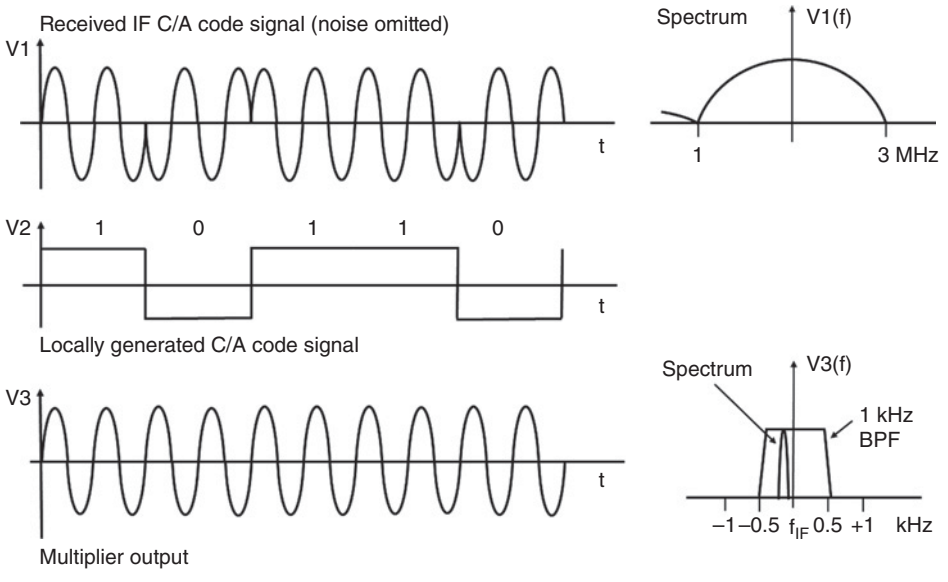


Figure 6.32 Operation of the correlator in a GPS C/A code receiver. The IF signal *V*1 has BPSK modulation by a C/A code at 1.023 Mcps. The IF signal is multiplied by a locally generated C/A code *V*2. When the two codes are the same and exactly aligned in time, the chip modulation is stripped from the IF signal and only the 50 bps navigation message modulation remains. The bandwidth of IF signal *V*3 at the correlator output is reduced to 100 Hz and the signal passes through the 1 KHz BPF. The BPF bandwidth of 1 kHz is chosen because GPS satellites are in a MEO orbit and the L1 signal has up to 4 kHz Doppler shift. The receiver must search in 1 kHz steps for the correct Doppler frequency offset before the satellite signal can be acquired.

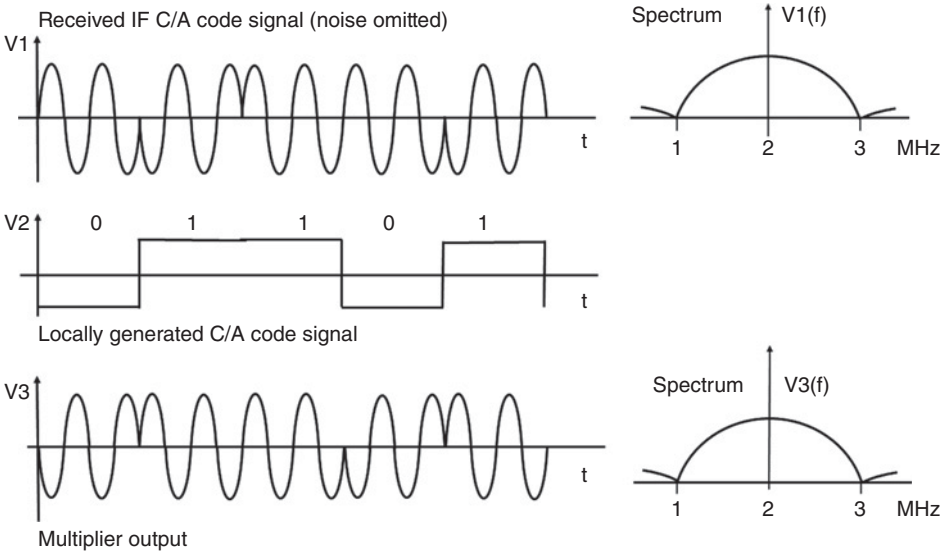


Figure 6.33 The locally generated C/A code does not match the received GPS C/A code signal, or is not correctly aligned in time. The 1.023 MHz chip modulation of the signal is not removed, so the correlator output V3 has a bandwidth of 2.046 MHz and very little energy passes through the 1 KHz filter at the correlator output. The receiver will not lock to the GPS signal.

either case the output of the multiplier has BPSK modulation present at the 1.023 Mcps chip rate. The spectrum of the *V*3 signal is the same as the *V*1 signal, 2.046 MHz between nulls, and very little energy passes through the 1 kHz BPF. If the locally generated C/A code does not match any of the codes present in the received signal, the result will be the same – no output from the BPF and no GPS satellite detected.

As noted earlier, it is very difficult to make a BPF with a bandwidth of 1 kHz at a center frequency of 2 MHz, as this requires a Q factor of 2000, and the filter needs to be implemented digitally. Instead, a phase locked loop is used, which has the characteristic of a narrow BPF centered at a specific frequency.

6.13.4 Acquiring a Spread Spectrum Signal

Spread spectrum signals, like the GPS C/A code signal at frequency L1 are typically buried in noise or from overlay of other CDMA signals. With a negative decibel CNR at the input to a BPSK demodulator there will be no signal output, just noise. The receiver must first use correlation as described in the previous section to lock to the signal by stripping off the chip sequence. A GPS receiver may be receiving signals from as many as 14 satellites at the same time, or as few as 4, the minimum number required to solve for the receiver's location.

If the GPS receiver does not know which satellites are visible, it must search all the possible 37 C/A codes until lock is achieved. Only 24 out of the 37 C/A codes are normally in use because the GPS constellation has 24 satellites. This process is complicated by the Doppler frequency shift caused by the GPS satellite motion in orbit, requiring eight different frequency settings of the receiver. The timing of the C/A code is not known to the receiver so all 1023 possible start positions of the locally generated C/A code sequence must be tried. When no information is available to the receiver (a cold start) it can take a long time to work through all the possible receiver frequencies, C/A codes, and code start positions. Most GPS receivers can complete the process in 30 seconds. If the receiver has been switched on recently, it can start by assuming the same satellites are visible and speed up the search process (a warm start).

From a cold start, the GPS receiver must typically work through the following steps to acquire its first satellite.

1. Select a receiver frequency.
2. Select a satellite (by SV number). Generate that satellite's C/A code.
3. Try all 1023 start positions of the code to attempt to correlate with the received signal.
4. Repeat step 3 several times. If correlation is obtained, repeat another three time to confirm.
5. If no correlation is obtained, try a different SV code. Repeat until all 24 operational SV codes have been tried.
6. If no correlation is obtained, change to a different receiver frequency.
7. Repeat steps 2 through 6 until a satellite is acquired.
8. Once the first satellite is acquired, decode the navigation message to find other visible satellites and their Doppler shifts, then repeat the acquisition process, which will be much quicker given the additional information available.
9. Change to tracking mode in which Doppler shift and code rate are tracked through frequency and code lock loops.

For further details of the GPS CDMA multiple access spread spectrum system, read Chapter 12, on GPS and GNSS.

6.13.5 Processing Gain and CDMA System Capacity

Processing gain is the ratio of chip rate at the input of the correlator to bit rate at the correlator output, usually quoted in decibels. The SNR at the correlator output is equal to the CNR at its input with the processing gain added. The despreading process using a correlator to recover the original signal adds a *processing gain* equal to the $(\text{CNR})_{\text{SS}}$ ratio of the received spread spectrum signal. Hence the $(\text{SNR})_{\text{out}}$ in the spread spectrum receiver after the correlator is given by

$$(\text{SNR})_{\text{out}} = (\text{CNR})_{\text{SS}} + 10\log_{10} M \tag{6.32}$$

where M is the ratio of chip rate to bit rate. The SNR must be sufficiently high for the receiver to recover the bits of the transmitted signal with a reasonable BER. For example, if a BER no larger than 10^{-6} is required, $(\text{SNR})_{\text{out}}$ must be greater than 11.0 dB, allowing a 0.4 dB implementation margin with no forward error correction.

In a DSSS CDMA system where there are a number of CDMA signals present at the input to each receiver, it is usual to treat the unwanted (interfering) CDMA signals as noise. If a receiver has an input containing S signals, each at a power level C watts, and the receiver thermal noise power is N_{t} watts, the $(\text{CNR})_{\text{in}}$ ratio for the wanted signal is approximately

$$(\text{CNR})_{in} = 10\log_{10}\left[\frac{C}{N_{\text{t}} + (S-1)\times C}\right] \text{dB} \tag{6.33}$$

where $(N_{\text{t}} + [S-1]\times C)$ watts is the total noise at the receiver input. The term $(S-1)\times C$ is the power N_{i} watts of the $(S-1)$ interfering CDMA signals. (Note that N_{t} and C must be added in watts, not decibel units.) The correlator in the receiver adds a processing gain of $10\log_{10} M$ dB to the input CNR, as seen in Eq. (6.32), and outputs a correlated signal with a $(\text{SNR})_{\text{out}}$. Hence the output SNR for the bit stream in the receiver is given by

$$(\text{SNR})_{\text{out}} = 10\log_{10}\left[\frac{C}{N_{\text{t}} + (S-1)\times C}\right] + 10\log_{10} M\, dB \tag{6.34}$$

If S is a large number, it is probable that

$$[N_{\text{t}} + (S-1)\times C] \approx (S-1)\times C \text{ W} \tag{6.35}$$

and then Eq. (6.35) reduces to

$$(\text{SNR})_{\text{out}} = 10\log_{10}\left[\frac{1}{S-1}\right] + 10\log_{10} M = 10\log_{10}\left(\frac{M}{S-1}\right) \text{dB} \tag{6.36}$$

If S is also large such that $S \gg 1$ then

$$(\text{SNR})_{\text{out}} \approx 10\log_{10}\left(\frac{M}{S}\right) \text{dB} \tag{6.37}$$

Examination of Eq. (6.37) shows that M, the number of chips in the spreading code must be 10 times larger than S if the output SNR is to be greater than 10 dB, and that the system capacity is independent of the thermal noise power in the receiver.

If M must be 10 times larger than S to allow demodulation of the spread signal without many bit errors, the total bit rate through the transponder in bits per hertz using CDMA will be numerically less than one tenth of the bandwidth in hertz. This results in poor utilization of the RF bandwidth when CDMA is used, compared to FDMA or TDMA, as the following example demonstrates.

Example 6.8 CDMA in a Fixed Earth Station Network

A DSSS CDMA system has a number of earth stations sharing a single 54 MHz bandwidth Ka-band transponder. Each station has a different 1023 bit PRN sequence, which is used to spread the traffic bits into a bandwidth of 45 MHz. The transmitters and receivers use SRRC filters with $\alpha = 0.5$ and the chip rate is 30 Mcps. Determine the number of earth stations that can be supported by the CDMA system if the correlated output SNR = 12 dB.

Equation (6.37) gives

$$(\text{SNR}) = 12\ \text{dB} = 10\log_{10}\left(\frac{M}{S-1}\right) = 30.1 - 10\log_{10}(S-1)$$

Hence

$$10\log_{10}(S-1) = 18.1\ \text{dB}$$

$$S = 64 + 1 = 65$$

Each of the carriers has a data bit rate of 30 Mbps/1023 ≈ 29.325 kHz, so the transponder carries a total bit rate of 65 × 29.325 kbps = 1.906 Mbps. A 54 MHz bandwidth transponder operated in FDMA or TDMA would have a much higher capacity, typically 90 Mbps with QPSK modulation and no FEC.

The capacity of the system can be improved by adding half rate forward error correction (FEC) coding to the baseband signal to reduce the SNR required for detection of the bits in the receiver. If the FEC system has a coding gain of 6 dB, we can use SNR = 12–6 = 6 dB. Using Eq. (6.37), because we now know M » S.

$$6\ \text{dB} = 10\log_{10}\left(\frac{M}{S}\right)$$

This gives $S \approx M/4 = 255$ channels. The data bit rate of each channel (before application of half rate FEC) is now 14.66 kbps and the total throughput of the transponder is 255 × 14.66 kbps = 3.74 Mbps. This is still well below the capacity of a FDMA or TDMA system.

We can conclude that CDMA is useful in commercial systems only where efficient use of satellite capacity is not important, or where the ease with which stations can leave and join the network outweighs the loss of efficiency, or where power limitations in the transponder ensure that it cannot be heavily loaded. One example of such use is the tracking of wild animals and birds fitted with collars that have a GPS receiver to determine the location of the animal or bird and a Globalstar transmitter to report data back to a gateway station. The collar collects data in a memory over an extended period, 12 or 24 hours for example, and then transmits a CDMA data packet to a Globalstar satellite at a random time. Data is forwarded daily from the gateway station to researchers who can then plot the track of the subject animal or bird. One such system has been proposed to track the location of reindeer in the north of Norway that are at risk of wandering onto railroad tracks (Liveview 2017). It is estimated that the reindeer herd in the area

has 600 000 members, so it may be some time before they are all fitted with tracking collars.

Example 6.9 CDMA in a LEO Satellite Network

A LEO satellite communication systems uses direct sequence CDMA as the multiple access method for groups of terminals within each of its multiple antenna beams. The terminals generate and receive compressed digital voice signals with a bit rate of 9.6 kbps. The signals are transmitted and received at a chip rate of 5.0 Mbps as BPSK modulated DSSS-CDMA. In the absence of any other CDMA signals, the input power level at the receiver input is −116.0 dBm (−146.0 dBW) for one CDMA signal, and the noise temperature of the receiving system is 300 K. The satellite transmits 31 simultaneous CDMA signals. Find the SNR for the 9.6 kbps BPSK signal after despreading, and estimate the BER of the data signal, given a system implementation margin of 1 dB. If two of the multiple beams from the satellite overlap, so that a second group of 31 DSSS-CDMA signals is present at the receiver, find the BER of the wanted signal.

Answer

The thermal noise power in the receiver is $N_t = k\,T_s\,B_n$. For the chip rate of 5.0 Mbps with BPSK and ideal SRRC filters, $B_n = 5.0$ MHz. Hence

$$N_t = -228.6 + 24.8 + 67.0 = -136.8 \text{ dBW} = 2.09 \times 10^{-14} \text{ W}$$

There are 30 interfering CDMA signals overlaid in the 5 MHz bandwidth of the receiver filter. The total interfering power is

$$I = 30 \times P_r = -146.0 + 14.8 \text{ dB} = -131.2 \text{ dBW} = 7.59 \times 10^{-14} \text{ W}$$

The carrier to noise plus interference ratio must be calculated in watts, not dBW, because we cannot add noise and interference in decibel units, only in watts. The CNR in the receiver for the wanted CDMA signal is

$$\frac{C}{N_t + I} = \frac{2.51 \times 10^{-15}}{2.09 \times 10^{-14} + 7.59 \times 10^{-14}} = \frac{2.51}{96.8} = 0.0259 \text{ or } -15.9 \text{ dB}$$

The carrier power is well below the noise plus interference power, so the wanted carrier is hidden below the noise and interference. This is called a *low probability of intercept* signal.

CDMA was first used by military radio communication systems to reduce the likelihood of a tactical radio transmitter being detected.

The coding gain, G_c, for the CDMA receiver is given by the chip rate divided by the bit rate

$$G_c = R_c/R_b = 5.0 \text{ Mcps}/9.6 \text{ kbps} = 520.8 \text{ or } 27.2 \text{ dB}$$

Hence, after correlation of the wanted code (despreading), the SNR ratio of the 9600 bps BPSK signal is

$$\text{SNR} = -15.9 + 27.2 = 11.3 \text{ dB}$$

With an implementation margin of 1 dB, the effective SNR is 10.3 dB = 10.7 as a power ratio. For BPSK, the BER is

$$P_e = Q\left(\sqrt{2(\text{CNR})_{\text{eff ratio}}}\right) = Q\,(4.63) = 1.6 \times 10^{-6}$$

If a second group of 31 signals is present at the receiver from an overlapping satellite beam, there will be additional interference, which lowers the C/(N + I) ratio. The interfering power from 31 signals is

$$I = 31 \times P_r = -116.0 + 14.9\ \text{dB} = -101.1\ \text{dBm} = 7.76 \times 10^{-14}\ \text{W}$$

Hence the new $C/(N_t + I)$ ratio is

$$\frac{C}{N_t + I} = \frac{2.51 \times 10^{-15}}{\left(2.09 \times 10^{-14} + 7.59 \times 10^{-14} + 7.76 \times 10^{-14}\right)}$$
$$= 2.51/174.4 = 0.0144\ or\ -18.4\ \text{dB}$$

Note that the interfering CDMA signals dominate over thermal noise in this system. After correlation of the wanted code the SNR ratio of the 9600 BPSK signal is

$$\text{SNR} = -18.4 + 27.2 = 8.8\ \text{dB}$$

With an implementation margin of 1 dB, the effective SNR is 7.8 dB = 6.02 as a power ratio. For BPSK the BER is

$$P_e = Q\left[\sqrt{2(\text{CNR})_{\text{eff ratio}}}\right] = Q\,(3.47) = 3 \times 10^{-4}$$

We would need to add FEC to the baseband signal to improve the BER. To achieve a BER of 10^{-6} in this case, a coding gain of about 3 dB would be adequate. With half rate convolutional coding, a coding gain of 5.5–6 dB is typical, which would provide a margin of 3 dB over a BER of 10^{-6} and a baseband data rate of 4.8 kbps. This bit rate can support a single digital speech channel with linear predictive encoding (LPC) compression.

The advantage of overlapping beams in a mobile satellite system is that the wanted signal can be transmitted by both satellites (using different CDMA codes) and blockage of one beam by an obstruction on the ground does not cause loss of the signal if the other beam can still be received. The wanted signal from both satellites can be combined at IF using a *rake receiver*, which improves the BER (Rake receiver 2017). With optimum combining of the same baseband signal, the BER will be the same as for a single beam with 31 users.

Example 6.10 GPS

The global positioning system (GPS) uses direct sequence CDMA for both the C/A and P code transmissions. The design and operation of GPS is discussed in detail in Chapter 12, from which this example of a direct sequence CDMA system is drawn.

The C/A code transmissions from GPS satellites are 1023 chip PRN sequences, formed into 37 *Gold codes* where the bits are referred to as *chips* to distinguish them from data bits. At any given time, up to 14 GPS satellites may be visible, but we will consider the case where 10 GPS satellites are visible and interference with the wanted signal is limited to nine overlaid CDMA signals. In this example we will assume for simplicity that the signals are all received with equal power. There are variations in the transmitted power

Table 6.7 Downlink budget for GPS C/A code signal

Satellite EIRP	26.8 dBW
Path loss	186.8 dB
Receive antenna gain	0 dB
Received power P_r	−160.0 dBW

between satellites, and a satellite close to the horizon has a longer path length, so there will be variations in the received power level of individual satellite signals in practice.

The received power level for a typical C/A signal from an early GPS satellite is given by the downlink budget in Table 6.7, assuming 0 dB receiving antenna gain. Later GPS satellites achieved better performance – see Chapter 12 for details. The C/A code is transmitted at a bit rate of 1.023 Mcps using BPSK modulation. The receiver IF noise bandwidth is assumed to be 2 MHz.

The interference from nine satellite spread spectrum signals of equal power for the C/A code is given by

$$I = -160.0 + 9.5 = -150.5 \text{ dBW} = 8.91 \times 10^{-16} \text{ W}$$

The thermal noise power in a noise bandwidth of 2 MHz for a system noise temperature of 273 K is $k\, T_s\, B_n$ where

$$N_t = -141.2 \text{ dBW} = 7.59 \times 10^{-15} \text{ W}$$

The noise and interference powers must be added in watts, not in decibels:

$$N_t + I = 8.48 \times 10^{-15} \text{ W} = -140.7 \text{ dBW}$$

The nine interfering satellites have lowered the CNR ratio for the wanted signal by 0.7 dB, a relatively small decrease. This is different from the previous example, 6.9, in which interference power from other CDMA signals exceeded the thermal noise power in the receiver. In a GPS C/A code receiver, thermal noise is the dominant factor, not interference from other satellites.

The CNR for one GPS C/A code signal with nine interfering signals is

$$C/(N_t + I) = -160.0 - (-140.7) = -19.3 \text{ dB}$$

The theoretical coding gain for a 1023 chip code sequence is $10 \log_{10} 1023 = 30.1$ dB as discussed above. Hence the SNR ratio of the correlated C/A signal in a bandwidth of 1 kHz, assuming ideal correlation, is

$$\text{SNR} = -19.3 + 30.1 = 10.8 \text{ dB}$$

Navigation data are modulated onto the C/A code signal at 50 Hz, so there are $20+V$ or $20-V$ samples in succession for each 1 or 0 data bit of the navigation message. Typically, once the GPS satellite signal has been acquired, the BPSK filtering bandwidth can be reduced toward 50 Hz, which adds 13 dB to the effective SNR to give SNR = 23.8 dB. At a 50 bps data rate, errors in the navigation message will rarely occur.

In this example we see that GPS satellites make excellent use of CDMA as a way to obtain a high speed bit stream from which timing information can be obtained, an essential ingredient for any time of arrival position location system, and also a low speed data stream that provides the navigation information for the solution of the location problem.

The example here uses nine interfering satellites with the same power level. Most GPS receivers select the four strongest satellite signals to use in the position location solution. A more realistic scenario would have four satellites at the maximum receive power level and the remainder at a lower level, since GPS satellites orbit in constellations of four, with one constellation always visible, to improve the accuracy of position location measurements. Thus we should expect less than 0.7 dB degradation in CNR due to interference by other satellites' CDMA signals, and the probability of a bit error in the navigation message then becomes very small.

Interestingly, the eastern European satellite navigation system known to the western world as GLONASS uses FDMA for multiple access, with 15 channels spaced by 0.5625 MHz centered at 1602 MHz. Satellites on opposite sides of the earth use the same 15 frequencies since they cannot interfere. Additional channels using CDMA have been added to later satellites in the GLONASS constellation (GLONASS 2017).

6.14 Summary

Analog modulation, primarily FM, has been replaced by PSK as the digital modulation in satellite links. QPSK and 8-PSK are the most widely used forms where CNR values at the receiving earth station are low. With higher CNRs, 8-APSK, and 16-APSK can be employed. QPSK and 8-PSK have constellations with circular shape because all states have the same magnitude. This is useful when satellite transponders are driven close to saturation. APSK modulations employ two or more amplitude levels, which creates problems with non-linear transponders and restricts their use to satellite links that can operate transponders in a quasi-linear condition. MSK, which has characteristics of both FSK and PSK is used in some SCPC applications in the form of Gaussian MSK, where it has some implementation advantages to QPSK, but with a small CNR penalty.

Multiple access is the process by which a number of earth stations interconnect their links through a satellite. In FDMA stations are separated by frequency, while in TDMA they are separated in time. In CDMA, stations use spread-spectrum transmissions with orthogonal codes to share a transponder without interference. Multiple access may be preassigned or demand assigned (DAMA), depending on whether or not it responds to changing traffic loads. In each case, the signals transmitted from a satellite are labeled in a way that allows an earth station receiver to separate out the content.

FDMA is a widely used multiple access scheme. Each transmitting earth station is assigned a frequency band for its uplink transmissions. Because of the back off required to reduce intermodulation distortion with bent pipe transponders, the spectral efficiency (i.e., the number of channels that can be carried per megahertz of bandwidth) degrades as the number of stations that access a transponder increases. FDMA is widely used with VSAT earth stations and SCPC systems where the uplink from the earth station is at a low power level.

In TDMA, earth stations transmit in turn. Since only one carrier is present at a time, only a small TWTA backoff is required and thus almost full transponder EIRP is available. TDMA performance does not degrade with the number of accesses. TDMA transmissions are organized into frames; a frame may contain one or two reference bursts that synchronize the network and identify the frame, and a series of traffic bursts. Each participating station transmits one traffic burst per frame. Frames and individual

traffic bursts are identified by standardized bit sequences called unique words. One of the major technical problems in implementing TDMA is synchronization. Once synchronization is acquired, it must be maintained dynamically to compensate for orbital motion of the spacecraft. TDMA is often combined with FDMA, so that a small number of earth stations share a TDMA frame forming one FDMA access to a transponder. This is called MF-TDMA.

In CDMA stations transmit at the same time and in the same frequency bands using spread-spectrum (SS) techniques. CDMA avoids the centralized network control required for synchronization in TDMA, but tends to achieve poor spectral efficiency. The Globalstar LEO satellite system uses CDMA, with the advantage that an earth station can receive the same signal from more than one satellite at the same time, allowing soft handoff between satellites. Globalstar handsets can transmit at any time and all share the same carrier frequency, a simpler system than either SCPC-TDMA or MF-TDMA.

Random access is used in systems, which have low traffic requirements and can tolerate less than 18% utilization of the RF channels. The advantage of random access is that no central network control is needed.

Digital links between computers require protocols to ensure efficient transfer of data, and invariably use some form of packet communication. Satellite systems have tended to use proprietary protocols, with the result that different satellite systems are not compatible.

Exercises

6.1 The power in RF waveforms can be compared by assuming that the waveform is supplied to a 1 Ω load. For a sine wave $V\cos(\omega t+\varphi)$, the power in the 1 Ω load is $\frac{1}{2}V^2$ watts. QPSK and 8-PSK signals have a constant magnitude V for all symbols, so relative power per symbol is $\frac{1}{2}V^2$ watts. The 16-QAM and 16-APSK constellations shown in Figure 6.3 have I and Q voltage levels of $\pm V$ and $\pm 3V$. Calculate the relative power for a 16-APSK signal compared to an 8-PSK signal, and for 16-QAM compared to QPSK, in dB. How do these results relate to the CNR values required for BER of 10^{-6} for these signals in Table 6.1?

6.2 Generate a mapping table for the QAM constellation shown in Figure 6.3 using I and Q channel voltages of $\pm V$ and $\pm 3V$.

6.3 Generate a mapping table for the 16-APSK constellation shown in Figure 6.3 that creates APSK signals with magnitudes ±V and ±3V volts.

6.4 Generate a mapping table for the demodulator logic for the 16-QAM constellation shown in Figure 6.3 with I and Q signal magnitudes ±V and ±3V volts. The logic circuit uses two-bit ADCs with the following outputs: +3V → 01, +1V → 00, −1V → 10, −3V → 11.

6.5 Five earth stations share one transponder of a 6/4 GHz satellite. The stations all operate in a TDMA mode. Speech signals are sampled at 8 kHz, using

8 bits/sample. The sampled signals (PCM) are then multiplexed into 2.0 Mbps streams at each station, using QPSK without FEC.

a. Find the bit rate for each PCM signal.

b. The number of speech signals (as PCM) that can be sent by each earth station, as a single access, with no overhead (i.e., no header or CRC) or guard times. This is TDM data stream.

c. The shortest frame time for any TDMA scheme.

d. The number of speech signals (as PCM) that can be sent by each earth station, as a single access, with a 1 ms frame, 10 μs guard time and a 10 μs header per frame.

6.6 Five earth stations share one transponder with a bandwidth of 36 MHz on a 6/4 GHz satellite using TDMA. The earth stations each transmit bursts at a symbol rate of 7.5 Msps. A TDMA system is established using a 125 μs frame time. Find the symbol rate at an earth station within the TDMA frame when:

a. No time is lost in overheads or preambles.

b. A 5 μs preamble is added to the beginning of each earth station's transmission.

c. A 5 μs preamble is added to each station's transmission and 2 μs guard band is allowed between every transmission.

6.7 Repeat the analysis of Question 6 with a 2 ms frame instead of a 125 μs frame.

6.8 A satellite link is established between eight 6 GHz uplink earth stations and a 4 GHz receiving earth station with the properties detailed in Table Ex 6.1 Each uplink station transmits a 5 Mbps data channel using QPSK modulation with half rate FEC coding.

Calculate the uplink transmitter power at each earth station, uplink CNR, downlink CNR, and overall CNR in the receiver when the link is operated:

i) in FDMA mode with 3 dB transponder output backoff and 6 dB input backoff.

ii) in TDMA mode with 1 dB transponder output backoff transponder and 3 dB input backoff.

Table Ex 6.1 System parameters for Problem 6.8

Transponder BW	54 MHz
Transponder gain	110 dB
Transponder noise temperature	550 K
Transponder saturated output power	50 W
Satellite transmit antenna gain 4 GHz	20.0 dB
Satellite receive antenna gain 6 GHz	22.0 dB
Earth station receive antenna gain 4 GHz	50.0 dB
Earth station transmit antenna gain 6 GHz	53.0 dB
Earth station receive system noise temp.	60 K
Path loss at 4 GHz, L_p	196 dB
Path loss at 6 GHz, L_p	200 dB

Ignore all losses except path loss. Remember to share the transponder output power between the transmitted signals in FDMA mode.
Note: The output power of the transponder is reduced from the saturated power level by the output backoff value. The input power to the transponder is equal to the saturated output power minus the transponder gain minus the input backoff value.

iii) Comment on the difference in power transmitted by each earth station between the FDMA and TDMA multiple access cases. What is the difference in overall CNR at the receiving earth station between the two multiple access methods? Is this difference significant in a 6/4 GHz link?

6.9 A digital communication system uses a satellite transponder with a bandwidth of 54 MHz. Several earth stations share the transponder using QPSK modulation with half rate FEC coding using either FDMA or TDMA. Standard message data rates used in the system are 100 kbps and 2.0 Mbps. The transmitters and receivers in the system all use ideal SRRC filters with $\alpha = 0.25$, and FDMA channels in the satellite are separated by 50 kHz guard bands. When TDMA is used, the TDMA frame is 250 μs in length, and a 5 μs guard time is required between each access. A preamble of 148 bits must be sent by each earth station at the start of each transmitted data burst.

a. What is the symbol rate for the 100 kbps and 2.0 Mbps QPSK signals sent using FDMA?

b. What is the symbol rate of each earth station's transmitted data burst when TDMA is used?

c. Calculate the number of earth stations that can be served by the transponder when 100 kbps channels are sent using (i) FDMA and (ii) TDMA.

d. Calculate the number of earth stations that can be served by the transponder when 2.0 Mbps channels are sent using (i) FDMA and (ii) TDMA.

6.10 A LEO satellite system transmits compressed digital voice signals to handheld terminals (satphones). Data is sent using QPSK modulation without FEC encoding. The satphones work in groups of 10. The inbound bit stream from the satphone to the satellite is at 10 kbps. The outbound bit stream from the satellite is TDM at a bit rate of 100 kbps, and consists of packets addressed to each of the 10 satphones in a group. All 10 satphones receive the 100 kbps TDM bit stream. The system operates in L-band between the satellite and the satphones where rain fading can be ignored, but blockage from buildings and trees is a significant factor. The satellite uses onboard processing (OBP) and multi-beam antennas. The links use SRRC filters with $\alpha = 0.35$. In this question we will be concerned only with the links between the satellite and the satphones. The CNR on the Ku-band links between the satellite and the gateway station are much higher than CNR on the L-band links. An implementation margin of 1.0 dB must be applied to each link.

a. What is the noise bandwidth of the narrowest BPF in (i) the satphone receiver and (ii) the satellite OBP receiver for the inbound link?

b. What is occupied RF bandwidth of the radio signals of (i) the inbound link (phone to satellite) and (ii) the outbound link (satellite to phone)?

c. The inbound uplink has clear air $(CNR)_{up} = 16.0$ dB. What is the clear air BER in the baseband of the satellite's OBP receiver?
d. What is the available fade margin for $(CNR)_{up}$ on the uplink to the satellite if the inbound link operating threshold is set at BER = 10^{-4}?
e. The outbound link has clear air $(CNR)_{dn} = 16.0$ dB. What is the clear air BER?
f. What is the available fade margin for the overall $(CNR)_{dn}$ on the downlink to the satphone if the outbound link operating threshold is set at BER = 10^{-5}?

6.11 Repeat the analysis of parts (c) through (f) of Question #10 when half rate FEC coding is applied to the signals producing an equivalent CNR improvement of 7 dB in the receivers.

6.12 A Ka-band satellite broadcasts digital television signals over the United States. The nominal bit rate of the signal is 27 Mbps. The digital signal can convey up to 20 pre-recorded MPEG-2 video signals. QPSK modulation is used with half rate FEC coding. The error mitigation techniques employed provide an effective coding gain of 8 dB. (Coding gain of 8 dB means that when the $(CNR)_o$ value of the received signal is X dB, the BER corresponds to CNR = (X + 8) dB.) The link has an implementation margin of 1.6 dB. The transmitters and receivers use SRRC filters with $\alpha = 0.25$.
a. What is the occupied bandwidth of the RF TV signal?
b. What is the symbol rate of the transmitted QPSK signal, and the noise bandwidth of the earth terminal receiver?
c. The minimum permitted BER after error mitigation in the receiver is 10^{-6}. What is the minimum permitted $(CNR)_o$ for the digital TV receiver?
d. The Ka-band link suffers rain attenuation that reduces $(CNR)_o$ in the receiver by 7 dB for 0.1% of the year. If the BER is 10^{-6} under the 0.1% year conditions, what is the clear air $(CNR)_o$ value?
e. A new coding algorithm is developed that provides a coding gain of 9 dB with a bit rate that increases to 30 Mbps allowing several TV channels to be broadcast in HD. Assuming that the SRRC filters in the system can be changed to match the new symbol rate, does implementation of the new coding algorithm improve the system performance? If so, what is the new $(CNR)_o$ margin?

References

Allnutt, J.E. (2013). *Satellite-to-Ground Radiowave Propagation*, 2e, Chapter 6. London, UK: IET.

Alohanet (2017). Wikipedia.org/wiki/alohanet (accessed 26 November 2017).

Aviation week (2018). http://aviationweek.com/aircraft-interiors/competition-ramping-airline-satellite-connectivity (accessed 25 March 2018).

Betaharon, K., Kinuhata, P.P., and Nuspl, R.P. (1987). On-board processing for communication satellites: technologies and implementations. *International Journal of Satellite Communications* 5 (2): 139–145.

Concorde (2017). https://en.wikipedia.org/wiki/Concorde (accessed 25 March 2018).

Couch, L.W. (2007). *Digital and Analog Communication Systems*, 7, Upper Saddle River, NJ. Pearson Education Inc.

Davidoff, M.R. (1998). *The Satellite Experimenter's Handbook (ARRL Publication No. 50 of the Radio Amateur's Library)*, 2e. Newington, CT: ARRL.
ETSI (2009). Digital Video Broadcasting DVB-S2. ETSI EN 302 307 V1.2.1 (2009–08).
Everett, J.L. (ed.) (1992). *VSATs*. Hemel Hempstead, UK: Peter Peregrinus, IEE.
Globalstar (2017). http://www.globalstar.com/en/index.php?cid=8600 (accessed 12 February 2018).
Globalstar (2018). en.wikipedia.org/wiki/Globalstar (accessed 16 February 2018).
GLONASS (2017). https://en.wikipedia.org/wiki/GLONASS (accessed 15 May 2018).
Gray F. (1953). Pulse code communication, U.S. Patent 2,632,058 filed November 1947, awarded March 17, 1953.
Ha, T.T. (1990). *Digital Satellite Communications*, 2e. New York, NY: McGraw-Hill.
Haykin, S.S. and Moher, M. (2005). *Modern Wireless Communications*. Upper Saddle River, NJ: Pearson Education Inc.
Haykin, S.S. and Moher, M. (2009). *Communication Systems*, 5e. Hoboken, NJ: Wiley.
Lathi, B.P. and Ding, Z. (2009). *Modern Digital and Analog Communication Systems*, 4e. Oxford, UK: Oxford University Press.
Liveview (2017). http://www.liveviewgps.com/blog/gps-tracking-reindeer-safe (accessed 23 February 2018).
Maral, G. and Bousquet, M. (2002). *Satellite Communication Systems*, 4e. Chichester, UK: Wiley.
O3B (2017). en.wikipedia.org/wiki/O3b_(satellite) (accessed 23 February 2018).
Perrot, B. (2015). Satellite transmission, reception, and Onboard processing, Signaling, and switching. In: *Handbook of Satellite Applications* (eds. J. Pelton, S. Madry and S. Camacho-Lara), 213–247. New York, NY: Springer.
Pratt, T., Bostian, C.W., and Allnutt, J.E. (2003). *Satellite Communications*, 2e. Hoboken, NJ: Wiley.
Pseudorandom noise (2018). https://en.wikipedia.org/wiki/Pseudorandom_noise (accessed 8 June 2018).
Pseudo-random noise codes (2013). http://www.idc-online.com/technical_ences/pdfs/electronic_engineering/Pseudo_Random_Noise_Codes.pdf (accessed 23 February 2018).
Rake receiver (2017). https://en.wikipedia.org/wiki/Rake_receiver (accessed 12 June 2018).
Rappaport, T.S. (2002). *Wireless Communications – Principles and Practice*, 2e. Upper Saddle River, NJ: Prentice Hall Inc.
SES-17 (2017). https://www.ses.com/press-release/ses-and-thales-unveil-next-generation-capabilities-onboard-ses-17 (accessed 25 March 2018).
Spaceway (2017). http://wikipedia.org/wiki/Spaceway (accessed 12 February 2018).
Spaceway 3 (2015). https://www.hughes.com/technologies/hughes-high-throughput-satellite-constellation/spaceway-3 (accessed 8 June 2018).
ViaSat (2017). http://www.ViaSat.com/about (accessed 22 February 2018).
ViaSat 2 (2017). http://www.exede.com/blog/status-update-ViaSat-2-newest-satellite (accessed 22 February 2018).
ViaSat 3 (2018). https://www.ViaSat.com/products/high-capacity-satellites (accessed 10 June 2018).
Webster (1980). *Webster's New World Dictionary*, 2e. Oviedo, FL: World Publishing Company.
Ziemer, R.E. and Tranter, W.H. (2015). *Principles of Communications*, 7e. Hoboken, NJ: Wiley.

7

Propagation Effects and Their Impact on Satellite-Earth Links

Communications system design requires the development of a link budget between the transmitter and the receiver that provides an adequate signal level at the receiver's demodulator input to achieve the required level of performance and availability. Just being able to detect the signal is not sufficient: it is akin to a person who is hard of hearing knowing someone is shouting at them, but not knowing what that person is actually saying. For audio signals, the distance between the speakers will usually determine the quality of the received information. For electronic signals, the distance between the transmitter and receiver is also important, but there may be potential impairments on the path that will reduce the quality of the received signal, causing the demodulated information to range from essentially perfect (high performance) to unavailable (an outage).

7.1 Introduction

The performance of a link is usually defined for time percentages in excess of 99% over periods of at least a month and is, for digital systems, determined by the bit error rate (BER) that provides the minimum level of service. For analog systems, the carrier to noise ratio (CNR) at the demodulator input that provides the minimum signal quality required defines the performance level for that link. The availability of a link is usually defined for low outage time percentages (typically between 0.04% and 0.5% of a year, or between 0.2% and 2.5% of the worst month, for satellite systems) and is, for digital systems, specified by the BER at which an outage is declared for the link. For analog systems, the CNR at the demodulator input at which no usable signal can be demodulated defines the limit of availability. Figure 7.1 illustrates the concept of performance and availability for a digital communications system with BER as the determinant.

The link budget was covered in Chapter 4, as was link *margin*: the difference in power level between clear sky conditions (essentially the performance level) and that which exists at the threshold of the demodulator when the link is under impaired conditions (the *availability* level). Actually, there are two margins to consider in a link budget: (i) the margin between the *clear sky level* (the signal level when there are no impairments on the link) and the *performance threshold* (the minimum signal level at which the required performance is achieved); and (ii) the difference between the performance threshold and the *availability* threshold when the incoming signal cannot be

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

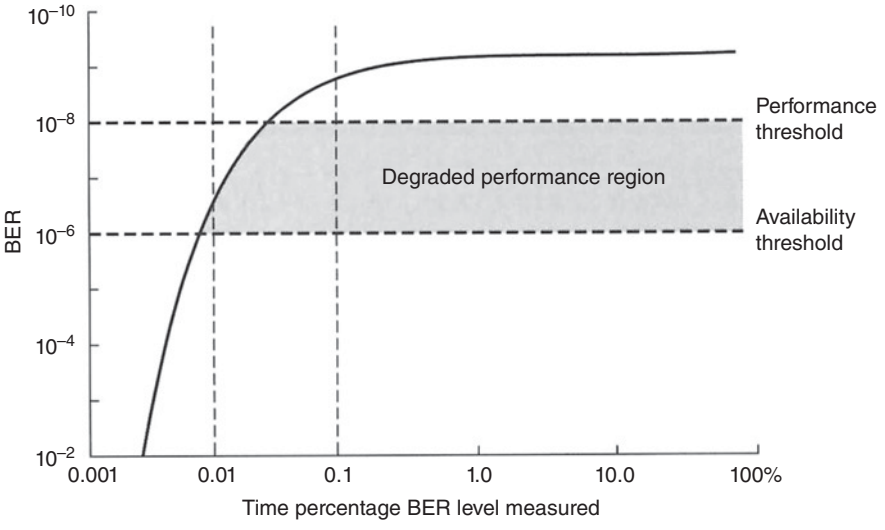


Figure 7.1 Schematic of the bit error rate (BER) statistic for a typical communications link. A link is normally designed to provide a given performance specification for a very high percentage of the time. In this example, a BER of 10^{-8} is the performance required for 99.9% of the time. The time period over which the statistics are taken is usually a year or a month. Atmospheric constituents (gases, clouds, rain, etc.) will cause the BER to degrade. At some point, the BER will reach the level at which an outage is declared. This point defines the availability specification. In this example, a BER of 10^{-6} is the availability threshold and it must be met, in this example, for a minimum of 0.01% of the time.

demodulated successfully. Figure 7.2 illustrates these two concepts of margin for a typical digital Ku-band downlink (11 GHz) located in the Mid-Atlantic region of the United States. As can be seen, the attenuation experienced on the link varies with time percentage, gradually falling through the performance threshold and then the availability threshold. It is the link designer's task to ensure that loss of signal occurs for no longer than the time permitted for that service. The development of an accurate link budget, which includes losses due to the passage of the signal through the atmosphere, is therefore critical.

The key equation in the development of the link power budget in Chapter 4 was Eq. (4.11), repeated here in modified form as Eq. (7.1).

$$P_r = EIRP + G_r - L_p - L_a \text{ dBW} \tag{7.1}$$

This equation indicates how the received power, P_r, in dBW depends on the transmitter effective isotropically radiated power (EIRP) (which is a combination of the output amplifier power, the gain of the transmitting antenna in the wanted direction, and the losses associated with that antenna system), the receiving antenna gain, G_r, (which includes, in this case, all losses associated with the receiving antenna), the path loss, L_p, (given by $20 \log_{10} [4\pi R/\lambda]$, with λ being the wavelength of the signal and R the distance between the transmitting antenna and the receiving antenna) and the attenuation contribution due to the atmosphere, L_a. Of the terms on the right hand side of Eq. (7.1), the only one that is not essentially constant with time for a satellite in geostationary orbit, is the atmospheric loss, L_a. The component L_a, usually referred to as propagation loss,

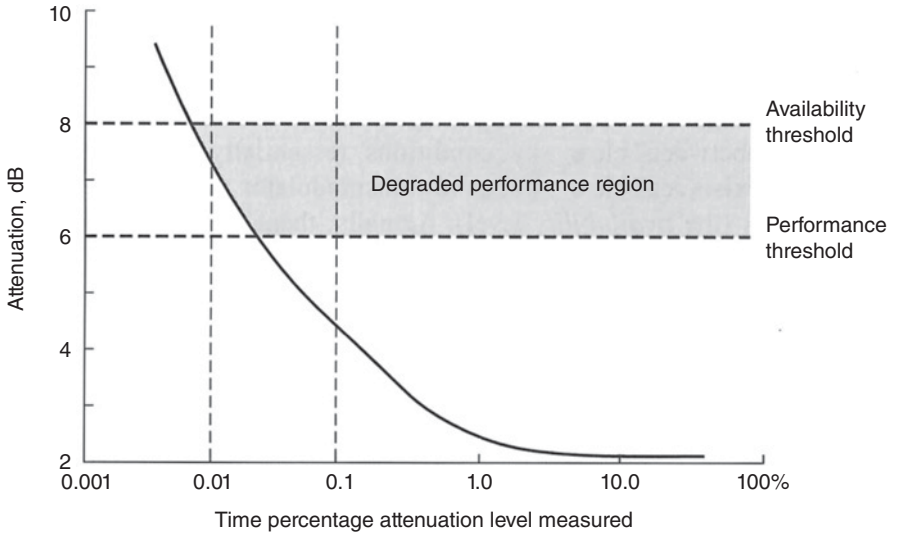


Figure 7.2 Schematic of the loss statistics encountered by a signal on transmission through the atmosphere for a typical Ku-band communications link. In most communications links, an allowance in power margin is built into the link so that the received signal is above the threshold for satisfactory demodulation and decoding. This power margin is commonly referred to as the *fade margin* since the signal, on occasion, appears to fade below the level established in clear sky conditions. In the schematic above, the link experiences an equivalent fade of about 6 dB before it reaches the performance threshold level established for the link (see Figure 7.1). A further fade of 2 dB, making a total reduction in signal level of 8 dB, takes the link below the availability level established for the link (see Figure 7.1). The relationship between power level, fade margin, and BER, will depend on the modulation used. It will also depend on the amount of channel coding used. In the example above, no inner (FEC), or outer (Reed-Solomon, interleaved) coding has been assumed for the link and the modulation is quadrature phase shift keying (QPSK). For most heavily coded links, the difference between good performance and an outage (a change on the order of two to three decades in BER) will occur for a change in signal level of less than 1 dB.

determines the margin required by the communications link to meet both the performance and availability specifications.

A wealth of radiowave propagation data and models can be found in the International Telecommunications Union (ITU-R) texts. Some are held in *reports* (ITU (2018) (a) https://itu.int./pub/R-REP). The ITU-R has an open access policy for retrieving the latest *approved Recommendations* (ITU (2018) (b) https://www.itu.int/pub/R-REC), in addition to providing information on past Reports and deleted Recommendations. The Reports were generally large documents that provided detailed information, including source publications, on a particular area, for example, Report 564, which dealt with Radiowave Propagation. The Reports were generated by *Working Parties* of the ITU-R, which are now referred to as *Study Groups* (ITU (2018) (c) https://www.itu.int/en/ITU-R/study-groups/Pages/default.aspx). Currently (August 2018) there are six Study Groups, and Study Group 3 is responsible for radiowave propagation. In between the regular meetings of Study Group 3, there are meetings of smaller groups, called *Working Parties*, in which detailed aspects of a particular propagation phenomenon are decided on, prior

to submitting the results for approval at a Working Party, and if approved, from thence to a full meeting of the Study Group. Within Study Group 3, there are four Working Parties:

- Working Party 3 J (Propagation Fundamentals)
- Working Party 3K (Point-to-Area Propagation)
- Working Party 3L (Ionospheric Propagation)
- Working Party 3M (Point-to-Point and Earth-Space Propagation)

7.2 Propagation Phenomena

There are many phenomena that lead to signal loss on transmission through the earth's atmosphere. These include: atmospheric absorption (gaseous effects); cloud attenuation (aerosol and ice particle effects); tropospheric scintillation (refractive effects); Faraday rotation (an ionospheric effect); ionospheric scintillation (a second ionospheric effect); rain attenuation; and rain and ice crystal depolarization. Rain attenuation is by far the most important of these losses for frequencies above 10 GHz, because it can cause the largest attenuation and is usually, therefore, the limiting factor in satellite link design in Ku-band and at higher frequencies. Raindrops absorb and scatter electromagnetic waves. In the Ku- and Ka-bands, rain attenuation is almost entirely caused by absorption. At Ka-band, there is a small contribution from scattering by large raindrops. The various propagation loss mechanisms are illustrated in Figure 7.3. We will discuss each of these loss mechanisms briefly; for a detailed treatment the reader should refer to (Allnutt 2011).

Figures 7.1 and 7.2 introduced the concept of a time varying BER (or excess link attenuation). Figure 7.3 indicates where each of the loss mechanisms can be found along the slant path to the satellite. It is also very useful to develop an appreciation for the various time percentages over which each of the propagation loss mechanisms is significant. Figure 7.4 illustrates this schematically, using the same curves from Figure 7.1.

Signal loss – that is, attenuation – affects all radio systems; those that employ orthogonal polarizations to transmit two different channels on a common, or partially overlapping, frequency band may also experience degradations caused by depolarization. This is the conversion of energy from the wanted (i.e., the *co-polarized*) channel into the unwanted (i.e., the *cross-polarized*) channel. Under ideal conditions, depolarization will not occur. When depolarization does occur, it can cause co-channel interference and cross-talk between dual-polarized satellite links. Rain is a primary cause of depolarization.

Both attenuation and depolarization come from interactions between the propagating electromagnetic waves and whatever is in the atmosphere at the time. The atmospheric constituents may include free electrons, ions, neutral atoms, molecules, and *hydrometeors* (an arcane term that conveniently describes any falling particle in the atmosphere that contains water: raindrops, snowflakes, sleet, hail, ice-crystals, graupel, etc.); many of these particles come in a wide variety of sizes. Their interaction with radio waves depends strongly on frequency, and effects that dominate 30 GHz propagation, for example, may be negligible at 4 GHz. The converse is also true. With one major exception (ionospheric effects) almost all propagation effects become more severe as the frequency increases.

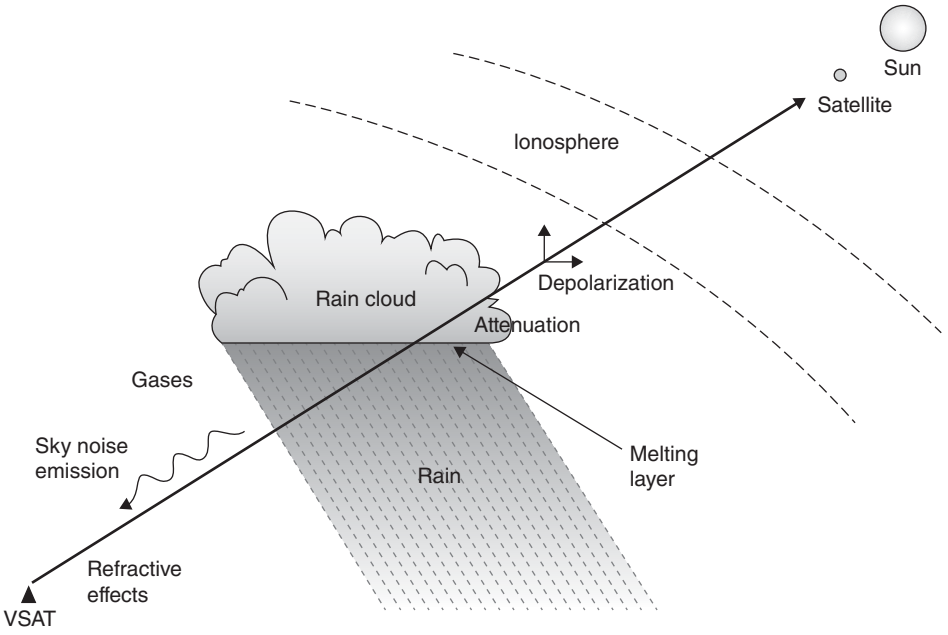


Figure 7.3 Illustration of the various propagation loss mechanisms on a typical earth-space path. The earth terminal (in this example a very small aperture terminal or VSAT) is directed toward the satellite. Refractive effects (causing tropospheric scintillation); gases; a rain cloud, melting layer, and rain all exist in the path and cause signal loss. The absorptive effects of the atmospheric constituents cause an increase in sky noise to be observed by the VSAT receiver. While atmospheric gases and tropospheric scintillation do not cause signal depolarization, collections of nonsymmetrical ice crystals and rain particles can depolarize the transmissions through them. Above the lower (neutral) atmosphere is the ionosphere, which begins at about 40 km and extends well above 600 km. The ionosphere can cause the electric vector of signals passing through it to rotate away from their original polarization direction, hence causing signal depolarization, At certain times of the day, year, 11-year sunspot cycle, the ionosphere can cause the amplitude and phase of signals passing through it to change rapidly, that is to scintillate, about a general mean level. The ionosphere has its principal impact on signals at frequencies well below 10 GHz while the other effects in the figure above become increasingly strong as the frequency of the signal goes above 10 GHz. Finally, if the sun (a very "hot" microwave and millimeter wave source of incoherent energy) is in the VSAT beam, an increased noise contribution results that may cause the CNR to drop below the demodulator threshold. Note: The above picture is not drawn to scale. Most rainstorms occur below 10 km altitude and the ionosphere is not normally present below 40 km and extends to more than 1000 km above the earth.

7.3 Quantifying Attenuation and Depolarization

Attenuation, A, is the decibel difference between the power received, P_r, at a given time t and the power received under ideal propagation conditions (often referred to as *clear sky* conditions). With all values in decibel units, we have

$$A(t) = P_{r_{\text{clearsky}}} - P_r(t) \tag{7.2}$$

Attenuation, $A(t)$, on satellite communications links operating at C-, Ku-, and Ka-band is primarily caused by absorption of the signal in rain. On most satellite links above 10 GHz, rain attenuation limits the availability of the system and, to develop an adequate

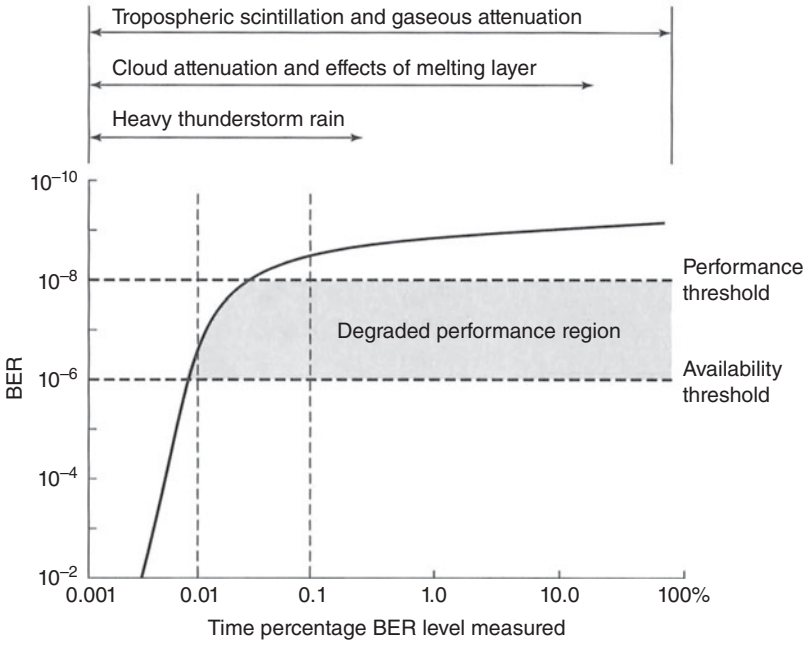


Figure 7.4 Approximate range of annual time percentages that various atmospheric impairments affect a link. (Source: After figure 2 of Allnutt 1999, © Wiley & Sons, Inc., Reprinted with permission). Tropospheric scintillation (a refractive effect in the lower atmosphere) and gaseous attenuation are pervasive phenomena that occur all of the time, but at different levels of impact depending on the climate, elevation angle, and time percentage of interest. Clouds exist at various time percentages, depending on the climate, but are generally present for at least 30% of the time at most locations. As the concentration of the frozen particles in the cloud increases, many will start to fall and will melt on reaching the 0°C isotherm. This will lead to enhanced attenuation in the melting layer. Drizzle rain will fall when the water vapor concentration reaches saturation levels. Such rain is usually stratiform and falls for between 1% and 10% of the time, depending on the climate. During hot periods, convective rain will fall, often in the form of thunderstorms. Heavy thunderstorms account for the highest rainfall rates, and hence the highest path attenuations encountered, but they exist for only small time percentages in a year. Not shown in the above figure are ionospheric effects, which have diurnal, seasonal, and 11-year cyclical impact, again depending on where the earth station is and the precise earth space path used.

link margin, the rain attenuation to be expected for a given percentage time needs to be calculated. This can be a complicated process, but there are basically three steps: (a) determine the rainfall rate for the percentage time of interest; (b) calculate the specific attenuation of the signal at this rainfall rate in dB/km; and (c) find the effective length of the path over which this specific attenuation applies. The difficult part of this process is part (c) because rain falls in two broad categories: stratiform rain and convective rain. These two separate atmospheric mechanisms have different effects on satellite paths. *Stratiform rain* is generated in cloud layers containing ice, and results in widespread rain or snow at rainfall rates of less than 10 mm per hour. *Convective rain* is generated by vertical air currents that can be very powerful, leading to thunderstorms and high rainfall rates. Convective rain is very important for satellite communication systems because it is the major cause of link outages. Stratiform rain consists of a generally constant rainfall rate over a very large area while convective rain is generally confined to a narrow, but

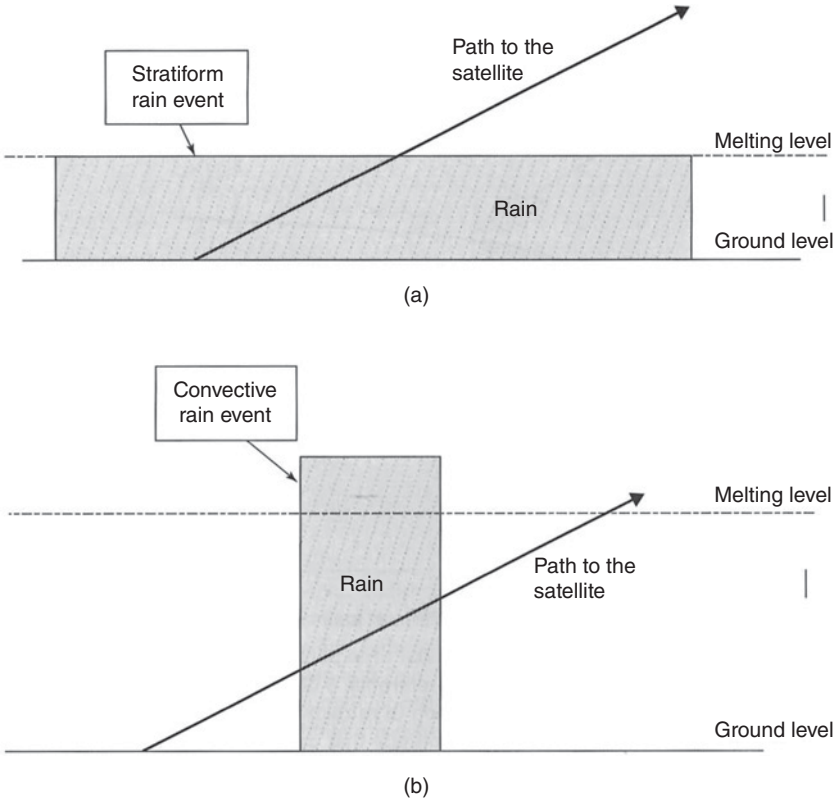


Figure 7.5 (***a***) Stratiform rain situation. In this case, a widespread system of stratiform rain – that is rain that appears to be stratified horizontally – completely covers the path to the satellite from the ground up to the point where the rain temperature is 0°C. This level is called the melting layer because, above it, the precipitation is frozen and consists of snow and ice crystal particles. Frozen precipitation causes negligible attenuation. In general, the signal path in stratiform rain will exit the rain through the top of the rain structure. (***b***) Convective rain situation. In this case, a tall column of convective rain enters the satellite-to-ground path. In some cases, the storm will be in front of the earth station; in others, behind it. Convective storms normally occur in the summer; thus, the melting level is much higher than in winter. In many cases, the melting level is not well defined, as the strong convective activity inside the storm will push the liquid rain well above the melting level height. Except for paths with very high elevation angles (>70°), the signal path in convective rain will most often exit from the side of a convective storm.

tall, column of rain. Figure 7.5 illustrates the two rain processes and Figure 7.6 gives the concept of the path attenuation calculation procedure for both rain types.

Stratiform rain occurs typically ahead of a warm front in an area of low pressure. Large areas of cloud exist in which ice crystals are sufficiently large to slowly fall and join other ice crystals to form snowflakes, which fall more quickly as their size increases. If there is a high concentration of moisture in the clouds, in the form of ice, large snowflakes may form. The snow falls until it reaches the *melting layer*. The melting layer is simply the region of the atmosphere where the temperature transitions from below 0°C to above 0°C. Snow falling into air at a temperature greater than 0°C melts and forms raindrops. If the air at the earth's surface is below 0°C the snow does not melt, but continues to the ground. The stratiform cloud mechanisms that generate snow result in low rainfall rates,

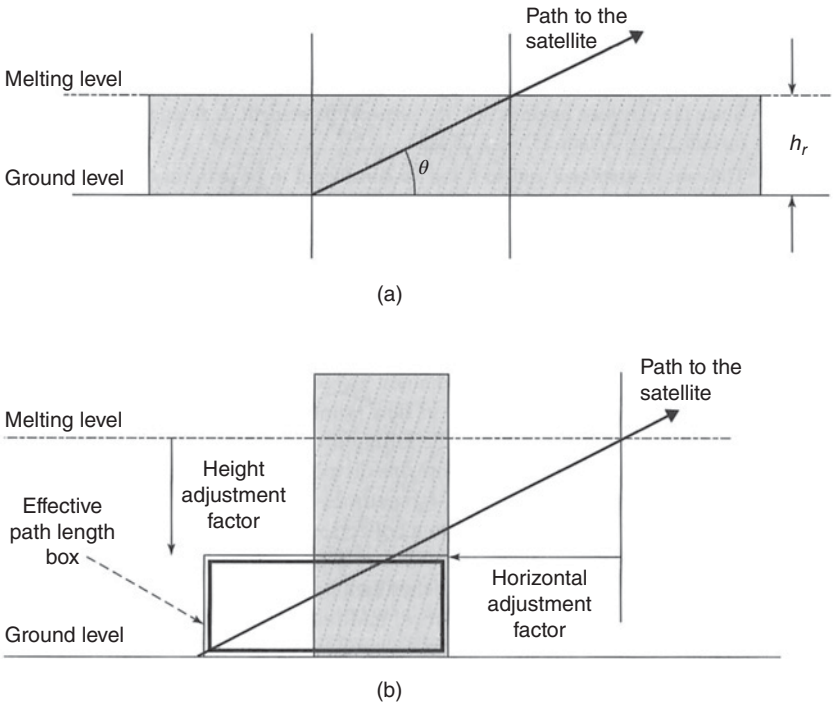


Figure 7.6 (***a***) Stratiform rain attenuation calculation procedure. In the case of stratiform rain, the rainfall rate along the path can be considered to be uniform and the path completely immersed in the rain. The *effective* path through the rain – the path over which the rain may be considered to be uniform – is therefore the same as the physical pathlength in stratiform rain. The path attenuation *A* is therefore the specific attenuation (i.e., dB attenuation per km) multiplied by the physical pathlength in the rain (i.e., $h_r / \sin\theta$). (**b**) Convective rain attenuation prediction procedure. In the case of convective rain, the melting level and elevation angle are used to develop two adjustment factors: a height adjustment factor and a horizontal adjustment factor. Once these factors have been used, a smaller box is created inside which it can be assumed that the rainfall rate is uniform. The length of the path that exists inside the box is the effective pathlength and it is this that is used to multiply the specific attenuation. In this case, the path exists through the top of the effective pathlength box. In other cases, it may exit through the side.

always less than 10 mm per hour, and widespread (stratiform) rain or snow. This leads to generally constant attenuation of the slant path signals over the entire path length from the ground to the melting layer.

Example 7.1

Question: An earth station at sea level communicates at an elevation angle of 35° with a geostationary earth orbit (GEO) satellite. The melting level height of the stratiform rain is 3 km. Find (a) the physical pathlength through the rain; (b) find the path attenuation if the specific attenuation is 2 dB/km.

Answer

a. The vertical height, h_r, of the rain is the difference between the melting level height (3 km) and the height of the earth station (0 km, since it is at sea level), which gives

$h_r = 3$ km. Since the elevation angle is 35°, the physical pathlength, L, through the rain is given by $L = h_r/(\sin 35) = 3/(\sin 35) = 5.23$ km.

b. The rain is stratiform and so it can be considered to be uniform over the path. The specific attenuation is therefore uniform along the path through the rain. If the specific attenuation is 2 dB/km, the path attenuation, A is given by $A = (2\text{ dB/km}) \times (5.23\text{ km}) = 10.46\text{ dB} \approx 10.5\text{ dB}$.

Convective rainstorms are very complex, and have both a horizontal and vertical structure. A convective cell becomes established when a mass of warm moist air is pushed up into colder air at a higher altitude. Adiabatic expansion of the air mass occurs, which cools the air. When the air is cooled below its dew point, it condenses forming clouds, and drops of water start to fall under gravity. The falling drops collide and coalesce with other drops to make larger drops, leading to a drop size distribution. The maximum size of stable raindrops is about 6 mm – larger drops (sometimes exceeding 10 mm in average diameter) are unstable and quickly break up into a collection of smaller drops under wind shear conditions. Large raindrops fall quickly, with terminal velocities up to 8 or 9 m/s. If the falling drops encounter supercooled water as they fall, hailstones can form. Hailstones can exceed the 10+ mm diameter limit of raindrops, and may reach golf-ball size in severe thunderstorms in the plain states of the United States. The accretion process can occur in an updraft as well as for falling drops, and in a vigorous thunderstorm, updraft velocities can exceed 100 mph. Since cold air is denser than warm air, once an updraft dies away at the top of a thunderstorm, cold air tends to flow downward, and can create a s*treamer*, a narrow region of intense rain and cold air. Streamers can be a few hundred meters wide or a kilometer wide. At the surface, the streamer is observed as a *microburst*, which has strong *wind shear* as the vertical down flow of cold air hits the ground and spills out in all directions. We are all familiar with microbursts. Shortly before heavy rain falls there is often a cold wind, followed by a downpour. The cold wind we feel is the outflow of cold air as it hits the earth's surface. The effect of convective rain on a satellite slant path depends on the angle at which the path intersects a streamer. Streamers are rarely vertical, so if a slant path is parallel to a streamer it will suffer very heavy attenuation if the streamer envelops the path. If the slant path cuts across the streamer, the path length within the heavy rain may be quite short, leading to relatively little attenuation despite the high rainfall rate.

Figure 7.7 shows an example of a convective rain cell observed with an S-band radar at Virginia Tech's satellite tracking station in Southern Virginia. The radar was used to make vertical scans across the slant path to a satellite (known to radar people as an *RHI* scan, for range height indicator, a WWII radar display mode). The complex shape of the storm cell requires the use of artificial "adjustment factors" to covert the physical path through the rainstorm to an effective pathlength over which the rain may be considered to be uniform. As well as causing significant attenuation, rain and ice crystals can cause depolarization.

Microbursts are dangerous for aircraft flying close to the ground, especially when taking off and landing. Several serious accidents to passenger aircraft in the 1980s were attributed to the wind shear associated with microbursts, and extensive research was carried out to develop ways to detect microbursts and wind shear. Networks of

anemometers, which measure wind speed, can be deployed around an airfield to detect wind shear, and *terminal Doppler radar* can be used for the same purpose.

An aircraft on final approach to a runway is flying slowly and descending on a 3° glide slope. If the aircraft encounters a microburst, it first experiences a headwind, which increases its speed relative to the air and tends to slow the rate of descent. The natural reaction of the pilot is to reduce engine power to maintain a constant rate of descent on the 3° glide slope. However, as soon as the aircraft passes the center of the microburst, the wind direction is opposite, and is now a tailwind, which reduces the speed of the aircraft relative to the air and increases the rate of descent of the aircraft. If the engine power has been cut the airplane may sink into the ground before the engines have developed enough power to keep the airplane aloft. Wind shear detection equipment at airfields and improved pilot awareness of the dangers of microbursts has reduced the incidence of accidents caused by microbursts.

Depolarization is more difficult to quantify than attenuation. All signals have a polarization orientation that is defined by the electric field vector of the signal (see Figure 7.8). In general, signals are never purely polarized; the direction of the electric field will never be perfectly oriented or constant. Successful orthogonal polarization frequency sharing – usually called dual polarization frequency re-use – requires that there be sufficient isolation between two orthogonal polarization states to permit the separation of the wanted polarization (the co-polarized signal) from the unwanted polarization (the cross-polarized signal) at the receiving antenna (Appendix B). The difference between the co-polarized and the cross-polarized signal energy will determine the cross-polarization discrimination at the receiver, the *XPD*, and hence the level of interference between two orthogonally polarized signals.

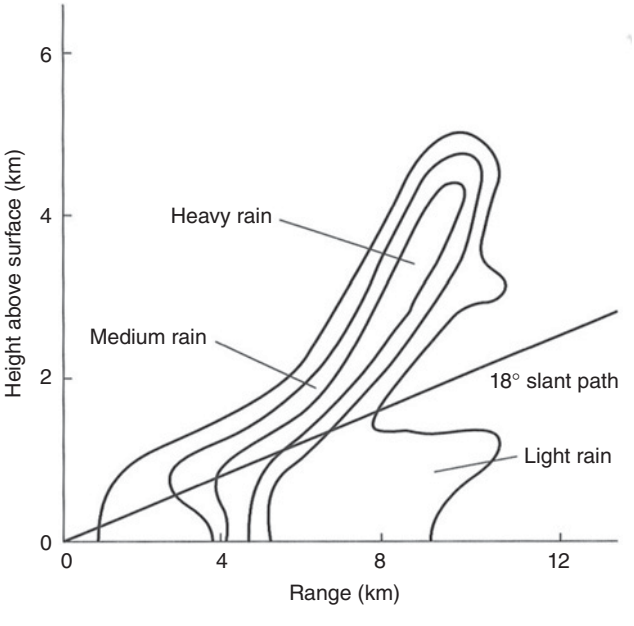


Figure 7.7 Example of an RHI scan through a rain storm. Radar reflectivity contours in a rainstorm on 15 June 1986, measured with an S-band radar in Blacksburg, Virginia. The contours represent light, medium, and heavy rain in a narrow vertical column. The radar and receiving station were collocated at the (0,0) point. Note the narrow extent of heavy rain in a sloping column, and the effect on the slant path to the satellite at an elevation angle of 18°.The statistical rain height H_i for Blacksburg is 4.1 km. In this example, rain is present up to an altitude of 5.6 km above sea level.

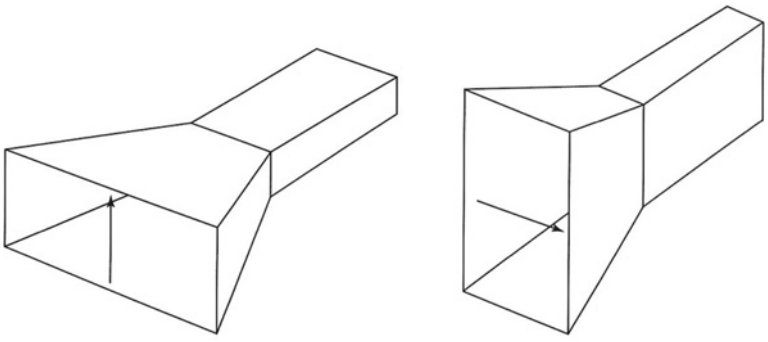


Figure 7.8 Orthogonally polarized waveguide horn antennas. The polarization of an electromagnetic wave is defined by the orientation of the electric vector. In the example above, two waveguide horns, excited in the TE_{10} mode, are radiating in the same direction. The left hand horn is oriented such that the electric vector is vertically polarized; the right hand horn is turned on its side compared with the left hand horn and so the electric vector is horizontally polarized. The arrows indicate the electric field vector. Since the electric polarization vectors are oriented 90° with respect to each other in the two horns, the transmitted signals are considered to be orthogonally polarized. Orthogonally polarized signals do not interfere with each other, even if they are at the same frequency, provided they are "purely" polarized (i.e., there is no component of the signal present in the other, orthogonal, polarization). In all cases, however, the transmitted signals are not purely polarized, due to antenna imperfections, so a component exists in the unwanted polarization. In addition, some of the energy in one polarization can "cross" over to the other polarization due to asymmetric particles (e.g., large oblate raindrops) existing in the propagation path. The cross-polarized energy can give rise to interference between the two, mutually orthogonal polarizations. The degree of cross-polarization to be expected along a given path is predicted using cross-polarization models that are usually based on the rain attenuation along the path.

To illustrate the process by which depolarization is measured; imagine a dual-polarized antenna transmitting orthogonally polarized signals. We will call the two polarizations *V* (for vertical) and *H* (for horizontal) for convenience, although there are infinitely many orthogonal polarization pairs. Let the complex phasor amplitudes of the transmitted electric field vectors with polarization *V* and *H* be **a** and **b**, respectively, as shown in Figure 7.9. The transmitting antenna is excited so that **a** and **b** are equal.

If the transmission medium between the transmitting and receiving antennas were clear air, phasor **a** would give rise to a *V* polarization wave of amplitude $\mathbf{a_c}$ at the receiving antenna and phasor **b** would cause an *H* polarization wave of amplitude $\mathbf{b_c}$. The subscript *c* stands for co-polarized; these fields have the same polarization sense as their transmitted counterparts (see Figure 7.10).

If asymmetrical rain or ice crystal particles exist in the transmission medium, some of the energy in **a** will couple into a small (cross-polarized) *H* polarized field component whose amplitude at the receiving antenna is $\mathbf{a_x}$, and **b** will give rise to a small (cross-polarized) *V* polarized component $\mathbf{b_x}$. An ideal receiving system that introduces no cross polarization will have a *V* channel output ($\mathbf{a_c} + \mathbf{b_x}$) and an *H* channel output ($\mathbf{b_c} + \mathbf{a_x}$). The unwanted $\mathbf{b_x}$ term represents interference with the wanted signal $\mathbf{a_c}$ and the unwanted $\mathbf{a_x}$ term is interference with the wanted signal $\mathbf{b_c}$. This interference will cause cross talk on an analog link and increase the BER on a digital link. This generation of unwanted cross-polarized components is called depolarization.

The measure of depolarization that is most useful for analyzing communications systems is the *cross-polarization isolation*, *XPI*. In terms of the complex phasor field

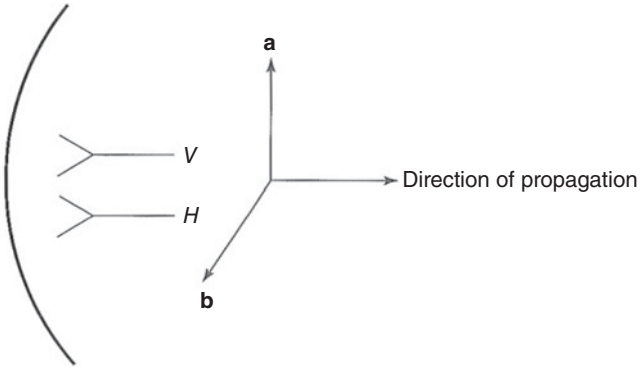


Figure 7.9 Fields excited by a dual-polarized antenna. The field radiated by the *V* horn has the vertically polarized electric field vector indicated by **a** and the field radiated by the *H* horn has the horizontally polarized vector indicated by **b**. In most antenna systems, one horn is used to radiate both polarizations simultaneously rather than two. This permits the single feed horn to be located at the prime focus of the antenna to generate the best far field pattern (see Appendix B). The two polarization senses to be transmitted are excited in separate parts of the transmitter and are then coupled together via an ortho-mode transducer into a single waveguide section. This waveguide section, which can support both polarizations simultaneously, is used to couple the signals into a waveguide horn that is capable of radiating both polarizations equally.

amplitudes, XPI is given by Eq. (7.3) for the *V* polarized channel and by Eq. (7.4) for the *H* polarized channel.

$$XPI_V = \mathbf{a}_c/\mathbf{b}_x \tag{7.3}$$

$$XPI_H = \mathbf{b}_c/\mathbf{a}_x \tag{7.4}$$

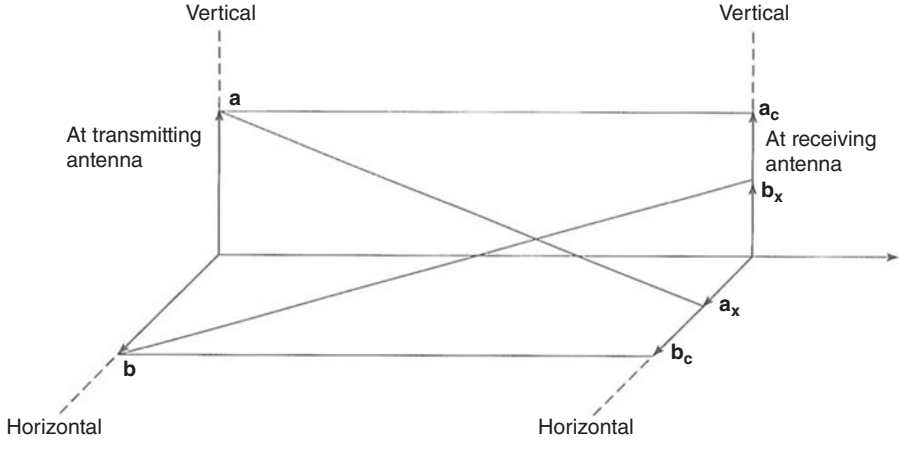


Figure 7.10 Illustration of signal depolarization in the transmission path. The transmitted fields **a** and **b** produce co-polarized components $\mathbf{a_c}$ and $\mathbf{b_c}$ at the receiving antenna. The transmission medium in this instance is not clear sky, nor is the transmitting antenna perfectly polarized, and the anisotropy of the transmission medium and the imperfections in the transmitting antenna induce cross-polarized components of the transmitted signal to be received. These cross-polarized components at the receiving antenna are $\mathbf{a_x}$ and $\mathbf{b_x}$. With perfect antennas and in the absence of depolarization, $\mathbf{a_x}$ and $\mathbf{b_x}$ would be zero.

The *XPI* values are commonly expressed in decibels; for example

$$XPI_V = 20\log_{10}|\mathbf{a}_c/\mathbf{b}_x|\text{dB} \tag{7.5}$$

Physically, the *XPI* is the decibel ratio of wanted power to unwanted power in the same channel. The larger the *XPI* value, the less interference there is and the better the communications channel will perform. *XPI* is difficult to measure. It requires the simultaneous transmission of signals at the same frequency in both polarization senses. The COMSTAR series of satellites had a beacon that rapidly switched between two orthogonal polarization senses, thus permitting the measurement of XPI (e.g., Cox and Arnold 1984). The European Space Agency (ESA) satellite OLYMPUS and the US satellite ACTS (ACTS 1999) also incorporated a switched beacon to permit XPI measurements. Most propagation experiments are much simpler than this and measure simultaneously the wanted (the co-polarized) and the orthogonal, unwanted (the cross-polarized) signals that are received from a satellite beacon that transmits in only one polarization. In this case (referring to Figure 7.10) the experiment would measure (say) signals $\mathbf{a_c}$ and $\mathbf{a_x}$ that are derived from a singly polarized signal **a** that is transmitted from the satellite. Measuring received signals $\mathbf{b_c}$ and $\mathbf{b_x}$ simultaneously from a singly polarized signal **b** would provide the same result. This process allows the *cross-polarization discrimination*, *XPD* to be derived

$$XPD_V = \mathbf{a}_c/\mathbf{a}_x \tag{7.6}$$

or in decibels

$$XPD_v = 20\log_{10}|\mathbf{a_c}/\mathbf{a_x}|\text{dB} \tag{7.7}$$

In most transmission situations encountered in practice, the values calculated for XPI and XPD are the same (Watson and Arabi 1973) and they are sometimes simply called the *isolation*. In practice, real antennas do not transmit polarization pairs that are exactly orthogonal, nor does the isolation remain the same over the 3 dB beamwidth of the antenna. Receiving antennas can also introduce cross-polarization. There is therefore a residual XPD component present even in clear sky conditions. This must be accounted for in the link budget of a dual-polarized, frequency re-use system. The residual XPD on axis is normally better for linearly polarized (LP) antennas (~30 to 35 dB) than for circularly polarized antennas (~27 to 30 dB). These values represent antennas carefully designed for dual-polarized operation; inexpensive antennas will typically exhibit about 20 dB XPD for linear or circular polarizations.

7.4 Propagation Effects That Are Not Associated With Hydrometeors

In this section we will discuss propagation effects that are not associated with raindrops or ice crystals: atmospheric absorption, cloud attenuation, refractive effects that include tropospheric scintillation and low angle fading/multipath effects, Faraday rotation, and ionospheric scintillation.

7.4.1 Atmospheric Absorption

At microwave frequencies and above, electromagnetic waves interact with molecules in the atmosphere to cause signal attenuation. At certain frequencies, resonant absorption occurs and severe attenuation can result. Figure 7.11 (from Figure 5 of (ITU-R

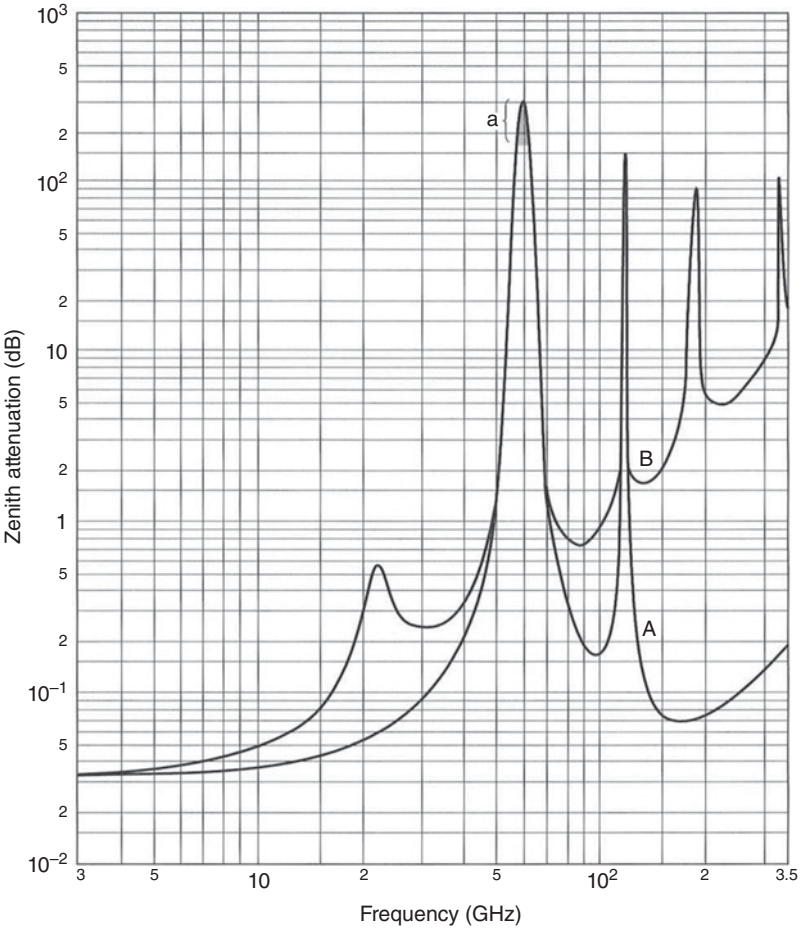


Figure 7.11 Total zenith attenuation due to atmospheric gases calculated from 3 to 350 GHz. (Source: From figure 5 of reference ITU-R Recommendation P.676-11 2016, reproduced with permission). The three curves represent water vapor absorption, dry atmospheric absorption, and the total absorption that would be observed looking straight up from sea level (i.e., on a zenith path) right through the neutral atmosphere on a satellite-earth path. The water vapor curve is for a standard atmosphere that consists of a surface pressure of 1013 hPa (a hPa has the same numerical value as the old pressure unit of millibars), a surface temperature of 15 °C, and a surface relative humidity of 7.5 mg/m^3. The dry atmospheric curve shows only the resonant absorption peaks of oxygen molecules (a broad peak at 60 GHz and a narrow peak at 118.75 GHz). The total absorption curve includes the resonant absorption peaks due to the water vapor molecule at 22.235, 183.31, and 325.153 GHz. The broad peak absorption of the dry atmosphere around 60 GHz consists of many individual peak absorption lines that do not become apparent until the atmospheric pressure is reduced (i.e., the data are obtained at a high altitude). Specific attenuation due to atmospheric gases. (Pressure = 1013.25 hPa; Temperature = 15 °C; Water Vapor Density = 7.5 g/m^3). Source: Reproduced with permission of ITU-R.

P.676-11, 2016)) shows these resonant absorption peaks on a zenith path (that is a path at an elevation angle of 90°) from a sea level location right through the neutral atmosphere. Neutral means that no ionization is present.

The first absorption band in Figure 7.11 is that due to water vapor at 22.235 GHz. The K-band sets of frequencies are on both sides of this absorption band, which has led to the terminology of Ku-band (signifying frequencies under the absorption band) and Ka-band (signifying frequencies above the absorption band). It is common to specify a satellite frequency band by the uplink frequency. From Figure 7.11, it can be seen that gaseous absorption accounts for less than 1 dB on most paths below 100 GHz that lie outside the absorption bands. However, in many new systems that employ very small system margins, it is important to account for the gaseous losses along the anticipated path. New prediction procedures that attempt to account for all attenuating phenomena along the path (e.g., Dissanayake et al. 1997) include gaseous absorption.

7.4.2 Cloud Attenuation

Once considered to be largely irrelevant for satellite communications paths, clouds have become an important factor for some Ka-band paths and all V-band (50/40 GHz) systems. The difficulty with modeling cloud attenuation is that clouds are of many types and can exist at many levels, each type having a different probability of occurrence. The water droplet concentrations in each cloud will also vary, and clouds made up of ice crystals cause little attenuation. Typical values of cloud attenuation for water-filled clouds are between 1 and 2 dB at frequencies around 30 GHz on paths at elevation angles of close to 30° in temperate latitudes (ITU-R 840-5). In warmer climates, where clouds are generally thicker in extent and have a greater probability of occurrence than temperate latitudes, cloud attenuation is expected to be higher. As with most propagation effects, the lower the elevation angle, the higher the cloud attenuation.

7.4.3 Tropospheric Scintillation and Low Angle Fading

The atmosphere close to the ground, sometimes called the boundary layer, is rarely still. Energy from the sun warms the surface of the earth and the resultant convective activity agitates the boundary layer. This agitation results in turbulent mixing of different parts of the boundary layer, causing small-scale variations in refractive index. Figure 7.12 illustrates the process.

When a signal encounters a turbulent atmosphere, the rapid variation in refractive index along the path will lead to fluctuations in the received signal level. These fluctuations are generally about a fairly constant mean signal level and are called *scintillations*. Because the bulk of the fluctuations are caused within 4 km of the earth's surface, that is, within the troposphere, they are referred to as *tropospheric scintillations*. Tropospheric scintillations occur in all weather conditions, the drier the air, the smaller the scintillation amplitude, and vice versa. Tropospheric scintillation is an *on axis* effect, that is, it is not caused by variations in angle-of-arrival of the signal at the receiving antenna, and it does not cause signal depolarization. The magnitude of the scintillations becomes generally larger as the frequency increases, the path elevation angle reduces, and the climate becomes warmer and more humid. Prediction models exist to calculate this phenomenon with good accuracy (ITU-R P.618-13, 2003). On

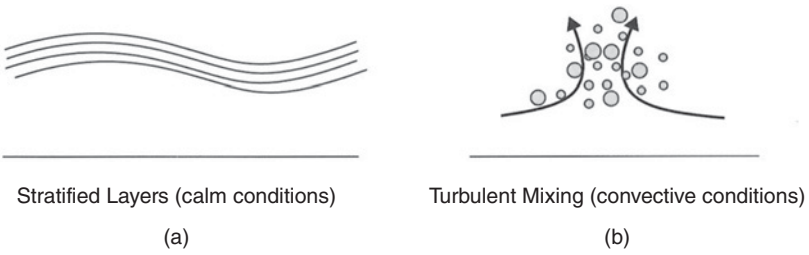


Figure 7.12 Schematic of stratified and turbulent conditions in the *boundary layer* of the atmosphere. In (*a*), the air is calm and the lower atmosphere next to the earth's surface (the boundary layer) forms into layers. Each layer has a slightly different refractive index, decreasing in general with height. In (*b*), the earth's surface has become heated by energy from the sun and the resultant convective activity has mixed the formerly stratified layers into "bubbles" that have different refractive indices. The turbulent mixing of the lower atmosphere will cause relatively rapid fluctuations in a signal passing through it, which are called *scintillations*.

paths below 10° elevation angle, tropospheric scintillation can be performance limiting; below 5° elevation angle it can become availability limiting.

When the elevation angle falls below 10°, a second propagation effect becomes noticeable: low angle fading. Low angle fading is the same phenomenon as multipath fading on terrestrial paths. A signal transmitted from a satellite arrives at the earth station receiving antenna via different paths with different phase shifts. On combination, the resultant waveform may be enhanced or attenuated from the normal clear sky level. Signal enhancement has been observed to exceed 8 dB on a 3.3° path at 11.198 GHz (Johnston et al. 1991), while cancelation can cause complete link dropout. The mechanism for low angle fading has been interpreted as atmospheric multipath and also as the "defocusing and focusing" of the incoming signal. Both explanations have merit: the received signal is made up of components that have arrived via different paths (i.e., multipath), but the mechanism for developing the different paths is one of refraction rather than reflection at the atmospheric layer boundaries. Low angle fading is only significant in very still air on very low elevation angle paths. It is normally not considered for satellite paths when the elevation angle is above 10°. Note that the multipath effect referred to here is occurring in the atmosphere, and is therefore different from multipath effects in terrestrial radio links which are caused by reflections from the ground, buildings, trees, and so on. Examples of tropospheric scintillation, both in apparent clear air and in rain, plus a significant rain attenuation event are depicted in Figure 7.13.

7.4.4 Faraday Rotation in the Ionosphere

The ionosphere is that portion of the earth's atmosphere that contains large numbers of electrons and ions. At its lowest, it reaches down to close to 40 km above the earth; there is no distinct upper boundary, but it exists well above 600 km above the earth. The ionosphere completely dominates radio propagation below about 40 MHz, but its effects on the frequencies used by most communications satellites are minor, except in periods around the equinox and in high sunspot activity on the sun. Even then, the effects are only significant at frequencies around 4 GHz, or below.

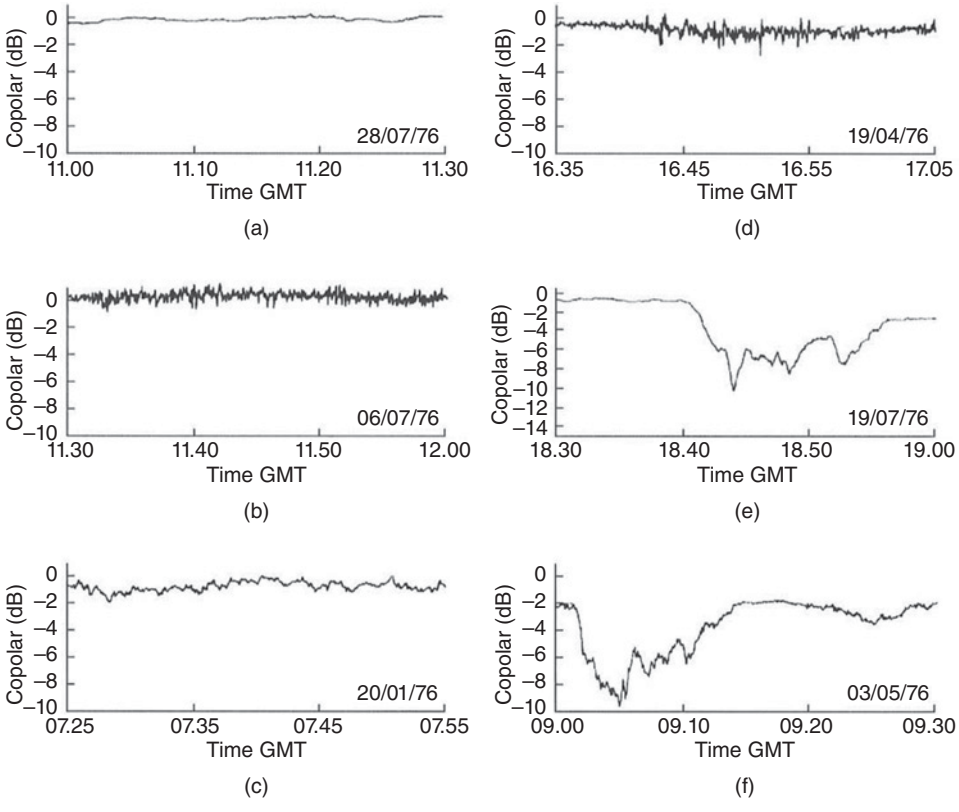


Figure 7.13 Scintillations observed under a variety of weather conditions on a 30 GHz downlink from ATS-6. Scintillations with various amplitudes can be observed under different weather conditions. Two of the data sets were taken in clear weather, two in cloud conditions, and two in rain, as follows: (*a*) clear-weather co-polar scintillation with low scintillation (*b*) clear-weather co-polar signal with high scintillation (*c*) co-polar scintillation in cloud; (*d*) co-polar scintillation in cloud; (*e*) co-polar scintillation and attenuation in rain; (*f*) co-polar scintillation and attenuation in rain. Note the difference in scintillation amplitude under what are apparently similar weather conditions along the path.

Electrically, the ionosphere is an inhomogeneous and anisotropic plasma and an exact analysis of wave propagation through it is extremely difficult. For a given frequency and direction of propagation with respect to the earth's magnetic field, there exist two characteristic polarizations. Waves with these polarizations, called *characteristic waves*, propagate with their polarization unchanged. Any wave entering the ionosphere can be resolved into two components with the characteristic polarizations. The phase shift and attenuation experienced by the characteristic waves can be calculated at any point along the propagation path, and the total field can be computed as the vector sum of the fields of the characteristic waves. This total field can be interpreted as an attenuated and depolarized version of the wave that entered the ionosphere. Thus, when a LP satellite path signal reaches the ionosphere, it excites waves with the two characteristic polarizations. These travel at different velocities, and when they leave the ionosphere their relative phase is different from when they entered. The wave that leaves the

ionosphere has a different polarization from the LP wave that was transmitted. This is called *Faraday rotation*, and its effect is essentially the same as if the field vector of the transmitted LP wave had been rotated by an angle φ. For a path length through the ionosphere of Z meters, the rotation angle φ is given by

$$\varphi = \int \left(\frac{2.36 \times 10^4}{f^2} \right) ZNB_0 \cos\theta dz \text{ rad} \tag{7.8}$$

Here, θ is the angle between the geomagnetic field and the direction of propagation, N is the electron density in electrons/cubic meter, B_0 is the geomagnetic flux density in Teslas, and f is the operating frequency in Hz. The rotation angle φ varies inversely with f^2. Table 7.1 gives the value of φ and some other parameters with frequency (ITU-R P.618-11, 2013).

The polarizations of an earth station antenna can be adjusted to compensate for the Faraday rotation observed under average conditions. However, the rotation of the uplink will be in an opposite sense to that on the downlink and so, to compensate in both directions at the same time, a feed will be required that is able to rotate the relevant sections in opposite directions. The *XPD* that results when the polarization angle of an LP wave changes by an amount $\Delta\varphi$ is given by

$$XPD = 20\log_{10}(\cot\Delta\varphi) \tag{7.9}$$

Hence, a 6° change from average conditions would reduce the *XPD* on the link to about 19.6 dB.

7.4.5 Ionospheric Scintillations

Energy from the sun causes the ionosphere to "grow" during the day, increasing the total electron content (TEC) by two orders of magnitude, or more. The TEC is the total number of electrons that would exist in a vertical column of area 1 m^2 from the surface of the earth all the way through the earth's atmosphere. Typical values of TEC range from $\sim 10^{18}$ during the day to $\sim 10^{16}$ during the night. It is the rapid change in TEC from the daytime value to the nighttime value, which occurs at local sunset in the ionosphere, that gives rise to irregularities in the ionosphere. The irregularities cause the signal to vary rapidly in amplitude and phase, which leads to rapid signal fluctuations that are called *ionospheric scintillations*. The magnitude of the ionospheric scintillations varies with time of day, month in the year, and year in the 11-year sunspot cycle. The greatest scintillation effects are observed just after local sunset in the equinox periods during the sunspot maximum years. The effects are also worst within about ±20° of the geomagnetic equator and over the poles. The length of the cycle averages at around 11 years, but has been as short as 9.5 years and as long as 12.5 (Mursula and Ulich 1998). Solar sunspot cycle 22 was from 1986.8 to 1996.4. Solar sunspot cycle 25 will start in late 2019, and the peak is expected to be in 2024.

7.5 Rain and Ice Effects

At frequencies above 10 GHz, rain is the dominant propagation phenomenon on satellite links. Many experiments have been conducted on geostationary satellite links, using experimental satellites such as SIRIO, OTS, and CTS (Hermes) at Ku-band and

Table 7.1 Estimated ionospheric effects for elevation angles of about 30° one-way traversal

Effect	Frequency dependence	0.1 GHz	0.25 GHz	0.5 GHz	1 GHz	3 GHz	10 GHz
Faraday rotation	$1/f^2$	30 rotations	4.8 rotations	1.2 rotations	108°	12°	1.1°
Propagation delay	$1/f^2$	25 μs	4 μs	1 μs	0.25 μs	0.028 μs	0.0025 μs
Refraction	$1/f^2$	<1°	<0.16°	<2.4′	< 0.6′	< 4.2″	< 0.12″
Variation in the direction of arrival (r.m.s,)	$1/f^2$	20′	3.2′	48″	12″	1.32″	0.12″
Absorption (auroral and/or polar cap)	$\approx 1/f^2$	5 dB	0.8 dB	0.2 dB	0.05 dB	6×10^{-3} dB	5×10^{-4} dB
Absorption (mid-latitude)	$1/f^2$	<1 dB	<0.16 dB	<0.04 dB	<0.01 dB	<0.001 dB	$<1 \times 10^{-4}$ dB
Dispersion	$1/f^3$	0.4 ps/Hz	0.026 ps/Hz	0.0032 ps/Hz	0.0004 ps/Hz	1.5×10^{-5} ps/Hz	4×10^{-7} ps/Hz
Scintillation	See Rec. ITU-R P.531	See Rec. ITU-R P.531	See Rec. ITU-R P.531	See Rec. ITU-R P.531	>20 dB peak-to-peak	≈ 10 dB peak-to-peak	≈ 4 dB peak-to-peak

The above estimates are derived from ITU-R P.531.
The values in the table are based on a TEC of 10^{18} electrons/m^2, which is a high value encountered in low latitudes in daytime with high solar activity.
The scintillation values are for near the equator during early night-time hours (local time) at equinox under conditions of high sunspot number.
Source: Reproduced with permission from Table 1 in ITU-R 618-11 2013.

ATS-6, OLYMPUS, and ACTS at Ka-band. One experimental satellite, ITALSAT, also allowed 50/40 GHz (V-band) experiments to be conducted in Europe. Allnutt (2011) provides detailed results of these experiments and explanations of all the propagation phenomena.

7.5.1 Characterizing Rain

Most farmers, hydrologists, and city planners need to know how much total rain will fall in a given period: that is, the rain *accumulation*. Indeed, most weather forecasts are given in terms of how much precipitation will fall (or accumulate) over a given region. Rain accumulation, unfortunately, is of little use to satellite link designers, since it is the rate at which the rain is falling that is important: that is, the *rainfall rate*. Rainfall rate is measured by a rain gauge, the most common of which is a tipping bucket rain gauge. This is fairly accurate between rainfall rates of 10–100 mm/h. Peak values of 100–150 mm/h may be expected for short periods during summer thunderstorms in the Mid-Atlantic region of the United States. Higher rainfall rates are observed in tropical regions.

The long-term behavior of rainfall rate is described by a *cumulative probability distribution* or by a *cumulative distribution function* (*cdf*). The cdf for rainfall rate is commonly referred to as an *exceedance curve*. This gives the percentage of time (usually the percentage of one year) that the rainfall rate exceeds a given value. Climate related parameters tend to be very variable, particularly as the earth seems to be entering a period of less predictable weather patterns. Rain accumulation can vary significantly from year to year, as can the exceedance curves, particularly at the low time percentages of interest to satellite link designers.

Four annual exceedance curves are shown in Figure 7.14. Three of them are individual annual exceedance curves in succeeding years (1979, 1980, and 1981), while the fourth is the average annual exceedance curve. The data were taken from an experiment performed at Blacksburg, Virginia, United States. It can be seen that the rainfall rate at the 0.01% point varies between 38 and 58 mm/h over the three years. Depending on the elevation angle of the link, this can make a significant difference in the attenuation measured at the same time percentage in each of the years. For this reason, link designers prefer to use values averaged over many years of measurements for their propagation models. Satellite path attenuation data do not exist over long time periods, so link designers have attempted to relate the attenuation exceedance curves to one parameter where long-term data exist: rainfall rate. The initial two approaches to developing these long-term attenuation statistics were rain climate maps and exceedance contour maps, both of which are discussed below. More recently, an approach to calculating rainfall rate exceedance using cumulative rainfall and mean surface temperatures as input parameters to a digital rainfall rate map (ITU-R P.837-7 2017).

7.5.1.1 Rain Climate Maps

Rain climate maps were the first approach to developing long-term rainfall rate statistics that could be used for both propagation predictions and interference calculations. These maps divide the world into regions where the average rainfall rate statistics are the same, within a margin of about ±10%. An example of a rain climate map is shown in Figure 7.15 for the Americas (from ITU-R P.837 1994). The rainfall rate statistics for

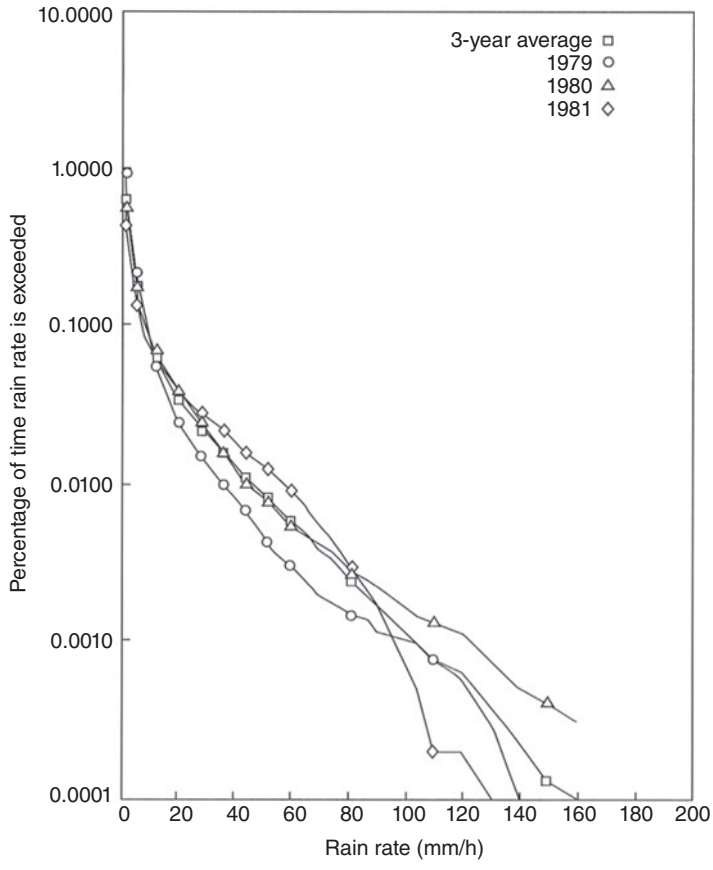


Figure 7.14 Typical rainfall rate cumulative probability distributions or *exceedance* curves. These sets were measured at Virginia Tech, Blacksburg, United States, as part of a three year experiment with the Italian satellite, SIRIO. The 1979 data indicate a relatively dry year, while those of 1981 indicate a relatively wet year. Despite this, a single, rare thunderstorm in 1979 produced much higher rainfall rates than those observed in 1981 at low time percentages. The availability level the link has to operate at will determine what rainfall rate is of most importance and it will also give a range over which the design must cope. For example, if 0.01% was the availability requirement, in 1979 the rainfall rate for this time percentage was 38 mm/h while in 1981 it was 58 mm/h. This shows the value of long-term statistics so that one year does not bias the link design.

the 15 rain climate regions worldwide are given in Table 7.2. (Note that not all of these regions exist in the Americas as can be seen in Figure 7.15). The ease with which the tables and rain climate maps can be used is offset by the clear inaccuracies that occur when large parts of the earth are given the same climate classification. The step-changes across the climate boundaries are also arbitrary and are not supported by measured data. Wherever possible, it is always best to use measured rainfall rate data as prediction model input whenever these data exist.

In an effort to overcome the inaccuracies of the rain climate maps, the ITU developed a set of comprehensive rainfall rate exceedance curves for the whole world. An initial set for the Americas is shown in Figure 7.16. These rain climate maps were superseded

Table 7.2 Rainfall Rate intensities in mm/h for the Rain Climatic Zones

Percentage of time (%)	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q
10	<0.1	0.5	0.7	2.1	0.6	1.7	3	2	8	1.5	2	4	5	12	24
0.3	0.8	2	2.8	4.5	2.4	4.5	7	4	13	4.2	7	11	15	34	49
0.1	2	3	5	8	6	8	12	10	20	12	15	22	35	65	72
0.03	5	6	9	13	12	15	20	18	28	23	33	40	65	105	96
0.01	8	12	15	19	22	28	30	32	35	42	60	63	95	145	115
0.003	14	21	26	29	41	54	45	55	45	70	105	95	140	200	142
0.001	22	32	42	42	70	78	65	83	55	100	150	120	180	250	170

Source: From Table 1 in ITU-R P.837-1 (1994). Reproduced with permission of ITU-R.

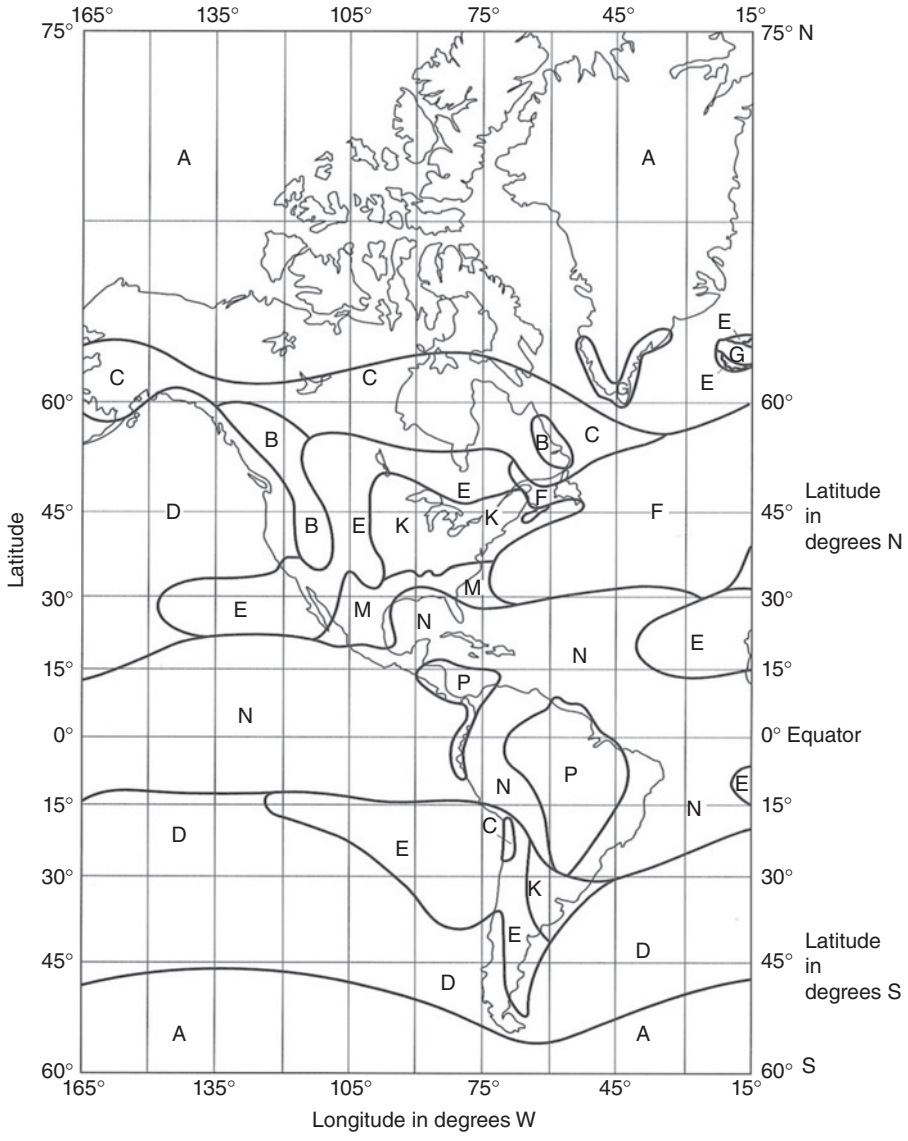


Figure 7.15 Rain climatic zones for the Americas. Source: From figure 1 of ITU-R Rec. P.837-1, 1994. Reproduced with permission of ITU-R.

by rain contour maps (ITU-R P.837-1 1994) and then by rain intensity contours (ITU-R P.837-2 1999). An example of the latter is shown in Figure 7.17.

7.5.1.2 Raindrop Size Distributions

Rain attenuation and depolarization occur because individual raindrops absorb energy from radio waves. The drops absorb some of the incident energy and some is scattered. The size and shape of raindrops have been measured (Laws and Parsons 1943). The most

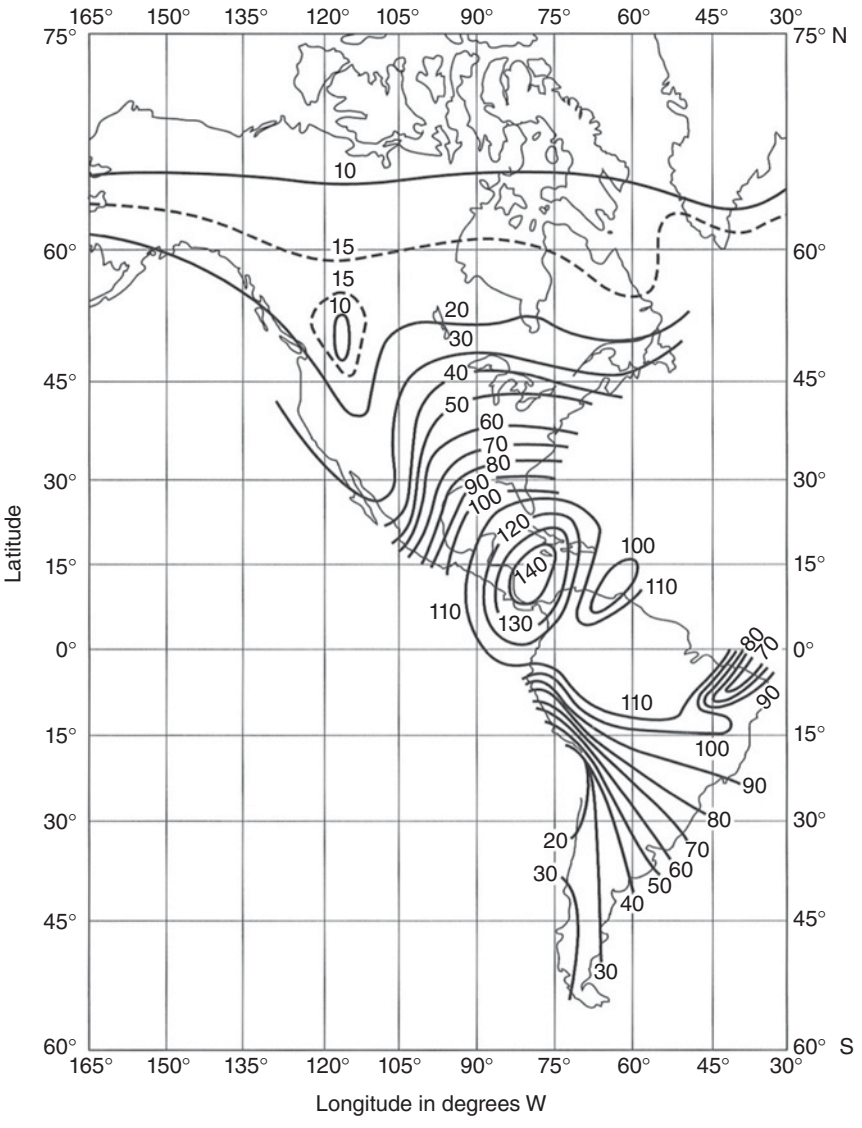


Figure 7.16 Rainfall rate exceedance contours for the Americas. (Source: Report 564 1982, © ITU-R, reproduced with permission). This was the first set of three rainfall rate exceedance contours that were developed for the world. In this version, the contours only existed over land. Later versions include data over the entire surface of the world (see Figure 7.17).

common mathematical description of the distribution of raindrop sizes is exponential and of the form

$$N(D) = N_0 e^{(-D/D_m)} \text{ mm}^{-1} \text{ mm}^{-3} \tag{7.10}$$

where D_m is the median drop diameter and $N(D)\,dD$ is the number of drops per cubic meter with diameters between D and $D + dD$ mm. The rainfall rate R is related to $N(D)$

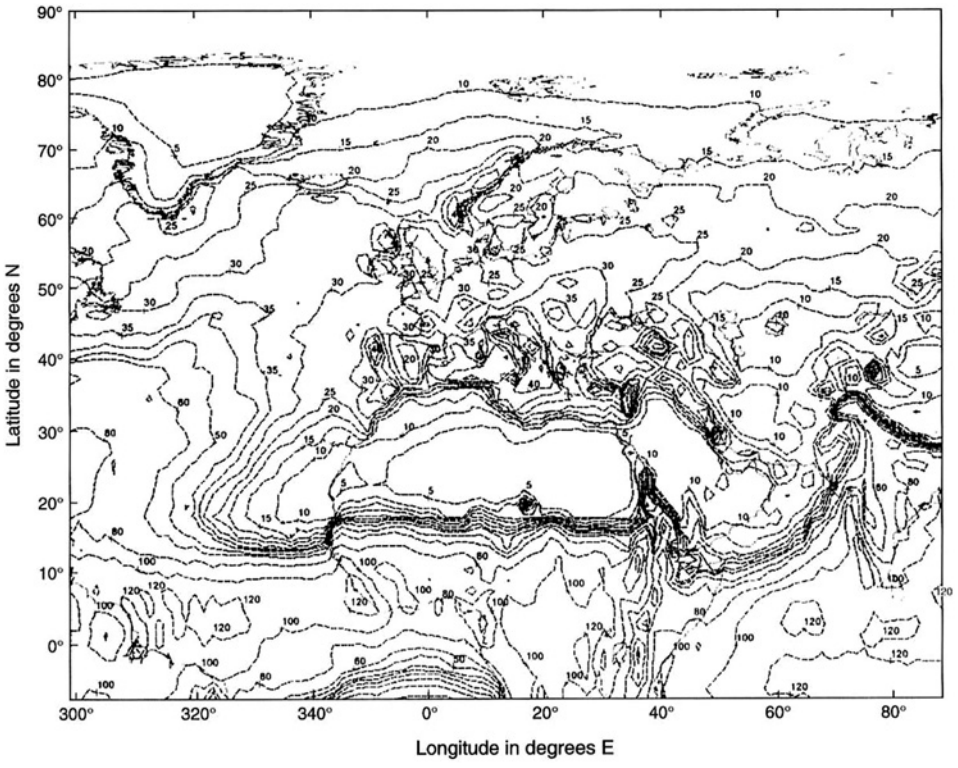


Figure 7.17 Rain intensity (mm/h) exceeded for 0.01% of the average year. Source: Figure 2 of ITU-R Rec. P.837-2, 1999. Reproduced with permission of ITU-R. This map provides rainfall rate contours for the Northern Hemisphere between longitude 300°E and 80°E (Europe, North Africa, the Middle East. And parts of Russia, India, and China.

and also to the terminal velocity $V(D)$ of the falling drops in meters per second with diameter D by (Hall 1979)

$$R = 0.6 \times 10^{-3} \pi \int D^3 V(D) N(D) dD \ \text{mm/h} \tag{7.11}$$

The details of scattering and absorption by a single raindrop, and the summation over the drop population that the calculation of path attenuation from the drop size distribution requires are beyond the scope of this text.

The first measurements of raindrop size distributions were made by Laws and Parsons (Laws and Parsons 1943) in 1943 with a very ingenious experiment. The experiment was designed to find the sizes of raindrops in typical rainstorms. A baking pan (typically 1 m × 0.5 m) was filled with flour and placed out in the rain for a minute. The pan of flour was then baked in an oven, and the loose flour sifted out. What remained were pellets of baked flour where raindrops had fallen and been absorbed by the flour, with dimensions that corresponded to the raindrops that hit the tray in that particular storm.

From their flour measurements, Laws and Parsons derived an empirical exponential relationship between the rain rate and the drop size distribution that is still used today. The experiment is repeated occasionally – by school children studying earth science – as a demonstration of rainfall characteristics.

7.6 Prediction of Rain Attenuation

Attenuation by rain can be predicted accurately if the rain can be precisely described all the way along the path. Path attenuation is essentially an integral of all the individual increments of rain attenuation caused by the drops encountered along the path. This is the physical approach to predicting rain attenuation. Unfortunately, rain cannot be described accurately along the path without extensive meteorological databases, which do not exist in most regions of the world. Most prediction models therefore resort to semiempirical approaches, which calculate an *effective pathlength* through the rain, L_{eff}, over which the rainfall rate is assumed to be constant. This constant rainfall rate leads to a constant specific attenuation, γ_{R}, and the path attenuation, A, is simply given by

$$A = \text{specific attenuation} \times \text{effective pathlength in rain} = \gamma_{\text{R}} L_{\text{eff}} \text{ dB} \tag{7.12}$$

The semiempirical approach is based upon two premises: (i) Rainfall rate measured at a point on the surface of the earth is statistically related (over a period of at least a year) to the attenuation encountered along the path to a satellite from that same point; (ii) The actual length of the path through the rain medium can be adjusted such that an effective pathlength is developed over which the rain can be considered to be homogeneous (see Figure 7.6).

The estimation of attenuation on the slant path to a satellite is essential to the process of establishing a margin in the link budget that ensures the required availability of the link. Over a period of many years, several attenuation models have been developed that have been widely used. These include the Crane Model (Crane 1980), the Simple Attenuation Model (Stutzman and Dishman 1982), the DAH model (Dissanayake et al. 1997) and several models published by the ITU, the latest of which is (ITU-R P.618-11 2013). The ITU-R model, based on the DAH model, is discussed in detail here, because it provides the most accurate statistical estimate of attenuation on slant paths, worldwide, at the time of writing (August 2018).

A power law equation describes the relationship between point rainfall rate R and *specific attenuation*, γ_{R}, the attenuation measured over 1 km (Olsen et al. 1978).

$$\gamma_{\text{R}} = k\,(\text{R}_{0.01})^{\alpha} \text{ dB/km} \tag{7.13}$$

In Eq. (7.13), the suffix 0.01 to R denotes the rainfall rate measured for 0.01% of the average year, a typical input time percentage for most models. Equation (7.13) holds for all values of rainfall rate, however. The parameters k and α are frequency dependent. Table 7.3 gives values for k and α for frequencies between 4 and 50 GHz (ITU-R P.838-3).

Example 7.2

Question: What is the specific attenuation at 10 GHz if the rainfall rate is 40 mm/h and linear vertical polarization is used?

Table 7.3 Regression coefficients for estimating specific attenuation

Frequency (GHz)	k_H	α_H	k_V	α_V
4	0.000 107 1	1.600 9	0.000 246 1	1.247 6
6	0.000 705 6	1.590 0	0.000 487 8	1.572 8
8	0.004 115	1.390 5	0.003 450	1.379 7
10	0.012 17	1.257 1	0.011 29	1.215 6
12	0.023 86	1.182 5	0.024 55	1.121 6
20	0.091 64	1.056 8	0.096 11	0.984 5
30	0.240 3	0.948 5	0.229 1	0.912 9
40	0.443 1	0.867 3	0.427 4	0.842 1
50	0.660 0	0.808 4	0.647 2	0.787 1

1. The suffices V and H refer to vertical and horizontal polarizations, respectively
2. Values of k and α at frequencies other than those in the table can be obtained by interpolation using a logarithmic scale for frequency and k, and a linear scale for α
3. Values have been tested and found to be accurate up to 50 GHz. ITU-R 838-3 has values for these parameters up to a frequency of 1000 GHz
4. For linear and circular polarization, and for all path geometries, the coefficients in Eq. (7.12) can be calculated using the values in the above table and the following equations [ITU-R P.838-3]

$$k = [k_H + k_V + (k_H - k_V)\cos^2\ \theta\ \cos\ 2\tau]/2$$

$$\alpha = [k_H\alpha_H + k_V\alpha_V + (k_H\alpha_H - k_V\alpha_V)\ \cos^2\theta\ \cos\ 2\tau]/2\,k$$

where θ is the path elevation angle and τ is the polarization tilt angle relative to the horizontal ($\tau = 45°$ for circular polarization).

Source: From Table 5 in ITU-R P.838-3. Reproduced with permission of ITU-R.

Answer

From Table 7.3, $k_V = 0.011\ 29$ and $\alpha_V = 1.2156$ at a frequency of 10 GHz. Using Eq. (7.13), we therefore have

$$\text{Specific attenuation} = \gamma_R = 0.011\ 29\ (40)^{1.2156} = 1.000\ 364\ 03 = 1.0\ \text{dB/km}$$

If the rainfall rate were constant along the path, as it generally is in light, stratiform rain (see Figure 7.5), then calculating the total attenuation for a given rainfall rate would be simple. The physical pathlength through the rain, L, would be the same as the effective pathlength and the total attenuation, A, is given by

$$A = \gamma_R \times \text{physical pathlength in rain} = \gamma_R \times \text{L dB}$$

On short terrestrial paths (<5 km, although this varies with rainfall rate: the lower the rainfall rate, the longer the path), the pathlength through relatively constant rain can be taken as the distance between the transmitting and receiving antennas. The path through the rain is also at almost the same height along the whole path. This is not the case with satellite paths where the signal follows a slanting path through the atmosphere, and encounters rain of different types and intensities on the way. Rain can take more than 10 minutes to fall from a height of 5 km (the approximate upper limit of liquid water in a severe thunderstorm) to the ground. If there are updrafts present, as is always the case in convective rain, it can take even longer. There is therefore no instantaneous relationship

between attenuation measured along a path to a satellite and the rainfall rate measured at the earth station site. However, there is a strong statistical relationship between the long-term cumulative statistics of rainfall rate and the long-term statistics of slant-path attenuation. Many models of rain attenuation use *equiprobable values* of rainfall rate and path attenuation to determine the cumulative statistics of attenuation from those of rainfall rate. Figure 7.18 illustrates the procedure for finding equiprobable values of rainfall rate and path attenuation.

The assumption that point rainfall rate on the ground is statistically related (over a period of at least a year) to the attenuation observed on a satellite path to that same point has been validated in many experiments worldwide. Since the path encounters

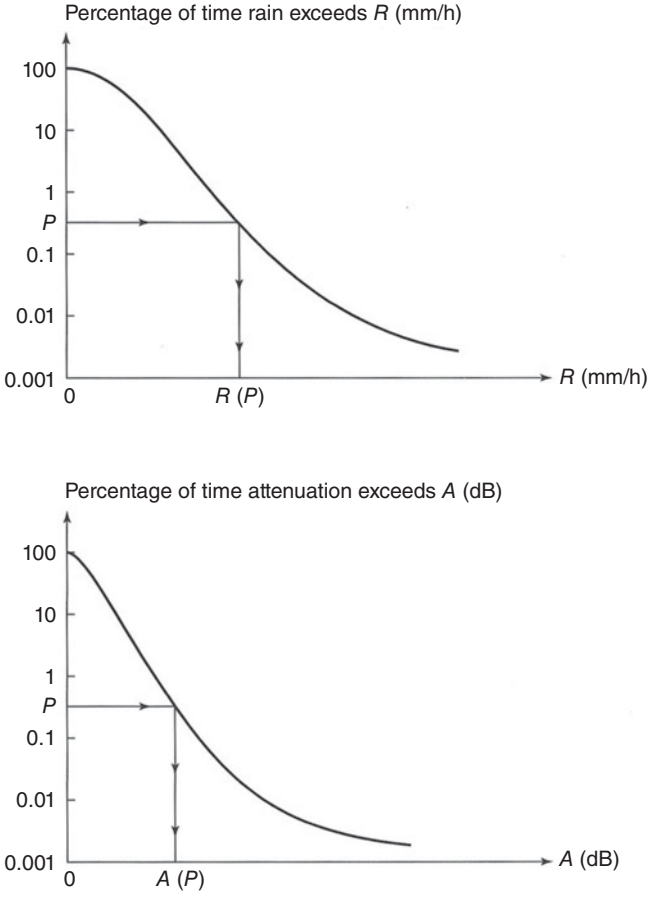


Figure 7.18 Cumulative statistics of rainfall rate and path attenuation illustrating equiprobable procedures. For a given time percentage, *P*, the rainfall rate is read off the rainfall rate statistics and the path attenuation is read off the path attenuation statistics. If the data for the two parameters have been taken over a long enough period (at least a year, longer periods in multiples of years), *R*(*P*) and *A*(*P*) are strongly related. Some models use the full rainfall rate statistics to develop path attenuation statistics. Others use a one-time percentage to relate the two statistics (e.g., the 0.01% point) and develop the second set of statistics from that single point. The disadvantage of this approach (i.e., it is non-physical) is outweighed by the improved accuracy obtained by extrapolating to both low and high time percentages, where the rainfall rate measurements are somewhat suspect.

highly variable drop sizes and rainfall rates, the physical length L used in Eq. (7.13) has usually to be replaced by an effective pathlength L_{eff}. We therefore find that the total path attenuation, A, for a given satellite link is given by Eq. (7.11), which is repeated below for completeness

$$A = \text{specific attenuation} \times \text{effective pathlength in rain} = \gamma_{\text{R}} L_{\text{eff}} \text{ dB} \tag{7.14}$$

The procedure by which the effective pathlength is calculated uses the statistical height of rain (i.e., the melting level height), the height of the earth station above mean sea level, and the elevation angle. (See Figure 7.19 and earlier Figures 7.4 and 7.5.)

In Figure 7.19, the rain is shown as filling the complete slant path up to the melting layer. This is a correct assumption in stratiform rain, which exists over large areas and has a relatively constant rain rate along the slant path. It is rarely correct when convective rain is present. The rain rate and drop distributions are not constant, and the path may not pass through the top of the rain cell. Figure 7.20 illustrates the problem.

The ITU-R procedure for predicting slant path rain attenuation for GEO satellite paths is contained in Section 2.2.1.1 of ITU-R Rec. 618 (P.618-13). It uses a semiempirical approach to the prediction of rain attenuation. Rather than attempt to predict attenuation by inputting rainfall rate at every time percentage, it inputs only the rainfall rate measured (or predicted) for 0.01% of a year. It then extrapolates from this time percentage to other time percentages. While this "one size fits all" approach is non-physical, it removes the inherent inaccuracies of using very low rainfall rates for time percentages of 0.1% (and higher) or very high rainfall rates for time percentages of 0.001%. The current procedure (August 2018) is reproduced below (the equation and figure number have been changed to correspond with those in this chapter). It is worth noting that the general form of this procedure has not changed since 2002 as it provides relatively accurate predictions in most climates. The geometry of calculation procedure is depicted in Figure 7.21 below.

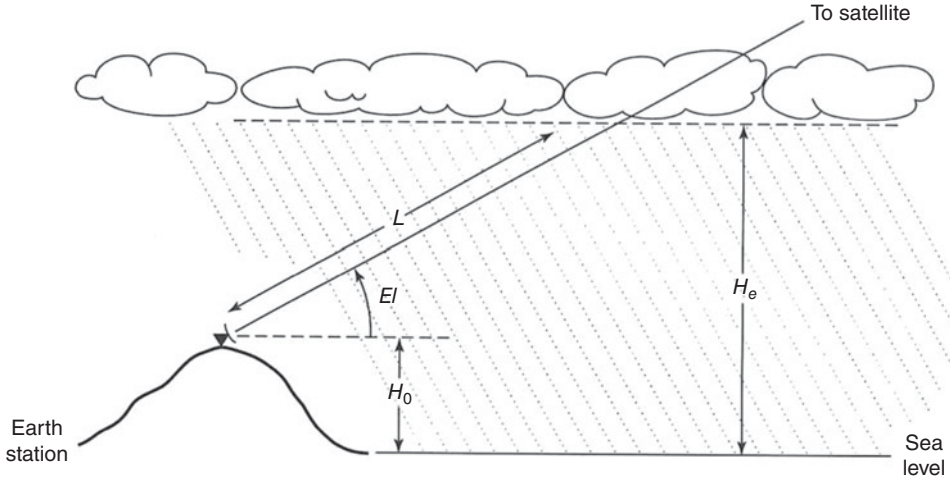


Figure 7.19 Geometry of a satellite path through rain. The height of the melting layer, shown as H_e here, is normally considered to be the highest point at which rain attenuation occurs. The rain fills the volume between the melting level height and the ground. The height of the earth station above mean sea level is given as H_0.

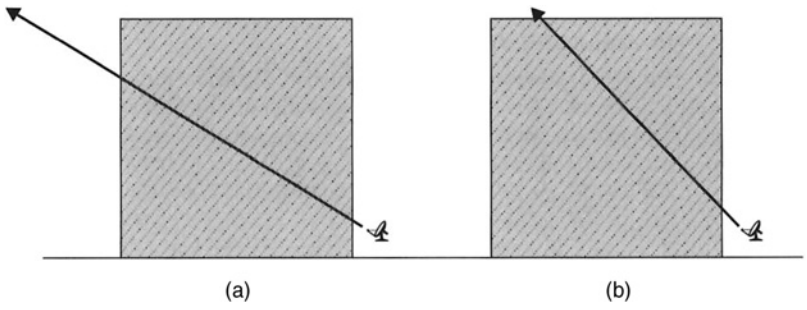


Figure 7.20 Example of different path length geometries. In both cases, a similar rainstorm exists in the slant path. In case **A**, the path to the satellite exits through the side of the storm cell while in case **B** it exits through the top (in a similar geometry as in Figure 7.5). The only difference between the two paths is the elevation angle to the satellite. To develop an effective path length to use in Eq. (7.12), use is made of both a *vertical adjustment factor* and a *horizontal adjustment factor* to account for the possibility of either case **A** or case **B** occurring.

The rain attenuation procedure below is reproduced from ITU-R Rec. P.618-13, with permission. The equation numbers have been changed from the ITU-R procedure to match those in this chapter.

"The following procedure provides estimates of the long-term statistics of the slant-path rain attenuation at a given location for frequencies up to 55 GHz. The following parameters are required:

$R_{0.01}$: point rainfall rate for the location for 0.01% of an average year (mm/h)
h_s: height above mean sea level of the earth station (km)
θ: elevation angle (degrees)

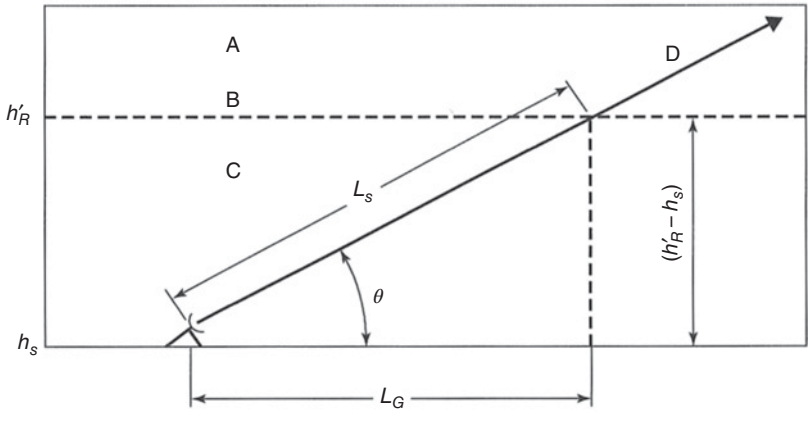


Figure 7.21 Schematic presentation of an earth-space path giving the parameters to be input into the ITU-R rain attenuation prediction procedure. Source: Figure 1 of ITU-R P.618-13, 2018. Reproduced with permission of ITU-R.

φ: latitude of the earth station (degrees)
f: frequency (GHz)
R_e: effective radius of the Earth (8500 km).

Step 1: Calculate the rain height, h_R, as given in Recommendation ITU-R P.839.
Step 2: For $\theta \geq 5^\circ$ compute the slant-path length, L_s, below the rain height from:

$$L_s = \frac{(h_R - h_s)}{\sin\theta} \text{ km} \tag{7.15}$$

For $\theta < 5^\circ$, the following formula is used:

$$L_s = \frac{2(h_R - h_s)}{\left(\sin^2\theta + \frac{2(h_R - h_s)}{R_e}\right)^{1/2} + \sin\theta} \text{ km} \tag{7.16}$$

If $h_R - h_s$ is less than or equal to zero, the predicted rain attenuation for any time percentage is zero and the following steps are not required.
Step 3: Calculate the horizontal projection, L_G, of the slant-path length from:

$$L_G = L_s \cos\theta \text{ km} \tag{7.17}$$

Step 4: Obtain the rainfall rate, $R_{0.01}$, exceeded for 0.01% of an average year (with an integration time of 1 minute). If this long-term statistic cannot be obtained from local data sources, an estimate can be obtained from the maps of rainfall rate given in Recommendation ITU-R P.837. If $R_{0.01}$ is equal to zero, the predicted rain attenuation is zero for any time percentage and the following steps are not required.
Step 5: Obtain the specific attenuation, γ_R, using the frequency-dependent coefficients given in Recommendation ITU-R P.838 and the rainfall rate, $R_{0.01}$, determined from Step 4, by using:

$$\gamma_R = k(R_{0.01})^\alpha \text{ dB/km} \tag{7.18}$$

Step 6: Calculate the horizontal reduction factor, $r_{0.01}$, for 0.01% of the time:

$$r_{0.01} = \frac{1}{1 + 0.78\sqrt{\frac{L_G\gamma_R}{f}} - 0.38(1 - e^{-2L_G})} \tag{7.19}$$

Step 7: Calculate the vertical adjustment factor, $v_{0.01}$, for 0.01% of the time:

$$\xi = \tan^{-1}\left(\frac{h_R - h_s}{L_G r_{0.01}}\right) \text{ degrees} \tag{7.20}$$

$$L_R = \frac{L_G r_{0.01}}{\cos\theta} \text{ km} \tag{7.21}$$

For $\xi \rangle \theta$

$$L_R = \frac{(h_R - h_s)}{\sin\theta} \text{ km} \tag{7.22}$$

else,

if $|\varphi| \langle 36^0, \chi = 36 - |\varphi|$ degrees

else, $\chi = 0$ degrees

$$v_{0.01} = \frac{1}{1 + \sqrt{\sin\theta}\left(31(1 - e^{-(\theta/(1+\chi))})\frac{\sqrt{L_R\gamma_R}}{f^2} - 0.45\right)} \tag{7.23}$$

Step 8: The effective path length is:

$$L_E = L_R v_{0.01} \text{ km} \tag{7.24}$$

Step 9: The predicted attenuation exceeded for 0.01% of an average year is obtained from:

$$A_{0.01} = \gamma_R L_E \text{ dB} \tag{7.25}$$

Step 10: The estimated attenuation to be exceeded for other percentages of an average year, in the range 0.001 to 5%, is determined from the attenuation to be exceeded for 0.01% for an average year:

If $p \geq 1\%$ or $|\varphi| \geq 36°$ $\beta = 0$
If $p < 1\%$ and $|\varphi| < 36°$ and $\theta \geq 25°$ $\beta = -0.005(|\varphi| - 36)$

Otherwise: $\beta = -0.005(|\varphi| - 36) + 1.8 - 4.25\sin\theta$

$$A_p = A_{0.01}\left(\frac{p}{0.01}\right)^{-(0.655+0.033\ln(p)-0.045\ln(A_{0.01})-\beta(1-p)\sin\theta)} \text{ dB.} \tag{7.26}$$

This method provides an estimate of the long-term statistics of attenuation due to rain. When comparing measured statistics with the prediction, allowance should be given for the rather large year-to-year variability in rainfall rate statistics (see Recommendation ITU-R P.678)."

Example 7.3

A Ku-band satellite is to be used in a video broadcasting system. The 17.8 GHz uplink will be from Miami, FL, in the United States, where the studios of the company are located. The elevation angle is 45°. Since the uplink will be used to feed more than a million home receivers, the uplink availability must be 99.99% in the average year.

Question: What is the rain attenuation on the Miami uplink path for 0.01% of the average year?

The information on the link is as follows:

Uplink frequency	17.80 GHz
Polarization	Vertical
Elevation angle	45°
Coefficients for calculating specific attenuation at 17.8 GHz:	$k_v = 0.08$
	$\alpha_v = 1.01$

We are not given the rainfall rate for 0.01% of a year, but Miami is in rain climate M and we can get the rainfall rate from Table 7.2. The height of the rain (h_R) in such a warm climate will be about 4 km. We will assume that the earth station height above sea level (h_S) is no more than 50 m. From Table 7.3 we can interpolate approximate values for k_v of 0.08 and α_v of 1.01.

Answer

Step 1: We already know the rain height (given as 4 km).

Step 2: Find L_S, the slant-path length below the rain height.

$$L_S = \frac{(h_R - h_S)}{\sin\theta}, \text{ thus } L_S = \frac{4.0 - 0.05}{\sin 45^\circ} = \frac{3.95}{0.7071} = 5.5861971 = 5.5862 \text{ km}$$

(Note: keep all the significant figures at present.)

Step 3: Find L_G, the horizontal projection of the slant-path length.

$$L_G = L_S \cos\theta = 5.5862 \times \cos 45^\circ = 5.862 \times 0.7071 = 3.95 \text{ km}$$

Step 4: Find $R_{0.01}$, the rainfall rate for 0.01% of an average year (mm/h)

From the Rain Climatic Zone information (Table 7.2) we have R0.01 = 63 mm/h

Step 5: Find γ_R, the specific attenuation, along the path for Miami for the rainfall rate encountered at 0.01% of an average year.

$$\gamma_R = k(R_{0.01})^{\alpha} = 0.08 \times (63)^{1.01} = 0.08 \times 65.665 = 5.2532 \text{ dB/km}$$

Step 6: Find $r_{0.01}$, the horizontal reduction factor for Miami from Eq. (7.19), thus

$$r_{0.01} = \frac{1}{1 + 0.78\sqrt{\frac{L_{G\gamma R}}{f}} - 0.38(1 - e^{-2L_G})} = 0.6838533$$

Thus $r_{0.01} = 0.6838$ for Miami

Step 7: Calculate the vertical adjustment factor, $v_{0.01}$, for Miami.

To do this we need some intermediate parameters.

Part (a) Calculate ζ, where

$$\xi = \tan^{-1}\left(\frac{h_R - h_s}{L_G r_{0.01}}\right) = \tan^{-1}\left(\frac{4.0 - 0.05}{3.95 \times 0.6838}\right) = \tan^{-1}\left(\frac{3.95}{2.701}\right)$$
$$= \tan^{-1} 1.4624 = 55.6356 = 55.64^{\circ}$$

Thus, $\zeta = 55.64^\circ$ for Miami. This is greater than the elevation angle, θ, which is 45°.

Part (b): Find L_R, an intermediate parameter in calculating the effective path length. Since ζ is $> \theta$,

$$L_R = \frac{L_G r_{0.01}}{\cos\theta} = \frac{3.95 \times 0.683853}{\cos 45^\circ} = \frac{2.7012205}{0.7071} = 3.8201394$$

giving $L_R = 3.82$ km

Part (c): Find χ, the second intermediate parameter for calculating effective path length.

$$\chi = 36 - |\varphi|$$

where φ is the latitude of the site. Thus, $\chi = 36 - 25 = 11.0$ for Miami.

Finally, calculate $v_{0.01}$, from

$$v_{0.01} = \frac{1}{1 + \sqrt{\sin(\theta)}\left(31(1 - e^{-(\theta/(1+\chi))})\dfrac{\sqrt{L_R \gamma_R}}{f^2} - 0.45\right)}$$

$$= \frac{1}{1 + \sqrt{\sin 45^\circ}\left(31(1 - e^{-45/(1+11)})\dfrac{\sqrt{(3.8201394 \times 5.2532)}}{17.8^2} - 0.45\right)}$$

$$= 1.0188537 \text{ for Miami}$$

Step 8: Calculate L_E, the effective path length for Miami

$$L_E = L_R v_{0.01} = 3.8201394 \times 1.0188537 = 3.8921632 = 3.89 \text{ km}$$

Step 9: Calculate $A_{0.01}$, the predicted attenuation exceeded for 0.01% of an average year along the path to Miami.

$$A_{0.01} = \gamma_R L_E = 5.2532 \times 3.8921632 = 20.4463 = 20.4 \text{ dB}$$

The rain attenuation on the uplink from Miami for 0.01% of an average year is therefore predicted to be 20.4 dB, which is the answer to the question posed. This value, however, pertains to a fixed link that does not change significantly with time. Such a situation would not apply to non-geostationary orbit (NGSO) satellite systems. A double-probabilistic approach is required for estimating the statistical impact of rain attenuation on NGSO paths: the probability that attenuation will occur for a given elevation angle and the probability that the satellite will be at that elevation angle. The first approach was documented by the ITU-R and is abstracted below from ITU-R Rec. 618-13, with permission.

7.6.1 Calculation of Long-Term Statistics for NGSO Systems

"For non-GSO systems, where the elevation angle is varying, the link availability for a single satellite can be calculated in the following way:

a. calculate the minimum and maximum elevation angles at which the system will be expected to operate;
b. divide the operational range of angles into small increments (e.g., 5° wide);
c. calculate the percentage of time that the satellite is visible as a function of elevation angle in each increment);
d. for a given propagation impairment level, find the time percentage that the level is exceeded for each elevation angle increment;
e. for each elevation angle increment, multiply the results of c) and d) and divide by 100, giving the time percentage that the impairment level is exceeded at this elevation angle;
f. sum the time percentage values obtained in e) to arrive at the total system time percentage that the impairment level is exceeded.

In the case of multivisibility satellite constellations employing satellite path diversity (i.e., switching to the least impaired path), an approximate calculation can be made assuming that the spacecraft with the highest elevation angle is being used."

7.6.2 Scaling Attenuation with Elevation Angle and Frequency

Experience has shown that, if long-term attenuation data already exist at a site, it is more accurate to scale measured results to another frequency or another elevation angle, instead of predicting the path attenuation at the new frequency and/or elevation angle from rainfall rate data. Two fairly simple (and surprisingly accurate) rules of thumb exist for scaling over small changes in frequency and elevation angle:

i) For a uniform rainfall rate environment (i.e., stratiform rain) and assuming a "flat Earth," path attenuation in decibels scales with the path length through the rain (i.e., it follows a cosecant law);
ii) Between about 10 and 50 GHz, attenuation in decibels scales as the square of the frequency. These two laws are expanded below.

7.6.2.1 Cosecant Law

The attenuation in decibels at the same frequency at elevation angles El_1 and El_2 are approximately related by

$$\frac{A(El_1)}{A(El_2)} = \frac{\text{cosecant}(El_1)}{\text{cosecant}(El_2)} \tag{7.27}$$

This formula breaks down when the elevation angle is low (<10°) where its implicit flat earth and uniform rainfall rate assumptions fail to hold.

Example 7.4

A 12 GHz direct broadcast satellite link was found to experience 4 dB of rain attenuation at an elevation angle of 45° for 0.01% of the time in an average year.

Question: What would be the rain attenuation measured at the same time percentage for the same site if the elevation angle were 10°?

Answer:
Let suffix 1 in Eq. (7.27) refer to the new elevation angle (i.e., 10°) and suffix 2 to the old elevation angle. Thus,

$$\begin{aligned} A(10°) &= [\text{cosecant}(10°)/\text{cosecant}(45°)] \times A(45°) \\ &= [5.7587705/1.4142136] \times 4 = 4.0720657 \times 4 = 16.2882626 = 16.3\,\text{dB} \end{aligned}$$

The impact of elevation angle on a given link is clear from this example.

7.6.2.2 Squared Frequency Scaling Law

If $A(f_1)$ and $A(f_2)$ are the attenuations that would be measured on the same path at frequencies f_1 and f_2 GHz, they are approximately related by

$$\frac{A(f_1)}{A(f_2)} = \frac{(f_1)^2}{(f_2)^2} \tag{7.28}$$

This formula relates the long-term statistics (i.e., the annual statistics). It should not be used for short-term frequency scaling (i.e., from second to second) on a link or for frequencies that are close to any resonant absorption line (e.g., the water vapor absorption line around 22 GHz).

Example 7.5

A user measures rain attenuation statistics along a satellite link as 6 dB for 0.01% of a year when using a carrier frequency of 10.7 GHz. The satellite operator wants to move the user from the current transponder to a new one, which would change the carrier frequency to 11.4 GHz.

Question: What would be the rain attenuation value at 11.4 GHz, all other link parameters remaining the same?

Answer:
Let suffix 1 in Eq. (7.28) refer to the new frequency (i.e., 11.4 GHz) and suffix 2 refer to the old frequency (i.e., 10.7 GHz). Thus,

$$A(11.4) = [(11.4)^2/(10.7)^2] \times A(10.7)$$
$$= [129.9600/114.4900] \times 6 = 6.8107 = 6.8 \text{ dB}$$

A more accurate form of frequency scaling can be found in Section 2.2.1.3.2 of ITU-R rec. 618-13, and is summarized in the section below.

7.6.3 ITU-R Long-Term Frequency Scaling of Rain Attenuation

If A_1 and A_2 are the equiprobable values of rain attenuation, in dB, at frequencies f_1 and f_2 in GHz, respectively, the attenuation at frequency f_2 can be found from that at frequency f_1 from

$$A_2 = A_1(\varphi_2/\varphi_1)^{1-H(\varphi_1,\varphi_2,A_1)} \tag{7.29}$$

where:

$$\varphi(f) = \frac{f^2}{1 + 10^{-4}f^2} \tag{7.30}$$

$$H(\varphi_1, \varphi_2, A_1) = 1.12 \times 10^{-3}(\varphi_2/\varphi_1)^{0.5}(\varphi_1 A_1)^{0.55} \tag{7.31}$$

7.7 Prediction of XPD

Any particle that has spherical symmetry will cause no depolarization of an incident signal. Rain in the atmosphere starts as very small droplets. The surface tension within these droplets is so strong that they retain their spherical shapes. As the drops collide, they coalesce into larger drops. The larger the drop, the more likely it is to distort out of a spherical shape due to wind effects. In convective events, particularly severe thunderstorms, the drops can become relatively large (many millimeters in average diameter) and so they will distort into ellipsoidal forms, generally flattening out in the horizontal axis. Figure 7.22 illustrates the process.

If all of the ellipsoidal drops in a rainstorm were aligned, then waves propagating with their electrical field vectors parallel to the raindrops' minor axes (for all practical purposes, vertically polarized waves) would experience the minimum attenuation for that rainfall rate, and waves propagating with their electric field vectors parallel to the major axes (i.e., horizontally polarized waves) would experience the maximum attenuation. In these two special cases, no depolarization would occur. The difference between the attenuations experienced by waves with horizontal and vertical polarization is

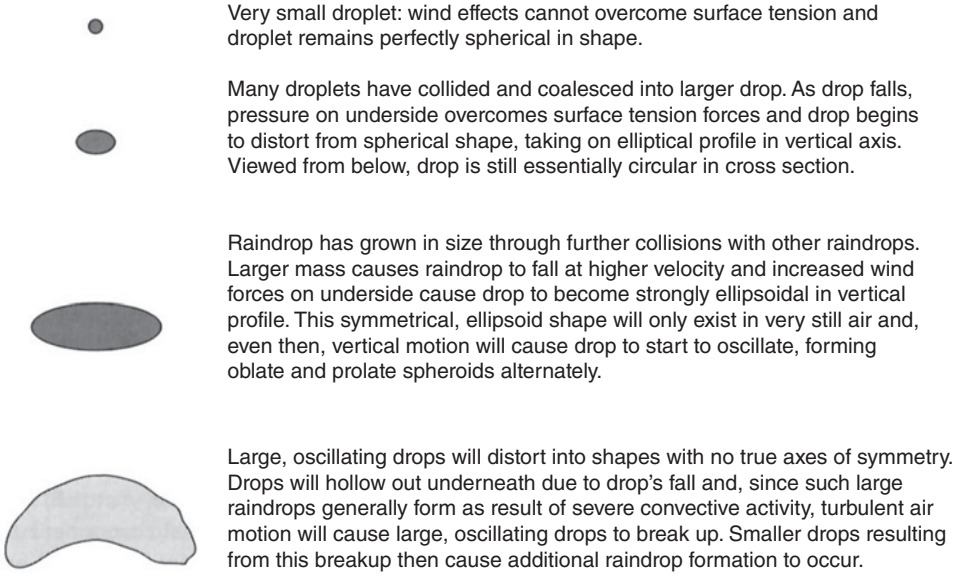


Figure 7.22 Schematic of the shape of an individual raindrop from formation to maturity.

small – rarely greater than a decibel. It is called the differential attenuation. In a like manner, waves with horizontal and vertical polarization can experience differential phase shift as they pass through an anisotropic medium. At frequencies below about 10 GHz, differential phase shift is the more important phenomenon. At frequencies above about 30 GHz, differential attenuation is more important. Between 10 and 30 GHz, either differential phase or differential attenuation will be the major effect, depending on the elevation angle of the link and the climate (Rogers and Allnutt 1986).

Imagine now the case of a wave whose linear polarization is intermediate between horizontal and vertical. We can resolve this wave into its vertically polarized and horizontally polarized components as in Figure 7.23. These components propagate through the rain with their polarizations unchanged, but the horizontal component is attenuated more than the vertical component. If at any point we recombine the vertical and horizontal components to reconstruct the wave, we find that its polarization has rotated toward the vertical and a cross-polarized component is now present. This process is a simplification of a complicated problem in electromagnetic wave scattering. For details of the process, the reader should consult the extensive publications of T. Oguchi, the pioneer researcher in the field (Oguchi 1983).

Depolarization, while it is dependent to a great extent on the volume of rain that is present in the path, the shape of the raindrops in the path and the orientation of their major and minor axes also significantly affect it. The orientation will have two independent features: one that is due to the rain medium, and is referred to as the *canting angle*; and one that is due to the path geometry, and is referred to as the *tilt angle*.

7.7.1 Canting Angle

Falling raindrops orient themselves so as to minimize the aerodynamic forces. In steady fall, the minor axis of the drop is parallel to the net wind force and so their major axis is

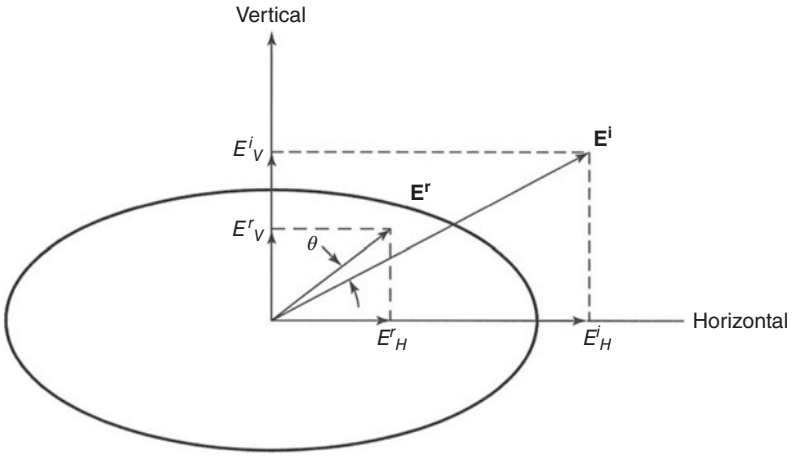


Figure 7.23 A simplified explanation of rain depolarization based on a drop with an elliptical cross section. An incident electromagnetic wave with electric field vector $\mathbf{E}^i$ strikes a raindrop. We resolve it into a horizontal component E^i_H and a vertical component E^i_V. The horizontal component is attenuated more than the vertical component because it encounters more water. Thus, when we combine the horizontal and vertical field components E^i_H and E^i_V that arrive at the receiver, we find that the received wave $\mathbf{E}^r$ has had its polarization rotated toward the vertical by the angle θ.

horizontal when the raindrop is falling in still air. Under windy conditions, the aerodynamic force will have two components: one due to the raindrop fall velocity (i.e., vertical) and one due to the prevailing wind direction (i.e., horizontal). The resultant of these two forces will lead to the raindrop's major axis being canted out of the usual horizontal orientation. The prevailing wind speed lessens with altitude, becoming zero at the ground. The raindrop orientation will therefore vary with altitude. Since the horizontal wind direction with respect to the path varies, the net horizontal component measured over a long interval will be close to zero. The canting angle will therefore have a mean of zero. In any given rainstorm, however, the canting angle will have a finite probability of being non-zero, thus leading to enhanced depolarization for horizontal or vertical polarized waves over short time intervals. Figure 7.24 illustrates the canting angle process schematically.

7.7.2 Tilt Angle

The tilt angle refers to the angle between the local horizontal (or vertical) and the actual orientation of the electric field vector of the transmitted signal. The orientation of the electric field vector transmitted by a geostationary satellite is referenced to the equator at the subsatellite point. Horizontal polarization is parallel to the equator and vertical polarization is perpendicular to the equator. An earth station that lies on the same longitude as the GEO satellite (say, to the north) would receive signals polarized in the local vertical direction if the satellite is transmitting a vertically polarized signal. If the location of the earth station is moved either east or west from the longitude of the GEO satellite, the vertically polarized signals transmitted by the satellite is now received out of the local vertical at the earth station. That is, the polarization vector would appear to be tilted away from their original orientation. The process is illustrated in Figure 7.25.

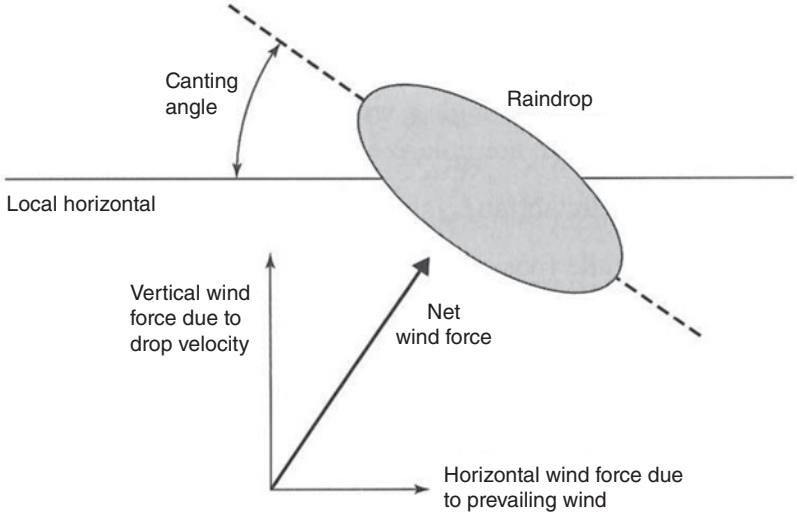


Figure 7.24 Illustration of canting angle. The resultant of the prevailing wind force and the force due to the raindrop fall velocity leads to a net wind force that is out of the vertical direction. The raindrop, which has already distorted into an ellipsoid due to the force induced by the drop velocity, will now orient itself to minimize drag forces. This means that the raindrop will cant out of the horizontal and orient its minor axis to be parallel to the net wind force.

A simple equation that gives the tilt angle τ with respect to the horizontal, assuming the transmissions from a GEO satellite are polarized in the north–south direction, is (Allnutt and Rogers 1986, (b))

$$\tau = \arctan(\tan L_e / \sin \beta) \text{ degrees} \tag{7.32}$$

where L_e is the earth station latitude (positive for the northern hemisphere and negative for the southern hemisphere) and β is the satellite longitude minus the earth station longitude (i.e., $L_s - L_e$), with longitude expressed in degrees east.

Example 7.6

Question: What is the perceived polarization tilt angle at an earth station located at 52°N, 1°E, for vertically polarized signals transmitted from a GEO satellite located at 60°E?

Answer:
Using Eq. (7.32)

$$\tau = \arctan(\tan L_e / \sin \beta) = \arctan(\tan 52 / \sin[60 - 1])$$
$$= \arctan(1.2799/0.8572) = \arctan(1.4932) = 56.19°$$

That is, the polarization vector will be tilted 56.19° away from the originally transmitted linear sense.

The ITU-R XPD prediction method (ITU-R P.618-13, Section 4.1) is based upon the attenuation measured (or predicted) at the frequency of interest, plus additional terms to take account of the canting angle distribution, the tilt angle, and ice crystal

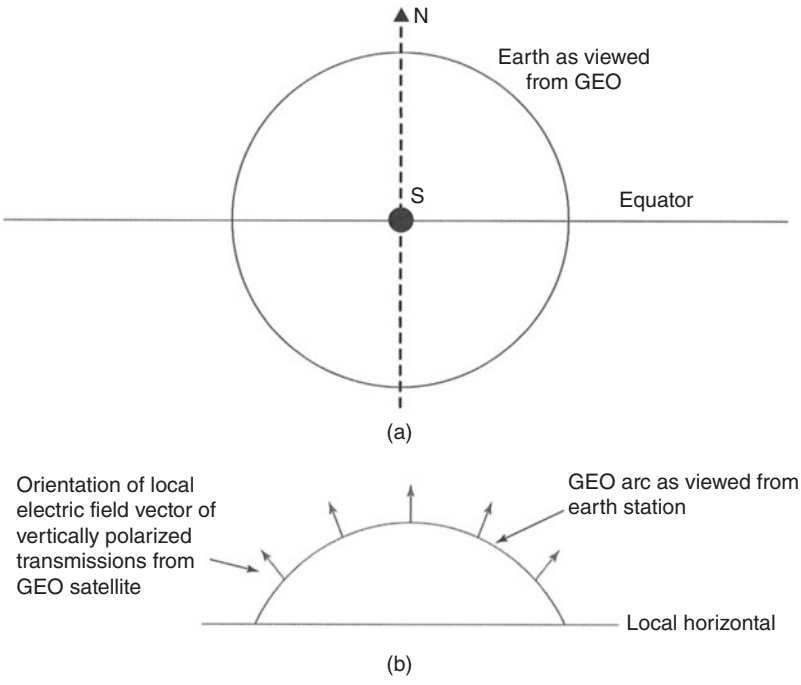


Figure 7.25 Schematic of tilt angle. In (*a*) above, ***S***, is the subsatellite point of a GEO satellite. Transmissions from the satellite will be horizontally polarized if they are parallel to the equatorial plane. Vertically polarized transmissions will be orthogonal to the equatorial plane. If an earth station were on the same satellite longitude (here shown by the broken line ***SN***) it would receive the polarization vector in the orientation transmitted – although the polarization would be undefined at the subsatellite point. In (*b*) above, the earth station is not on the equator. The arc shows how the GEO orbit would look from the earth station. In this instance, the satellite is transmitting a vertically polarized signal. The orientation of the vertically polarized transmission may not be received at the local vertical, however. The local orientation will depend on where the satellite is located on the GEO arc as seen by the earth station. The polarization vector may therefore be *tilted* out of the transmitted orientation by virtue of the link geometry. The polarization will only be vertical (or horizontal) at the earth station site to a GEO satellite if the azimuth to the satellite is 0° or 180° from true north.

depolarization (Bostian and Allnutt 1979). The step-by-step procedure of the ITU-R, reproduced with permission, is summarized below. The equation numbers follow the sequence of this chapter.

"To calculate the long-term statistics of depolarization from rain attenuation statistics, the following parameters are needed:

A_p: rain attenuation (dB) exceeded for the required percentage of time, p, for the path in question, commonly called co-polar attenuation (CPA)
T: tilt angle of the LP electric field vector with respect to the horizontal (for circular polarization use $\tau = 45°$)
f: frequency (GHz)
θ: path elevation angle (degrees)

The method described below to calculate XPD statistics from rain attenuation statistics for the same path is valid for $6 \leq f \leq 55$ GHz and $\theta \leq 60°$. The procedure for scaling

to frequencies down to 4 GHz is given in Section 4.3 of ITU-R P.618-13, and follows the section below.

Step 1: Calculate the frequency-dependent term:

$$C_f = 60\log f - 28.3 \text{ for } 6 \le f < 9 \text{ GHz} \tag{7.33a}$$

$$C_f = 26\log f + 4.1 \text{ for } 9 \le f < 36 \text{ GHz} \tag{7.33b}$$

$$C_f = 35.9\log f - 11.3 \text{ for } 36 \le f \le 55 \text{ GHz} \tag{7.33c}$$

Where f is the frequency in GHz

Step 2: Calculate the rain attenuation dependent term:

$$C_A = V(f)\log A_p \tag{7.34}$$

where $V(f)$ is given by

$$V(f) = 30.8f^{-0.21} \text{ for } 6 \le f < 9 \text{ GHz} \tag{7.35a}$$

$$V(f) = 12.8f^{0.19} \text{ for } 9 \le f < 20 \text{ GHz} \tag{7.35b}$$

$$V(f) = 22.6 \text{ for } 20 \le f < 40 \text{ GHz} \tag{7.35c}$$

$$V(f) = 13.0f^{0.15} \text{ for } 40 \le f \le 55 \text{ GHz} \tag{7.35d}$$

Step 3: Calculate the polarization improvement factor, C_τ:

$$C_\tau = -10\log\left[1 - 0.484\left(1 + \cos 4\tau\right)\right] \tag{7.36}$$

The improvement factor $C_\tau = 0$ for $\tau = 45°$ and reaches a maximum value of 15 dB for $\tau = 0°$ or 90°. The value $\tau = 45°$ corresponds to circular polarization.

Step 4: Calculate the elevation angle term, C_θ:

$$C_\theta = -40\log(\cos\theta) \text{ for } \theta \le 60^o \tag{7.37}$$

Step 5: Calculate the canting angle dependent term, C_σ:

$$C_\sigma = 0.0053\sigma^2 \tag{7.38}$$

σ is the effective standard deviation of the raindrop canting angle distribution, expressed in degrees; σ takes the value 0°, 5°, 10°, and 15° for 1%, 0.1%, 0.01%, and 0.001% of the time respectively.

Step 6: Calculate the rain XPD not exceeded for p% of the time, XPD_{rain}:

$$XPD_{\text{rain}} = C_f - C_A + C_\tau + C_\theta + C_\tau \text{ dB} \tag{7.39}$$

Step 7: Calculate the ice crystal dependent term, C_{ice}:

$$C_{ice} = XPD_{\text{rain}} \times (0.3 + 0.1\log p)/2 \text{ dB} \tag{7.40}$$

Step 8: Calculate the total XPD not exceeded for p% of the time, including the effects of ice crystals, XPD_p:

$$XPD_p = XPD_{\text{rain}} - C_{ice} \text{ dB} \tag{7.41}$$

The rain attenuation below 8 GHz is fairly low and so the attenuation-dependent XPD prediction method does not provide accurate results. A formula that allows XPD to

be scaled in frequency from a known XPD value (XPD_1) at frequency f_1 to a different frequency f_2 along the same path is given below (ITU-R P.618-13).

$$XPD_2 = XPD_1 - 20\log\left[\frac{f_2\sqrt{1-0.484(1+\cos 4\tau_2)}}{f_1\sqrt{1-0.484(1+\cos 4\tau_1)}}\right] \text{ for } 4 \le f_1, f_2 \le 30\,\text{GHz} \tag{7.42}$$

Unpublished results from the ITALSAT experiment (Barbaliscia et al. 1999) appear to show that it is possible to predict XPD between 35 and 50 GHz by amending the equations in Step 1 and Step 2 above such that

$$C_f = 26\log f \tag{7.43}$$

$$V(f) = 20 \tag{7.44}$$

Example 7.7

Question: What is the value of XPD at 0.01% of the time for a 12 GHz link operating at an elevation angle of 30° that experiences 7 dB attenuation for this period of time? Calculate the XPD for tilt angles of 20° and 0°.

Answer:
Using the step-by-step procedure we have:

Step1: $C_f = 26 \log f + 4.1 = 32.1587$
Step 2: $V(f) = 12.8\, f^{0.19} = 12.8 \times 1.6034 = 20.5236$

$$C_A = V(f)\log A_p = 20.5236 \times \log 7 = 17.3445$$

Step 3: tilt angle of 20°

$$C_\tau = -10\log[1 - 0.484(1 + \cos 4\tau)] \text{ giving}$$

$$C_\tau = -10\log[1 - 0.484(1 + \cos 80)] = 3.6456$$

Tilt angle of 0°

$$C_\tau = -10\log[1 - 0.484(1 + \cos 4\tau)] \text{ giving}$$

$$C_\tau = -10\log[1 - 0.484(1 + \cos 0)] = 14.9485$$

Step 4: $C_\theta = -40 \log(\cos\theta) = -40 \log(\cos 30°) = -40 \log(0.8660) = 2.4988$
Step 5: $C_\sigma = 0.0053\, \sigma^2 = 0.0053\, 10^2 = 0.53$
Step 6: $XPD_{\text{rain}} = C_f - C_A + C_\tau + C_\theta + C_\sigma$

For $\tau = 20°$

$$\text{XPD} = 32.1587 - 17.3445 + 3.6456 + 2.4988 + 0.53 = 21.4886 = 21.5\,\text{dB}$$

For $\tau = 0°$

$$\text{XPD} = 32.1587 - 17.3445 + 14.9485 + 2.4988 + 0.53 = 32.7915 = 32.8\,\text{dB}$$

Step 7: $C_{ice} = XPD_{rain} \times (0.3 + 0.1 \log p)/2$

For $\tau = 20°$

$C_{ice} = 21.4886 \times (0.3 + 0.1 \log p)/2 = 21.4886 \times (0.3 + 0.1 \log 0.01)/2$

$= 21.4886 \times (0.3 - 0.2)/2 = 1.0744$

For $\tau = 0°$

$C_{ice} = 32.7915 \times (0.3 + 0.1 \log p)/2 = 32.7915 \times (0.3 + 0.1 \log 0.01)/2$

$= 32.7915 \times (0.3 - 0.2)/2 = 1.6396$

Step 8: $XPD_p = XPD_{rain} - C_{ice}$

For $\tau = 20°$

XPD = 21.4886 − 1.0744 = 20.4142 = 20.4 dB

For $\tau = 0°$

XPD = 32.7915 − 1.6396 = 31.1519 = 31.2 dB

NOTE: The **single, best way** to reduce depolarization is to operate with polarization senses that are linear vertical or horizontal as perceived by the receiving antenna. This can be seen from the very different results calculated in the above example when the tilt angle was 0° (i.e., the signal is being received in linear, horizontal polarization) to those when the tilt angle is 20°.

7.7.3 Ice Crystal Depolarization

The calculation procedure for ice crystal depolarization incorporated in the calculation of XPD has been found to have wide variations in accuracy. At high elevation angles and at frequencies below 10 GHz, the procedure tends to agree with measured data. That is, ice crystal depolarization occurs only in severe thunderstorms and so it is a rare occurrence. However, on low elevation angle paths, the contribution due to ice crystals has been observed to occur for quite high time percentages. At frequencies above 30 GHz, it is expected that ice crystal depolarization will be a significant effect, particularly at elevation angles below 30°.

7.7.4 Rain Effects on Antenna Noise

At frequencies below about 50 GHz, rain attenuation is mostly caused by absorption rather than by scattering of the signal energy out of the path. Any absorber with a physical temperature greater than absolute zero (0 K) will act as a black body radiator. At frequencies below 300 GHz, the radiation is in the form of white Gaussian noise with a noise power given by kTB, where T is the equivalent noise temperature of the absorber. Raindrops are absorbers at microwave frequencies and, when the raindrops fall through the antenna beam, some of their isotropically radiated thermal energy will be detected by the receiver (see Chapter 4). Rain will therefore cause not only signal attenuation and depolarization; it will also cause an increase in sky temperature, which, in turn, will increase the overall system noise temperature. The impact of the increase in sky noise temperature can be high for low noise receiving systems at Ku-band, as illustrated in the examples in Chapter 4. Rain attenuation in the 1–3 dB range can cause the system noise level to increase by 1–3 dB, leading to a reduction in CNR ratio (in dB) in rain, which is twice the rain attenuation value.

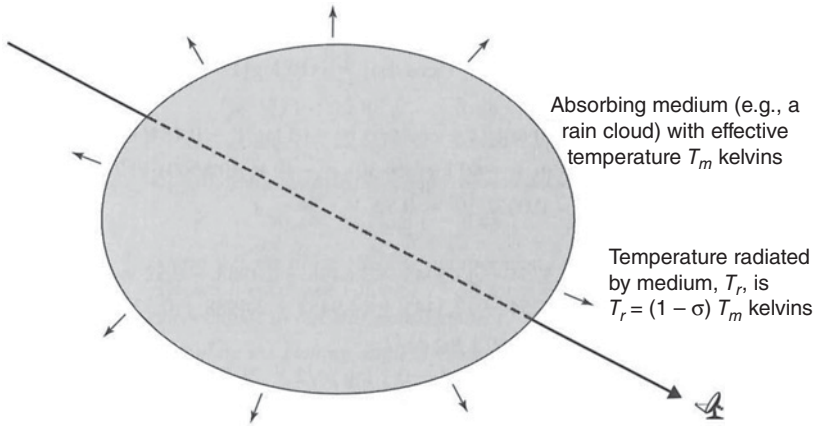


Figure 7.26 Schematic of the additional radiated sky temperature due to absorption in rain. The added temperature received by the antenna due to the radiation from the "hot" rainstorm will cause an additional component to be added to the system noise temperature. This additional component is similar to the noise temperature contribution from a lossy feed. In Chapter 4, in the analysis of system noise temperature, a noise temperature contribution due to signal loss, T_l, was calculated using a "gain" component G_l. where G_l was a linear value. For example, when the component at a physical temperature of 280 K caused a loss of 2 dB (which = 1/1.58 = 0.63 of the original value), $T_l = T_p(1 - G_l) = 280(1 - 0.63) = 103.6K$. The parameter G_l is identical to σ, that is, a loss of 2 dB is the same as a fractional transmission of 0.63 of the original signal.

The *increase* in antenna noise temperature due to rain, T_b, may be estimated by

$$T_b = 280(1 - e^{-A/4.34})\ \text{K} \tag{7.45}$$

where A is the rain attenuation in decibels and the value 280 K is an *effective temperature* of the rain medium in kelvins. Values between 273 and 290 K may be used, depending on whether the climate is cold or tropical.

An alternative approach is to treat the rain as a passive attenuator with a fractional transmission coefficient of σ. If the rain totally attenuates the signal, $\sigma = 0$ (that is to say no signal energy passes through the passive attenuator); if the rain medium is completely transparent and no attenuation takes place, $\sigma = 1$ (that is to say, all the incident energy passes through without any loss). Figure 7.26 illustrates the process.

Example 7.8

Question: What is the additional noise temperature contribution of an antenna compared with that in clear sky when there is 4 dB of rain attenuation in the path? You may assume that the rain medium is at a temperature equivalent to 285 K.

Answer

An attenuation of 4 dB causes the signal to be reduced by a factor of 2.5119. The fractional transmission coefficient, σ, would therefore be 1 / 2.5119 = 0.3981. (Another way of looking at this is to say that only 39.81% of the original signal power is being received during the 4 dB rain event). The additional sky temperature radiated would therefore be 285 (1–0.3981) which = 171.5395 K = 171.5 K. Note that, if the system noise temperature had been 200 K, the effective system noise temperature is now 200 + 171.5 = 371.5 K. In other words, the signal power has decreased by 4 dB and the noise power has increase by 2.7 dB. A 4 dB rain attenuation has thus led to a 6.7 dB reduction in *CNR*. This is

somewhat simplistic, since the receiving antenna efficiency is not 100%, and it therefore does not accept all of the radiation that is incident upon it. However, the enhanced sky noise contribution received by the antenna during rain conditions will be close to that radiated by the rainstorm. Careful attention must be paid in the system design to allow for enhanced sky noise contributions as well as signal degradations when developing link budgets. Put another way, the key in link budget calculations is to find the change in carrier-to-noise, *CNR*, rather than just the change in carrier power, *C*.

7.8 Propagation Impairment Countermeasures

7.8.1 Attenuation

Many research groups have investigated the use of fade countermeasures. Fade mitigation has been shown to fall into three main classes (Castanet et al. COST 255, 1998).

- Power control (i.e., varying the EIRP of the signal to enhance CNR)
- Signal processing (i.e., changing the parameters of a signal to improve BER)
- Diversity (i.e., choosing a different path or time to take advantage of decorrelated fading)

Interestingly, the three main classes of fade mitigation affect a link differently and are complementary in nature. For satellite systems that use frequencies at Ka-band and above, all three classes of fade mitigation techniques might be required for high availability links, particularly those that are *packet oriented*. See, for example, (COST 272, 2018) and the follow-on ESA COST Action (COST 280, 2011). We will look briefly at each technique.

7.8.1.1 Power Control

In adaptive power control, the transmitter power is adjusted to compensate for changes in signal attenuation along the path. At its simplest, it is like automatic gain control in a receiver, which adjusts to fluctuations in the incoming energy so as to hold the receiver output constant. Unless *reverse band operation* is being used, most satellite earth stations transmit in the higher band to the satellite, receiving in the lower band. An example at C-band is to operate with 6 GHz on the uplink and 4 GHz on the downlink. In this way, many satellite links are operated such that the uplink is the critical portion of the connection; that is, the first part of the overall link that will drop out in a rain fade is the uplink. The overall availability (and performance) of the connection is therefore enhanced if the uplink can operate with an increased *EIRP* in rain. This is referred to as *Uplink Power Control* (*ULPC*).

ULPC can operate *closed loop*, where the signal power is detected at the satellite and a control signal sent back to the earth station to adjust the power, or *open-loop*, where the fade on the downlink signal is used to predict the likely fade level occurring on the uplink. Closed-loop operation is always more accurate but is more expensive to implement and has an inherent delay due to the time the control signal takes to be received at the earth station. Hence most *ULPC* systems are, at present, open-loop.

Open-loop *ULPC* becomes more difficult the further apart the downlink and uplink frequencies are. It becomes even more difficult at Ka-band when the downlink (~20 GHz) and uplink (~30 GHz) frequencies are on either side of the 22 GHz water vapor absorption line. The ratio of 30 GHz attenuation to 20 GHz attenuation is less than

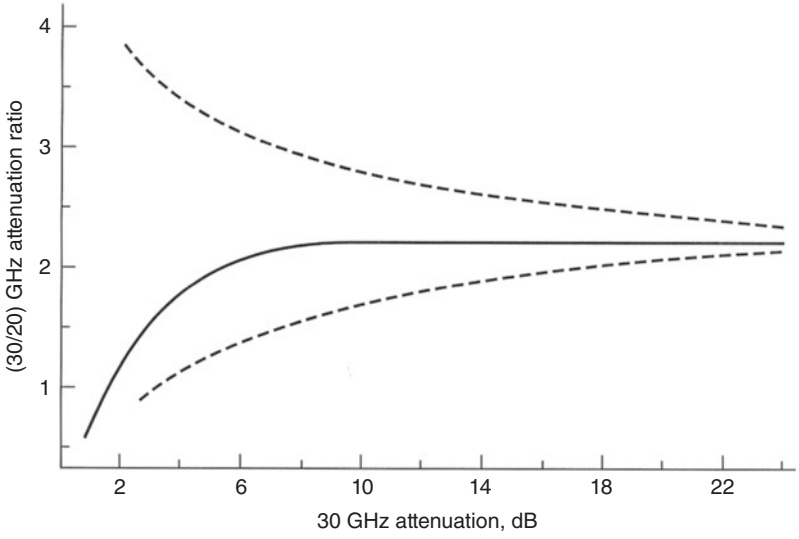


Figure 7.27 Instantaneous 30 : 20 GHz attenuation scaling ratio with 20 GHz attenuation as parameter. The solid curve above is a prediction of the scaling ratio that takes into account both rain attenuation and tropospheric scintillation. The pair of broken curves are the bounds of individual instantaneous measurements of the uplink and downlink attenuation values. The large range of scaling ratios shows that great care must be taken in developing open-loop ULPC algorithms that use only a measure of the amplitude of the downlink signal.

1 for 20 GHz attenuation values of less than 1.0 dB, since cloud attenuation (i.e., essentially water vapor absorption) is higher at 20 GHz than at 30 GHz due to the proximity of the 22 GHz water absorption line to the 20 GHz downlink. Figure 7.27 gives the average 30 : 20 GHz attenuation ratios, with downlink attenuation as parameter. Note that the long-term 30 : 20 GHz attenuation scaling ratio does not become established until the downlink attenuation is above 7 dB. Another major consideration is power flux density variations at the satellite. If many earth stations are operating under rain fade conditions with the same satellite, as could happen in a very small aperture terminal (VSAT) network with many hundreds or thousands of earth stations, implementing ULPC can lead to significant received power fluctuations at the satellite, and this has capacity implications. Some of the advanced Ka-band satellites with multiple switched beams can also implement downlink power control, if sufficient bandwidth and power are available.

7.8.1.2 Signal Processing

The move from very large earth stations (e.g., the INTELSAT Standard A) to a multiplicity of small earth stations has been accompanied by a shift in the median traffic stream. It is rare to find a non-video or non-internet network distribution link via a satellite at a rate of more than 2 Mbps. The need to make small traffic streams economic by using VSATs has led to the introduction of *onboard processing* (*OBP*) techniques. This process typically translates the digital carriers arriving at the satellite to baseband for processing and onward transmission back to earth. The process is generically called *MCDDD – multi-carrier demodulation, demultiplexing, and decoding*. The *OBP* process is carried out at baseband and allows each individual traffic packet to be switched to the correct output port of the satellite antenna for transmission down to earth following recoding and remultiplexing. By detecting the signal level of each packet on arrival at the *OBP*,

not only can most bit errors be removed but the transmitting earth station can also be alerted if the energy level of the received packet has fallen, so that *ULPC* can be used at the earth station to correct the signal level (within the power level range of the *ULPC* system). The use of *OBP* separates the uplink from the downlink and each part of the link can be treated separately in developing a link budget.

7.8.1.3 Diversity

Many diversity schemes have been proposed, but few have been implemented with large earth stations as yet due to the cost. Not only do you need to have two, very expensive earth stations, but a high capacity link is required to connect the two. This may change with *NGSO* systems operating with a large number of satellites (see Chapter 9) and advanced phased array antennas on the ground that can switch rapidly from a satellite with a bad link to one with a good link. If *OBP* techniques are being used on a satellite, a form of time diversity can be used. In this approach, additional slots in the time division multiple access (*TDMA*) *frame* can be assigned to the rain-affected link so that the same signal can be sent at a slower rate, essentially lowering the bandwidth and raising the *CNR*. The forward error correction (*FEC) rate* could also be changed in the *OBP payload*. If the satellite operates in a number of frequency bands (e.g., C-band and Ku-band), a rain affected Ku-band link could be switched to C-band, which is not attenuated significantly by rain. To be able to do this, spare C-band capacity must be held in reserve on the satellite so that it can be used when required. Similarly, each Ku-band earth station would need to have a dual-band antenna and receiving system so that they could switch between the two bands. The added cost has not justified this approach to date. However, the *NGSO V-band systems* in design at present may find it economic to include a low capacity Ka-band or Ku-band payload to use in those traffic streams that are the highest priority. Of all the diversity schemes, that of *site diversity* appears to offer the most significant gain in availability.

Site diversity is a technique whereby two, or more, earth stations are located sufficiently far apart to ensure that the rain impairments observed at each of the stations are generally uncorrelated. More exactly, it is the *paths* through the rain that are uncorrelated and so the technique is more accurately described as path diversity. The earth stations are connected together so that any one earth station can be used to support the traffic stream while the other(s) is (are) suffering a rain fade. This technique is operational with the earth stations used to support the 30/20 GHz earth-space links for the Iridium constellation. Since Iridium is an NGSO constellation, there are four earth stations used, two working with one satellite in site diversity mode, while the other pair wait for the next satellite to move above the horizon.

If we assume that there are two earth stations, identified by suffixes 1 and 2, which are operated in a site diversity mode, then the *joint attenuation* $A_J(t)$ is defined by

$$A_J(t) = \text{minimum } [A_1(t), A_2(t)] \text{ dB} \tag{7.46}$$

The average single-site attenuation $A_S(t)$ is the mean of $A_1(t)$ and $A_2(t)$, namely

$$A_S(t) = [A_1(t) + A_2(t)]/2 \text{ dB}. \tag{7.47}$$

An ideal system that monitors the received downlink signals at both sites and always selects the stronger of the two experiences an attenuation of $A_J(t)$ and the diversity system would perform better than either site alone. How much better is measured by two statistical quantities, *diversity gain* and *diversity improvement*. Diversity gain is a measure of the difference in attenuation experienced between that at a single site and that

measured when the two earth stations are working in diversity mode. More precisely, diversity gain, $G_D(P)$, is the decibel difference between the average single-site attenuation $A_S(P)$ equaled or exceeded P% of the time and the joint attenuation $A_J(P)$ equaled or exceeded P% of the time.

$$G_D(P) = A_S(P) - A_J(P) \tag{7.48}$$

Diversity improvement $I_D(A)$ is the ratio between the percentage of time P_S that the average single-site attenuation A_S exceeds A dB to the percentage of time P_J that the joint attenuation A_J exceeds A dB.

$$I_D(A) = \frac{P_S(A)}{P_J(A)} \tag{7.49}$$

Diversity gain determines system margin, and it is the measure of diversity system performance that we will use here. In addition, diversity gain has been shown to be stable from year to year and, as such, is a reliable statistic to use in system design. Diversity improvement, on the other hand, is extremely variable from year to year. Section 2.2.4.1 of ITU-R P.618-13 details the procedure to use to calculate the joint outage probability. Figure 7.28 illustrates these two concepts.

The first, and still the best, diversity gain model is that due to Hodge (Hodge (a) 1976; (b) 1982), which has been adapted by the ITU-R [section 2.2.4.2 of P.618-13]. Hodge

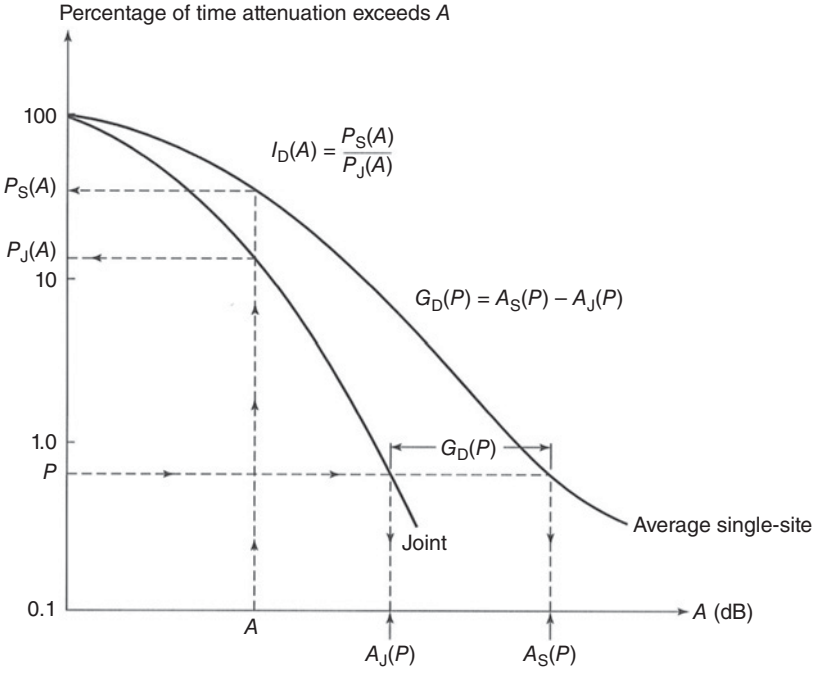


Figure 7.28 Illustration of diversity gain and diversity improvement (diversity advantage). At a given percentage of time, ***P***, the diversity gain $\boldsymbol{G_D(P)}$ is the decibel difference between the average single-site attenuation exceeded $\boldsymbol{A_S(P)}$ and the joint attenuation exceeded $\boldsymbol{A_J(P)}$. At a given attenuation, ***A***, the diversity improvement $\boldsymbol{I_D(A)}$ is the ratio of the percentage of the time $\boldsymbol{P_S(A)}$ that the single site attenuation exceeds ***A*** to the percentage time $\boldsymbol{P_J(A)}$ that the joint attenuation exceeds ***A***.

developed the diversity gain model through an iterative analysis of diversity data available. Intuitively, he assumed site separation was the key element. In this, he has been proved correct. Once two sites are separated by 20 km, or more, it is rare for heavy rain to fall simultaneously at both sites. The Hodge model, and the ITU-R diversity gain model, is therefore applicable to site separations of 20 km or less. The ITU-R procedure is abstracted below, with permission.

d: separation (km) between the two sites
A: path rain attenuation (dB) for a single site
f: frequency (GHz)
θ: path elevation angle (degrees)
ψ: angle (degrees) made by the azimuth of the propagation path with respect to the baseline between the two sites, chosen such that $\psi \leq 90°$

Step 1: Calculate the gain contributed by the spatial separation of the two earth stations from:

$$G_d = a(1 - e^{-bd}) \tag{7.50}$$

where:

$$a = 0.78A - 1.94(1 - e^{-0.11A})$$

$$b = 0.59(1 - e^{-0.1A})$$

Step 2: Calculate the frequency-dependent gain from:

$$G_f = e^{-0.025f} \tag{7.51}$$

Step 3: Calculate the gain term dependent on elevation angle from:

$$G_\theta = 1 + 0.006\theta \tag{7.52}$$

Step 4: Calculate the baseline-dependent term from the expression:

$$G_\psi = 1 + 0.002\psi \tag{7.53}$$

Step 5: Compute the net diversity gain, G, as the product:

$$G = G_d \times G_f \times G_\theta \times G_\psi \tag{7.54}$$

The use of a site diversity system is very expensive if traditional approaches are used. That is, two large earth stations connected together via a very high-speed terrestrial link. It has only been used operationally to date by the gateway stations of the Iridium network. These gateway earth stations operate in Ka-band and are single-point failures for the network. As such, the expense of a diversity setup was well justified. Another proposed approach to site diversity has been to use wide area diversity (Allnutt and Arbesser-Rastburg 1985), in which a multitude of VSATs are linked via routers to a metropolitan area network.

7.8.2 Depolarization

Depolarization compensation is a technique whereby the feed system of the antenna is adjusted in such a way as to correct for depolarization in the path. Alternatively, the orthogonal channels may be cross-coupled in the receiver and, provided good samples of

the signal in each channel can be obtained, the interfering (i.e., depolarized) signal may be removed by subtracting the correct amount of signal. Few earth stations have implemented depolarization compensation, as it is an expensive undertaking. Those earth stations that have implemented depolarization compensation have done so at C-band.

At C-band, differential phase is the primary cause of depolarization. As the frequency increases, differential attenuation becomes an increasingly significant cause of depolarization until, at V-band, differential attenuation dominates completely. For this reason, the amount of rain depolarization observed for each dB of rain attenuation on commercial communications satellite links is largest at C-band, reducing monotonically to V-band. Most Ka-band systems for direct-to-home (DTH) internet services have rain margins of less than 10 dB. At this attenuation level, depolarization effects are not significant.

7.9 Summary

The design of radio systems includes a link margin that is intended to provide for changes in the received signal level due to both equipment effects and random changes in the environment between the transmitter and the receiver. The link margin permits the communications system to operate with both the required performance, a measure of the service quality required for a significant fraction of the time, and availability, a measure of the time period when usable service is provided. Developing an adequate link margin is critical to the acceptance of the service. However, each additional dB of link margin that is provided comes with a cost associated with it. A lot of care, therefore, goes into developing an accurate estimate of the likely impairments on any given link that would cause the performance and availability of the service to fall below acceptable levels. A key to this estimate is an understanding of the propagation effects along the path between the satellite and the earth station.

Propagation effects cause two principal phenomena to be observed at the receiving terminal: a change in the wanted signal level, which is referred to as signal attenuation or fading (although care must be taken with using the term *fading* as terrestrial microwave engineers use this term to describe multipath fading); and a change in the unwanted signal level, which is referred to as depolarization or cross-polarization. Attenuation and depolarization effects are a function of the signal frequency, the atmospheric conditions, and the path geometry. In general, the higher the frequency, the warmer and wetter the weather, and the lower the operating elevation angle of the earth station, the worse the propagation effects are. The only time this is not true is for ionospheric effects, where the effects on commercial satellite systems are only of significance at C-band or below.

With the exception of ionospheric effects, propagation phenomena are dependent on the weather that occurs in the lower atmosphere, generally up to about 10 miles, 16 km. Weather is a *cyclostationary phenomenon*, that is the major seasons (summer, fall, winter, and spring) tend to be in the same periods of a year, as do monsoon and other large weather patterns such as cyclones, typhoons, and hurricanes. However, within this relatively predictable seasonal pattern, there are large fluctuations in the type and severity of weather systems on a day-to-day basis, and even hour by hour. To overcome the apparent randomness in the weather phenomena, particularly rain and *rain rate*, statistical models are used. Long-term measurements of rainfall rate are statistically related to

long-term path attenuation measurements when taken over the same period and at the same site. In this case, long term is at least one year so that all of the seasons normally experienced in a given year may be included.

The prediction of rain attenuation has taken two distinct paths: one uses measured data and develops an empirical model to predict the phenomenon on a worldwide basis; the other attempts to model the physics of the process. Statistical models of rain depolarization, tropospheric scintillation, gaseous losses, cloud attenuation, low angle fading, and related propagation effects have been developed. Most of these models provide usable predictions for frequencies between 4 and 55 GHz, but care must be taken when predictions for unusual path geometries (e.g., <5°) or severe climates (e.g., tropical regions) are required.

More recently, the impact of individual rain fades – their occurrence statistics, duration of individual events, time between fades of the same level – has become important for developing user perception for *direct to home* (DTH) services. Countermeasures to rain fades may take many forms – for example, increasing the TDMA frame allocation, changing the modulation index, changing the power level, changing the frequency – and it is likely that some of them will be included in the Ka-band and V-band services planned for the second decade and beyond of the twenty-first century. Paradoxically, individual rain attenuation events that cause a link to drop out are, on average, between three and five minutes, depending on the climate. For a livestreaming video service, this length of time for no service is unacceptable, but for internet users, this is well within the expected tolerance for receiving a reply from a called party.

Exercises

7.1 Which of the 48 states that form the Continental United States (Conus) has the highest occurrence of very heavy rain? What is the highest rainfall rate exceeded for 0.01% of an average year in that state?

7.2 Why is attenuation on a slant path in clear sky conditions greater at 22 GHz than at 30 GHz?
Why is the occurrence of attenuation in rain on a slant path at the 1–2 dB level often greater at 22 GHz than at 30 GHz?

7.3 A satellite system operator is looking for locations in North America to locate V-band earth stations to link to a constellation of low earth orbit satellites serving latitudes between 50°N and 50°S. Rain attenuation is a major factor at V-band, so the operator wants to find locations with very low occurrence of heavy rain.
a. Why might the system operator reject rain climate zone A? Give two reasons.
b. Why might the system operator decide that rain climate zone B is a better choice? Give two reasons.

7.4 An earth station at sea level communicates at an elevation angle of 30° with a GEO satellite in Ka-band at a downlink frequency of 20 GHz. Stratiform rain is present in the slant path with a melting level height of 2.5 km. Find
a. The path length through the rain.
b. The path attenuation for a specific attenuation of 1.5 dB/km.

c. Use the square law frequency scaling procedure to estimate the attenuation on the downlink at 42 GHz with an elevation angle of 30°.
d. Estimate the downlink attenuation for the 20 GHz link when the elevation angle is increased to 45°.

7.5 A satellite downlink has a frequency of 10.0 GHz.
a. What is the specific attenuation for a rainfall rate of 30 mm/h with linear vertical polarization?
b. What is the specific attenuation for a rainfall rate of 30 mm/h with linear horizontal polarization?
c. What is the specific attenuation for a rainfall rate of 30 mm/h and circular polarization?

7.6 An 18 GHz uplink to a GEO satellite is established from an earth station in northern Virginia. The elevation angle to the satellite is 35° and vertical polarization is used on this link.
The link is required to have an availability of 99.99% of an average year. Rainfall exceeds a rate of 40 mm/h for 0.01% of the year at the earth station location, and the effective rain height is 3.9 km. The estimated values of parameters k_v and α_v at 18.0 GHz are $k_v = 0.78$ and $\alpha_v = 1.02$.
Using the procedure set out in Section 7.6 of the text, assuming that the earth station is at sea level, determine:
a. The effective slant path length through rain.
b. The specific attenuation for rain on this path, γ_R, for rain that occurs for 0.01% of an average year.
c. The total path attenuation for rain that occurs for 0.01% of an average year.

7.7 Repeat the analysis of question #6 for an earth station located in California where rainfall exceeds a rate of 22 mm/h for 0.01% of the year at the earth station location, and the elevation angle to the satellite is 45°.

7.8 A DBS-TV receiving station is located in near Billings, Montana, at latitude 45.5°N, longitude 109.0°W. The earth station is in rainfall zone B, and receives circularly polarized signals from a satellite at longitude 119°W.
a. Calculate the look angles for the receiving antenna.
b. Using the procedure set out in Section 7.6 of the text, find the effective rain height for this earth station and the effective path length to the satellite.
c. Find estimated values of the parameters k and α for a circularly polarized signal at 12.0 GHz for rainfall that occurs for 0.01% of an average year. Hence find the specific rain attenuation γ_R and the slant path attenuation $A_{0.01}$ for this rainfall rate.
d. Using the procedure in Section 7.6 of the text, estimate the slant path attenuation for 0.1% and 0.3% of an average year.

7.9 Repeat the analysis of Question #8 for an earth station located near Orlando, Florida, at latitude 28.5°N, longitude 84.5°W, in rain zone N.

7.10 The operator of the DBS-TV satellite wants to ensure that all earth station locations have an availability of 99.7% of an average year. In designing the

coverage of the conus beam of the DBS-TV satellite at longitude 119°W, what is the difference in required satellite EIRP to meet this requirement between the Montana and Florida earth station locations, as calculated in Questions #8 and #9?

7.11 A radiometer is a useful device for estimating slant path attenuation when a suitable satellite beacon is not available. However, the accuracy of the estimate decreases as attenuation increases. The formula for calculating sky noise temperature T_{sky} for a slant path with attenuation A dB is

$$T_{sky} = 280\,(\,1 - 10^{-A/10})\ \text{K}$$

Show that accuracy decreases with increasing path attenuation by comparing the difference in sky temperature between $A = 0.5$ dB and $A = 1$ dB and the difference in sky temperature between $A = 9.5$ dB and $A = 10$ dB.

7.12 Another way that slant path attenuation can be estimated is with a radar that is pointed in the direction of a satellite. The radar output is reflectivity Z in units known as dBz at intervals along the slant path, called range gates in radar jargon. An estimate of rainfall rate can be obtained from a formula relating Z to rainfall rate r in mm/h

$$Z = 10 \log_{10} (A\, r^{b})\ \text{dBz}$$

Commonly used values for A and r with an S-band radar are $A = 200$ and $b = 1.6$, which work well for stratiform rain. Many other values for A and b have been suggested for convective rain.
Specific attenuation γ in dB/km for a rainrate of r mm/h can be estimated from the following formula

$$\gamma = a\, r^{b}\ \text{dB/km}$$

For frequencies below 25 GHz, the coefficients a and b are given by

$$a = 4.21 \times 10^{-5}\, f^{2.42} \text{ and } b = 1.41\, f^{-0.078}$$

where frequency f is in GHz. Adding up the calculated attenuation in each range gate along the slant path gives the total attenuation through the rain. However, the relationship between reflectivity Z in dBz and rainfall rate r in mm/h is logarithmic, so the radar distinguishes rainfall rates very well for rain falling at rates in the 1–10 mm/h and poorly for rain rates of 50–100 mm/h. This is illustrated in the following questions.

- **a.** An S band radar measures reflectivity values in dBz in 20 range gates of 250 m length along a slant path with an elevation angle of 20°. Rainfall falls at a constant rate of 1 mm/h in each range gate. Calculate the reflectivity in dBz using coefficients A = 200 and b = 1.6.
- **b.** Estimate the specific attenuation and the slant path attenuation at a frequency of 10 GHz for this rain event.
- **c.** Later in the rain event, the rain rate in the range gates increases to 2 mm/h. Find the corresponding reflectivity in the range gates and the slant path attenuation at 10 GHz.

Repeat the analysis in parts (a) through (c) with rain rate values of 50 and 100 mm/h.

References

ACTS (1999). ACTS experimental results, presented in a series of proceedings; e.g. "Proceedings of the Twenty-Third NASA Propagation Experimenters Meeting (NAPEX XXIII) and the Advanced Communications Technology Satellite (ACTS) Propagation Studies Workshop", Nasser Golshan and Christian Ho, editors, Falls Church, Virginia, June 2–4, 1999. For historical information on the ACTS program, see https://www.nasa.gov/centers/glenn/about/fs13grc.html (accessed 6 August 2018).

Allnutt, J.E. (2011). *Satellite-to-Ground Radiowave Propagation*, 2e. London, UK: The IET, ISBN-13: 978-1849191500; ISBN-10: 1849191506.

Allnutt, J.E. and Arbesser-Rastburg, B. (1985). "Low elevation angle propagation modeling considerations for the INTELSAT Business Service *ICAP* 85", *IEE conf. Pub.* **248**, 662–666.

Allnutt, J.E. and Rogers, D.V. (1986). System implications of 14/11 GHz path depolarization. Part II: reducing the impairments. *International Journal of Satellite Communications and Networking* 4 (1): 13–18.

Barbaliscia F., Paraboni, A., and Bousquet, M. (1999). Private Communication.

Bostian, C.W. and Allnutt, J.E. (1979). Ice crystal depolarization on satellite-to-earth microwave radio paths. *IEE Proceedings* 126: 951–960.

Castanet, L., Lemorton, J., and Bousquet, M. (1998). "Fade mitigation techniques for new SatCom services at Ku-band and above: a review", COST 255 First International Workshop on Radiowave Propagation Modelling for SatCom Services at Ku-Band and Above, **WPP-146**, October, pp. 243–251.

COST Action 272 (2018). https://www.dir.de/kn/en/desktopdefault.aspx/tabid-12748/22264_read-4900/admin-1 (accessed 13 August 2018).

COST Action 280 (2011). "Propagation Impairment Mitigation for Millimetre Wave Radio Systems", http://www.cost.eu/COST_Actions/ict/280 (accessed 13 August 2018).

Cox, D.C. and Arnold, H.W. (1984). Comparison of measured cross-polarization isolation and discrimination for rain and ice on a 19 GHz space-earth path. *Radio Science* 19: 617–628.

Crane, R.K. (1980). Prediction of attenuation by rain. *IEEE Transaction on Communications* 28 (9): 1717–1735.

Dissanayake, A.W., Allnutt, J.E., and Haidara, F. (1997). A prediction model that combines rain attenuation and other propagation impairments along earth-satellite paths. *IEEE Transactions on Antennas & Propagation* 45 (10): 1546–1558.

Hall, M.P.M. (1979). *Effects of the Troposphere on Radio Communication*. Stevenage, UK: Peter Perigrinus, Ltd.

Hodge, D.B. (1976). An empirical relationship for path diversity gain. *IEEE Transactions on Antennas and Propagation* 24: 250–251.

Hodge, D.B. (1982). An improved model for diversity gain on earth-space paths. *Radio Science* 17: 1393–1399.

ITU-R Recommendation P.618-11 section 2.4, (2013). (in force August 2018)

ITU-R Recommendation P.618-13 (2017). (in force August 2018)

ITU-R Recommendation P.676-11 (2016). (in force August 2018)

ITU-R Recommendation P.837-1 (1994). (superceded)

ITU-R Recommendation P.837-2 (1999). (superceded)

ITU-R Recommendation P.837-7 (2017). (in force August 2018) [Digital rainfall rate maps are available in the supplement file R-REC-P.837-7-Maps.zip]

ITU-R Recommendation P.838-3 (2005). (in force August 2018).
ITU-R Recommendation P.840-5 (2012). (in force August 2018).
Johnston, E.C., Bryant, D.L., Maiti, D, and Allnutt, J.E. (1991). 'Results of low elevation angle 11 GHz satellite beacon measurements at Goonhilly', IEE Conference Publication No. 333, ICAP 910, pp. 366–369.
Laws, J.O. and Parsons, D.A. (1943). The relation of raindrop size to intensity. *Transactions—American Geophysical Union* 24: 452–460.
Mursula, K. and Ulich, T. (1998). A new method to determine the solar cycle length. *Geophysical Research Letters* 25: 1837–1840.
Oguchi, T. (1983). Electromagnetic wave propagation and scattering in rain and other hydrometeors. *Proceedings of the IEEE* 71: 1029–1078.
Report 564 (1982). *Report 564 of the Recommendations and Reports of the CCIR, Volume V, Propagation in non-ionized Media*. Geneva, Switzerland: International Telecommunications Union (ITU).
Rogers, D.V. and Allnutt, J.E. (1986). System implications of 14/11 GHz path depolarization. Part I: predicting the impairments. *International Journal of Satellite Communications and Networking* 4 (1): 1–12.
Stutzman, W.L. and Dishman, W.K. (1982). A simple model for the estimation of rain induced attenuation along earth-space paths at millimeter wavelengths. *Radio Science* 17: 1465–1476.
Watson, P.A. and Arabi, M. (1973). Cross-polarization isolation and discrimination. *Electronics Letters* 9: 516–519.

8

Low Throughput Systems and Small Satellites

The denser a substance is, the faster it will pass vibrations through it. Seismic sounders can capture the movement of the *tectonic plates* well before there are any external indications of violence to come, but that movement can be very quick. Tidal waves when moving above a deep part of the ocean can travel at more than 500 mph, but their amplitude is so small, ships do not notice their passing. However, such *tsunamis* create havoc when they reach shallow water, and the need to provide warning to those on islands or the coast of continents in good time has led to hundreds of buoys being placed around the oceans to provide a warning to those in danger. Such information transfer is time critical.

8.1 Introduction

8.1.1 The Beginning of Long Distance Communications

For many thousands of years, the transfer of information between individuals was very slow. Hand signals and voice commands were passed between people who had, of necessity, to be fairly close. Even then, the chance of being misunderstood was quite high. If a simple warning or "call to arms" needed to be sent over a fairly wide area, smoke signals (Smoke Signals 2017) or signal drums (Drums 2017) could be used. Though the information content was low, such long distance communications could contain critical information. A visual signaling system that contained more complex information was invented in the late seventeenth century (Semaphore flags 2017) and perfected in the early nineteenth century in France into a semaphore line (Semaphore line 2017) that stretched for hundreds of miles. Orders could be sent from Paris to anywhere in Napoleonic France in three to four hours rather than the three to four days a dispatch rider on a horse would (BBC 2017). These fixed communications systems on land – that is the signal towers did not move – started to be adapted for ships that needed a mobility component.

8.1.2 Adding Mobility to Long Distance Communications

In the seventeenth century, communications began to be used between ships using signal flags (Maritime flags 2017), and it has been argued that those used by Nelson at Trafalgar made the difference between victory and defeat. Ships in line astern could

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

pass a message from one to another and it is probable that this was the first over-the-horizon transfer of complex information while on the move (Flaginstitute 2017). Long distance communications by Martello Towers on land using fires to warn of an invasion (Martello Towers 2017) was certainly over-the-horizon, but the bit rate was very slow: possibly one bit per hour (the time it took to restock the tower with fire wood). While the information sent by the Martello Towers was digital (it was either lit – *a one* – or not lit – *a zero*) and thus easy to interpret without error, the other forms of signaling were basically analog in nature and so at the mercy of the person interpreting that information. There is no easy way to introduce error correction techniques in analog messaging, other than to repeat the information many times and hope a majority of the received signals are correctly interpreted. Sending a written message was less prone to error than visual or voice communications, but the problem still remained of how to get the information from the writer (the *transmitter*) to the reader (the *receiver*) in a timely fashion. The discovery of radio waves by Heinrich Hertz in 1887 offered the possibility of signaling information both beyond visible line-of-sight and potentially over very long distances. Guglielmo Marconi (2017a) turned this possibility into fact on 12 December 1901, when he sent the letter "S" across the Atlantic (Marconi 2017b). The radio transmitter and receiver did not have to be on land, or indeed stationary, but all of these messages – whether voice, semaphore, letter, or radio – were essentially narrow bandwidth. When measured by the amount of real information being conveyed per unit time, they were all *low throughput systems*. They also lacked the ability to link, or de-link, with a neighboring node at will, or to search for the optimum path for any given message. The early mobile cell phone service was originally limited to dialing up through an operator, and the size of the so-called handheld devices tended to restrict their practical use to cars or trucks. The invention of the cell system for mobile phones, the move from analog to digital waveforms, and the incredible reduction in the physical size of the handheld units (Mobile Phones 2017), led to an explosion in mobile cell service traffic. The widespread move to digital rather than analog messaging (Digital 2017) over wired transmission links, albeit earlier than for wireless systems, was followed about a decade later by the adoption of Bluetooth for close range wireless operation (Bluetooth 2017a).

8.1.3 Adding Automation to Link Setup

Radio communications between two, or more, users require cooperation between both the sender and the intended receiver. Once an agreed frequency has been set up in both the transmitter and receiver, established protocols are both used to start the interchange of information and to regulate the flow. In voice communications, the same frequency is often used both to transmit and receive; a process called *simplex signaling*. This necessitates the standard "over" at the end of one user's transmission, signifying the other user may now speak. When the conversation is completed, the last speaker states "end." This process is both cumbersome and slow. To speed communications in dense air traffic operations, controllers use the call-sign of an aircraft (e.g., United 918) as a beginning marker for instructions to that aircraft. No marker is used to signify the end of the transmission to flight 918: calling a different flight number signifies the end of that particular transmission. Digital communications between machines is both faster to set up, and able to link more than one device. The IEEE 802 standards detail the frequency, bandwidth, and power levels for digital cable and wireless links in local area networks (LANs).

Bluetooth was conceived as a wireless alternative to RS232 cable and wireless links (Digital 2017). Bluetooth operates in the unlicensed band at 2.4 GHz, which means the operators of Bluetooth devices are not required to coordinate their devices with other possible users. Security is provided by the utilization of fast Frequency Hopping Spread Spectrum (Bluetooth 2017a). The frequencies are changed at a rate of 1600 per second so that any close-by user with a fixed frequency operation would not notice any interference. Another feature that initially provided additional levels of non-interference was that most Bluetooth devices were limited to a range of about 10 m at signaling rates of up to 25 Mbps (Bluetooth 2017b). Later standards permit a range of 240 m and rates of up 50 Mbps. The ability to link up automatically with any other similar device within range is critical to the deployment of small satellites, especially if these are to be launched in dual formation (Prisma 2006), or even cluster configurations (ESA 2017).

8.2 Small Satellites

Almost all small earth satellites are in non-geostationary satellite orbits (NGSOs). NGSO systems are considered in Chapter 9, including one of the first used to rescue sailors at sea (Orbcomm 2017a,b), and the first two global low earth orbit (LEO) communications satellite systems Globalstar (2017) and Iridiumnext (2017a). The National Oceanic and Atmospheric Administration (NOAA) has been active in developing Search And Rescue Satellite systems known generically as SARSAT (Sarsat 2017a). The SARSAT system using 121 MHz beacons predates Orbcomm and has saved lives in Alaska, but few to none in the contiguous US states. SARSAT has been updated to use 406 MHz and geostationary earth orbit (GEO) satellites to transmit global positioning system (GPS) coordinates, reducing the search area from kilometers to meters. All of these systems are low throughput since their maximum information rate per customer is 1 Mbps or less. Increasingly, satellites are becoming multiple-use platforms, with more than one dedicated mission. One such critical add-on payload is automatic dependent surveillance-broadcast (ADS-B). This will provide near-real time position location for any aircraft in flight equipped with the necessary transponder. ADS-B is considered in Chapter 12.

8.2.1 The Genesis of Smallsats

Sputnik 1 ushered in the era of small satellites in 1957. Thirty years after this, Utah State University convened a conference on small satellites. The following year, the conference gained the sponsorship of the American Institute of Aeronautics and Astronautics (AIAA), and the AIAA/USU conference on small satellites has been held annually ever since then. A list of proceedings from these conferences can be found at http://digitalcommons.usu.edu/smallsat. The popularity of these conferences led to a consensus forming around a need to have a degree of standardization on small satellites, in particular what is termed the spacecraft *bus*. The bus contains the orbital control equipment (e.g., three-axis control determination, thrusters) that is common for most missions. This approach follows that of the large satellite manufacturers who develop a standard spacecraft bus into which a variety payloads can be installed. Developing a brand new satellite bus for each mission is uneconomical, and so it is with smallsats: the economics of smallsats is enhanced if a standard smallsat bus can be used

for a variety of missions. In this way, the major design issues are with the new payload. However, while not all payloads will fit into a single size of smallsat, there was still a need to have a form of commonality of bus size for different missions.

8.2.2 Defining the Size of Small Satellites

Traditional communications satellites in geostationary orbit, particularly those providing high capacity links over wide geographic areas, tend to be large, have significant power requirements, and must of necessity be extremely reliable to permit full operational capability over 10–15 years. As we saw in Chapter 3, these demands make the design and construction of large communications satellites very expensive and time consuming. A typical timeframe between mission conception and launch is about five years. A key requirement for such large satellites is that they are multifunctional: they must be able to offer a range of operational frequencies, flexible coverages, numerous power settings, and high internal connectivity.

The first 30 years of geostationary communications satellites saw telephony and video streams over the Atlantic, Pacific, and Indian Oceans remain fairly static in terms of coverages. The Atlantic Ocean Region (AOR) had by far the highest traffic, followed by the Indian Ocean Region (IOR), and then the Pacific Ocean Region (POR). When an AOR satellite became saturated, it was moved to cover the IOR, and then on to the POR. Newer, and larger, satellites were launched into the AOR to fulfill the need for higher capacity. This progression of satellites from the AOR, to the IOR, and then to the POR persisted for three decades. The AOR, IOR, and POR coverages were similar enough for the same satellite to be used for each region. Almost all of the traffic carried by these satellites was analog until the advent of the Intelsat Business Service (IBS) in the mid-1980s. IBS was essentially narrow band with most links operating between the basic digital rate of 64 kbps to T-1 (1.544 Mbps). The rapid acceptance of TCP/IP protocols over terrestrial links in the same time frame, and the concomitant growth of internet traffic, led to a significant demand for regional and domestic traffic over geostationary satellites, rather than trans-oceanic links. Video distribution services to cable TV head ends in the United States began in the 1980s with large bandwidth analog streams. In the United States, such video distribution was the largest user of domestic satellite capacity between 1980 and 1995, when direct broadcast satellite (DBS) TV started to take over.

The design of large geostationary communications satellites to fulfill long-term needs is complicated by the fact that traffic patterns in the twenty-first century are usually not stable for more than a few years. The reaction time between identifying a requirement and launching a large satellite to fill that need is much longer than is desired in a rapidly evolving environment. This is particularly true of the US military acquisition system, and possibly many other services in the western world, where the necessary long lead items cannot be ordered until the overall program has been authorized (Procurement 2018). And the overall program cannot even be submitted for approval to the funding authorities until a need or, in some cases, a threat has been identified. The purchasing process needs to be agile, able to respond quickly to new service requirements, or in the military sphere, new threats. Developing new satellite payload architectures is difficult at the best of times, particularly in the former analog world of telecommunications. The advent of powerful digital signal processing has enabled satellite payloads to be much more flexible, and have significantly less volume and mass than heretofore without sacrificing capabilities. The telecommunications world, as with almost all other aspects of

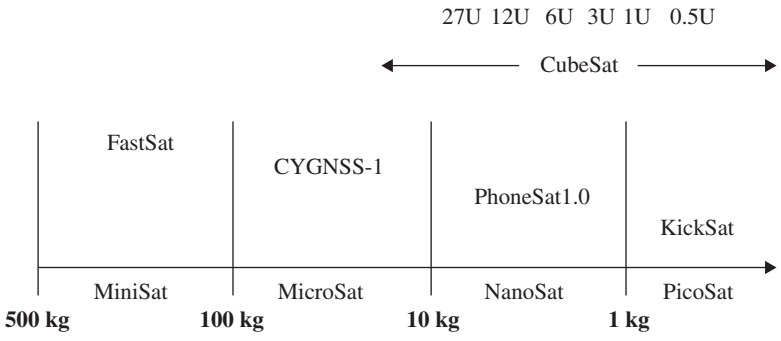


Figure 8.1 Classification of small satellites (smallsats). Examples of some small satellites, the range of the mass of such spacecraft, and their overall physical dimensions, expressed in "U" units. Source: From Fig. 3 of Planet Labs 2017, data © 2018 Planet Labs, Inc. Planet Team (2017) Planet Application Program Interface: In Space for Life on Earth, San Francisco, CA. https://api.planet.com. A U unit (see text) is cubic in shape leading to such spacecraft being called cubesats.

electronic devices, has significantly downsized with regard to mass and volume. To permit the standardization of a range of smaller spacecraft, a unit of one "U" is used to describe the dimensions of the satellite.

8.2.3 Cubesats

A U is defined as a volume of $10 \times 10 \times 10\,\text{cm} = 1000\,\text{cm}^3 = 0.001\,\text{m}^3$. Because of this cubic definition, such small satellites (smallsats) also go under the generic title of cubesats. Figure 8.1 gives a schematic of the range of sizes and approximate volume for the small satellites taken from reference (Planet Labs 2017). Inserted into the figure are some actual smallsats that are within the ranges of mass and cubic volume indicated. One company has advertised that it can build and deliver a smallsat in five days or less (Satellitetoday 2018a) and there is also a proposal to launch a really small spacecraft that weighs approximately 7 oz (Satellitetoday 2018b). This type of ultra small satellite, dubbed *Microns* by the designer, needs to work cooperatively with larger nano and micro satellites (Satellitetoday 2018b). Not all cubesats are designed for LEO, however. One is aimed at cellular backhaul from GEO to serve Kentucky as the CEO of the company says there are still 1 000 000 people in that state without broadband access (Satellitetoday 2018c).

A *thin satellite* has been developed, called *Sprite*, that is even smaller than a cubesat, weighing 4 g and consisting of a printed circuit board (PCB) 2 in. on a side (Sprite 2018). The PCB has a small solar cell and most of the components of a cellular telephone, together with a wire antenna and a gyroscope. Fleets of tiny satellites can work together to achieve a mission that would otherwise require a much larger satellite.

Cubesats can be launched as a single unit, or stacked together somewhat like *Lego*© (Rahmat-Samli et al. 2017) to form a larger unit, and their flexibility has led to a significant increase in their use (Helvajian and Janson 2008). Smallsats are rarely launched singly as the prime payload of a rocket, although the upsurge in interest in smallsats has led to proposals for single launch systems either by direct ascent from a launch pad or on board a rocket taken to high altitude by an aircraft. Chapter 2 discusses these launch systems. Unless the mission calls for a large number of smallsats to be launched together to

form a constellation in orbit, smallsats tend to be launched as a secondary payload, usually being carried to orbit attached to the payload adaptor section of a large rocket that is launching a much bigger satellite. The small size and comparative light mass of a cubesat provides a larger range of launch options. Some smallsats are dispatched toward their intended orbit from the International Space Station (ISS) using a Japanese Experiment Module (JEM) Small Satellite Orbital Deployer developed by the Japanese Aerospace Exploration Agency (JAXA) (Cubesat 2017). Many others have been pushed out into their planned orbit from the ISS using an approach called a *Kaber deployer* developed by NanoRacks that had been used to bring them to the ISS on a delivery flight (Nanoracks 2017a). NanoRacks has increasingly been used to carry into orbit, and deploy, larger smallsats. The biggest in late 2017 was a 100 kg payload for the US Army Space and Missile Defense Command (SMDC) and Adcole-Maryland Aerospace program called Kestrel Eye 2M (KE2M), which is a technology demonstration microsatellite carrying an optical imaging payload designed to track severe weather systems and provide observations on natural disasters (Nanoracks 2017b).

NanoRacks offer both a safe environment for the launch of smallsats and their eventual deployment into the required LEO. Partnered with Boeing, Nano-racks is developing an airlock module to fit onto the ISS, or future in-orbit structures, to assist in exploring the external environment in orbit (Nanoracks 2017c). Part of this effort is investigating the conversion of spent upper rocket stages into useful living space in orbit. They have termed this the *ixion* concept (Nanoracks 2017c). Before sending anything into space, approval must be obtained from the relevant government department in that country that has authority to approve such a launch. In the United States, the relevant authority is the Federal Communications Commission (FCC), and petitioners for launch approval must obtain a *station license*. If the spacecraft to be launched is one of a system of similar satellites, a *blanket license* is given once one of the satellites has already been launched. No other authorization is needed for subsequent launches, but the FCC must be notified before the launch of any other satellites in the constellation. Coordination of satellite systems with other countries is the responsibility of the FCC, who takes the request to the Radio Regulations Board (RRB) and the Radiocommunication Bureau (BR) of the International Telecommunications Union (ITU). The FCC is also the body in the United States that has prime responsibility for spectrum allocations for all radio services (in this context *radio* means all forms of electromagnetic signaling).

8.2.4 Spectrum Allocations

The radio spectrum can be divided, and subdivided into various elements. Figure 8.2 lists the simple divisions of the electromagnetic spectrum (Spaceacademy 2017). Much of the electromagnetic spectrum given in Figure 8.2 does not pass easily through the

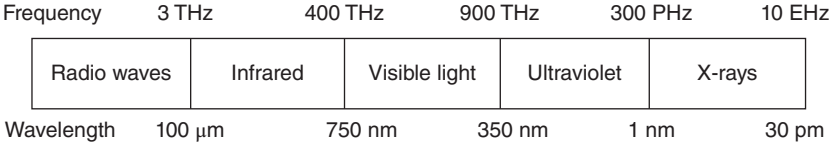


Figure 8.2 Division of the electromagnetic spectrum. In the above figure, abstracted from (Vasseur et al. 1998), the terms are as follows: μm = micrometers (10^{-6}), nm = nanometers, (10^{-9}), and pm = picometers (10^{-12}). THz = Terra Hertz (10^{12}), PHz = Peta Hertz (10^{15}), and EHz = Exa Hertz (10^{18}).

Table 8.1 ITU division of the radio spectrum

Acronym	Designation	Frequency range	Wavelength range
ULF	Ultra low frequency	Below 1 kHz	Longer than 300 km
ELF	Extremely low frequency	1–3 kHz	300–100 km
VLF	Very low frequency	3–30 kHz	100-10 km
LF	Low frequency	30–300 kHz	10–1 km
MF	Medium frequency	300 kHz–3 MHz	1 km–100 m
HF	High frequency	3–30 MHz	100–10 m
VHF	Very high frequency	30–300 MHz	10–1 m
UHF	Ultra high frequency	300 MHz–3 GHz	1–0.1 m
SHF	Super high frequency	3–30 GHz	0.1 m–10 mm
EHF	Extremely high frequency	30–300 GHz	10–1 mm
Sub mm	Sub-millimeter	Above 300 GHz	Less than 1 mm

Reproduced with permission of ITU-R.

atmosphere to space, or back down to earth from space. The portions of the spectrum that are somewhat transparent are termed *windows*. There are really only two such windows through the atmosphere: the visible spectrum, which humans can see through, and the radio spectrum. As we shall see later in this chapter, and in Chapter 7, the radio spectrum is not always readily transparent to a user.

The radio spectrum shown in Figure 8.2 has been further subdivided into portions that have been allocated specific user names by the ITU, such as high frequency (HF) and very high frequency (VHF) shown in Table 8.1. These designations are used mainly by government agencies and the ITU. The letter bands in Table 8.2 are from the relevant

Table 8.2 Frequency bands used in satellite communications (IEEE Std 521-2002)

Letter band	Frequency range
HF	3–30 MHz
VHF	30–300 MHz
UHF	300 MHz–1 GHz
L	1–2 GHz
S	2–4 GHz
C	4–8 GHz
X	8–12 GHz
Ku	12–18 GHz
K	18–27 GHz
Ka	27–40 GHz
V	40–75 GHz
W	75–110 GHz
mm wave	110–300 GHz

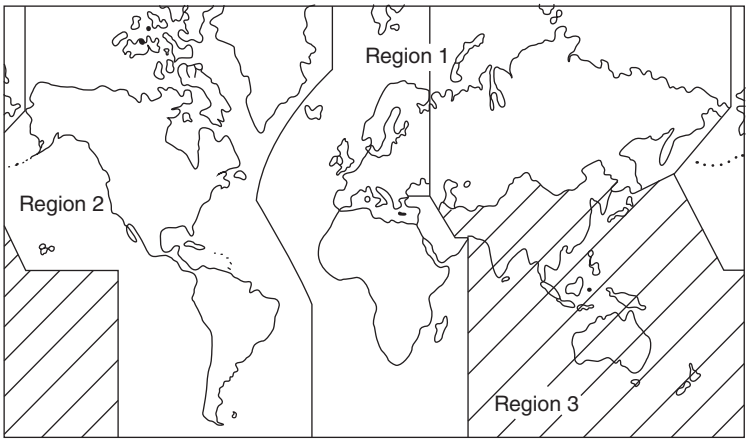


Figure 8.3 The three ITU regions.

IEEE standard and are the designations most widely used in the satellite communications industry, as they divide the spectrum into more useful frequency ranges (IEEE Std 521-2021).

During K-band radar experiments in World War II, researchers found a strong water absorption line close to 22 GHz. The range of the radar was so severely restricted at 22 GHz, K-band was split into two separate bands: one below the absorption line, which was called Ku-band (under K), and one above the absorption line, which was called Ka-band (above K).

The choice of a transmitting and receiving frequency for a satellite is determined by a variety of considerations. For spacecraft operating in the fixed satellite system (FSS) (*fixed* here means the earth stations are fixed on the surface of the earth and not moving around, as in mobile communications), the operating frequencies are closely specified by the ITU. The ITU divides the world into three main *regions*, which are basically Europe, Africa, the Middle East, and Russia (Region 1), the Americas and Greenland (Region 2), and Australia and Southern Asia (Region 3), as is depicted in Figure 8.3. Each of these regions differs slightly in the authorized frequency bands to be used for a variety of services. We will consider Region 2 in what follows. The key for smallsats is the choice of operating frequency.

8.2.5 Operating Frequencies for Smallsats

No specific frequencies have been allocated solely to smallsats. In Region 2, there are a plethora of frequency allocations, both earth-space and space-earth, for a variety of services from around 5 kHz to 75 GHz. In Chapter 7 we saw that ionospheric effects were significant for frequencies below about 4 GHz, and that rain effects dominated at frequencies above 10 GHz. At low frequencies, radio noise generated by the sun, the moon, and the galaxies, in particular our own galaxy, known as the *Milky Way*, can be

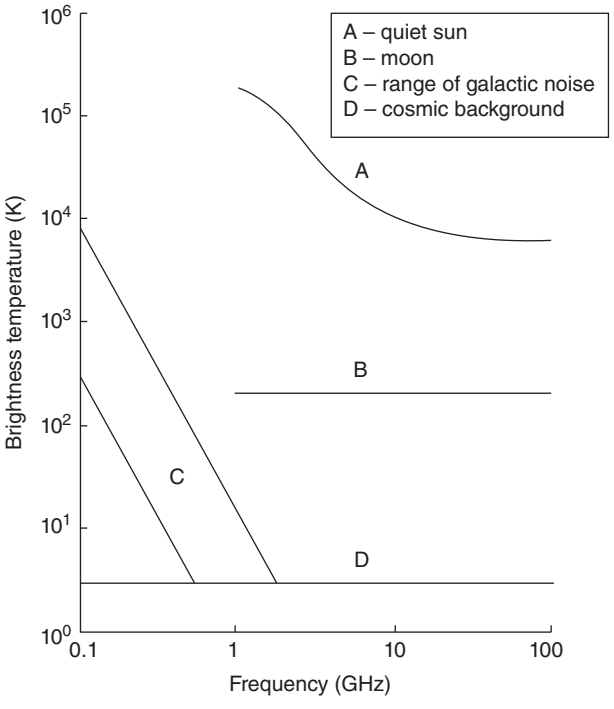


Figure 8.4 Extraterrestrial noise sources. Source: Figure 7 from ITU-R Report 720-1, 1976. Reproduced with permission of ITU-R.

a significant impairment to achieving adequate signal to noise ratios in a receiver at frequencies below 1 GHz.

The standard equation for noise power, kTB, where k is Boltzmann's constant, T is the noise temperature, and B is the bandwidth, can be used to calculate the received noise power. T, also called the *brightness temperature*, emitted by a variety of sources is depicted in Figure 8.4.

The noise temperature of the sun, T_{Sun}, in kelvins, can be calculated for any given frequency, f, using Eq. (8.1) below.

$$T_{\text{Sun}} = \frac{1.96 \times 10^{14}}{f} \text{ K} \tag{8.1}$$

Equation (8.1) assumes that the receiving antenna only includes the sun, which has an apparent diameter of about 0.5°. If the beamwidth of the antenna is much larger than 0.5°, then the noise temperature of the sun must be integrated over the full beamwidth. VHF and ultra high frequency (UHF) antennas usually have beamwidths much larger than 0.5°, and so the sun, while being a significant noise source, does not cause an unacceptable drop in received signal to noise ratio (SNR) at these frequencies. If d is the distance in meters between the transmitter and the receiver, the reduction in signal level, P_{loss}, is proportional to the reciprocal of the distance squared

$$P_{\text{loss}} \propto \frac{1}{d^2} \tag{8.2}$$

Thus, the further apart the transmitter and receiver are, the greater is the loss in the signal level. The overall signal loss in watts is determined by the wavelength of the signal,

λ. The wavelength of the signal is directly related to the frequency, and so care must be taken in choosing the communication frequency.

8.2.6 Selection of Frequency Band to Use

Free space path loss (FSPL) can be found from Eq. (8.3), where d is the distance between the transmitter and the receiver and λ is the wavelength of the signal. Both units are in meters.

$$\text{FSPL} = \left(\frac{4\pi d}{\lambda}\right)^2 \tag{8.3}$$

Example 8.1

Question: A LEO satellite has a distance to an earth station of 1000 km. What is the free space path loss for

a. A radio frequency of 40 MHz.
b. A radio frequency of 1.5 GHz.

Give the answers both in linear and decibel form. Comment on the answers.

Answer
Remembering that c is the velocity of light with

$$c = 3 \times 10^8 \text{ m/s} \tag{8.4}$$

and that the signal frequency f and the wavelength λ are related to c by

$$c = f\lambda \text{ m/s} \tag{8.5}$$

For part (a) the wavelength is 7.5 m and for part (b) the wavelength is 0.2 m.
The FSPL can be found from Eq. (8.3)

a. $$\text{FSPL} = \left(\frac{4\pi d}{\lambda}\right)^2 = \left(\frac{4\pi 1,000,000}{7.5}\right)^2 = 2.8074 \times 10^{12}$$
$$= 124.4830 \approx 124.5 \text{ dB}$$

b. $$\text{FSPL} = \left(\frac{4\pi d}{\lambda}\right)^2 = \left(\frac{4\pi 1,000,000}{0.2}\right)^2 = 3.9478 \times 10^{15}$$
$$= 155.96 \text{ dB} \approx 156.0 \text{ dB}$$

In this example, a frequency of 1.5 GHz has a path loss that is more than 1000 higher than a frequency of 40 MHz. In decibels, it is about 31 dB higher. If the satellite in this example employed a directional antenna with 30 dB gain for transmitting the 1.5 GHz signal, it would still provide a lower flux density on the ground than the 40 MHz transmission with an omnidirectional antenna and the directional antenna would have to be steered to track the earth station as the satellite flies by. However, path loss is not the only thing to consider.

At frequencies below about 6 GHz, the ionosphere causes a rotation of a linearly polarized signal that passes through it, which is why INTELSAT selected circular polarization for its communications satellites that operated in C-band. In certain latitudes and seasons of the year, the ionosphere can cause significant scintillation to occur (see Chapter 7). It can be seen from Eq. (8.3) that the larger the wavelength is, the smaller the FSPL

becomes. There is therefore a power advantage in choosing a relatively low frequency like *VHF*. Many terrestrial video transmissions have been allocated VHF and UHF channels that are no longer utilized in developed countries where DBS TV and cable systems now predominate. These vacant channels have now been called *White Space* due to the absence of utilization by the former services authorized (Urban WiFi 2018). In the United States, the FCC has essentially permitted the unlicensed use of these frequencies, and it is possible that smallsats may be able to use these channels. Heretofore, smallsat operators have tended to make use of frequencies allocated for space research or space exploration, but there is some doubt that this use will continue to be sanctioned as several of the smallsat systems have started to generate revenue streams, which is not considered to be either a research or an exploration operation. A number of smallsat operators began using their satellites as an amateur service, but in the United States, there are a number of rules governing what can be called an amateur service. For example, if a person builds a smallsat as part of her, or his, job as a university professor, this is considered to be part of the business of the university and so cannot be considered as an amateur service. It is strongly recommended that all smallsat activities are coordinated with the relevant governing body of the country, which is the FCC in the United States. The first amateur satellite in the United States was OSCAR 1 launched in 1961. The International Amateur Radio Union (IARU) (Amateur Radio 2018) has not agreed to coordinate experiments in the amateur radio bands, but it is a useful organization to contact prior to starting to build a smallsat that might be operated on an amateur satellite basis. The status of the coordination of any particular smallsat can be found at this AMSAT web site (Amateur Radio UK 2018).

8.2.7 Operational Considerations

The velocity of a LEO satellite is given by Eq. (2.5), which is repeated below as Eq. (8.6)

$$v = \left(\frac{\mu}{r}\right)^{\frac{1}{2}} \tag{8.6}$$

In Eq. (2.5), μ is the product of the gravitational constant G and the mass of the earth M_E. The product GM_E is called Kepler's constant and has the value 3.986004418×10^5 km^3/s^2. If a LEO satellite is in a circular orbit 500 km above the surface of the earth, the radius of the orbit from the center of the earth, r_s in Figure 8.5, is 500 + (the radius of the earth). The mean earth radius is 6378.137 km and so the radius of the satellite's orbit ≈ 500 + 6378 = 6878 km. From Eq. (8.6), the orbital velocity of the satellite is

$$v = \left(\frac{3.986004418 \times 10^5}{6878}\right)^{\frac{1}{2}} = (57.95295752)^{\frac{1}{2}} \approx 7.613 \text{ km/s} \tag{8.7}$$

The circumference of the orbit is $2\pi r$ where r is the radius of the orbit. Hence the circumference of the orbit is given by

$$2\pi \times 6878 = 43,216 \text{ km} \tag{8.8}$$

Given the circumference of the orbit is 43 216 km and the velocity of the satellite is 7.613 km/s then one orbit of the earth will take

$$\frac{43,216}{7.613} = 5,676.5 \text{ seconds} = 94.61 \text{ minutes} \tag{8.9}$$

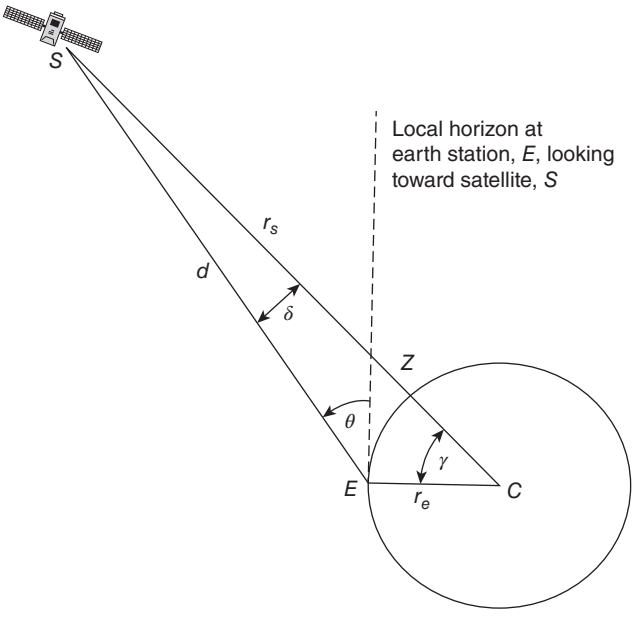


Figure 8.5 Geometry for calculating satellite look angles and coverage area. The satellite, earth station, and the center of the earth are in the same plane. Source: This figure is repeated as Figure 9.12.

With an orbital period of 94.61 minutes, it is instructive to calculate the time an observer on the earth can communicate with this satellite. In Figure 8.5, and from before, $r_e = 6378$, $r_s = 6878$ km, and δ is the angle at the satellite of the coverage of arc EZ. If the coverage from the satellite onto the surface of the earth is symmetrical about the *nadir* (the vertical direction down from the satellite onto the surface of the earth), then the total coverage arc is

$$Coverage = 2 \times EZ \tag{8.10}$$

If the minimum elevation angle at which the satellite can communicate with an observer on the surface of the earth is θ, and we also assume that the motion of the sub-satellite point is from Z to E in Figure 8.5, the satellite will initially be visible to an observer at point E_2, and will be out of communication when it moves to a place in the orbit where the observer at point E_1 is looking on the reverse direction at an angle θ. This geometry is expanded on in Figure 8.6 to show the coverage from the point of view of an observer on the surface of the earth at point G.

If the minimum operational elevation angle for the user at position G in the surface of the earth is 10°, then the observational arc subtends an angle of $(180 - 2\theta) = 160°$ as shown in Figure 8.7.

To determine the observation time available from E_2 to E_1, we need to find the angle β in Figure 8.7. Once the angle β is found in radians, the arc length from E_2 to E_1 can be found from

$$\text{arc length } = 2r_s \times \beta \tag{8.11}$$

The angle α is found using the law of sines, thus

$$\sin\alpha = \left(\frac{r_e}{r_s}\right) \times \sin 100 = \left(\frac{6378}{6878}\right) \times \sin 100 = 0.9132 \tag{8.12}$$

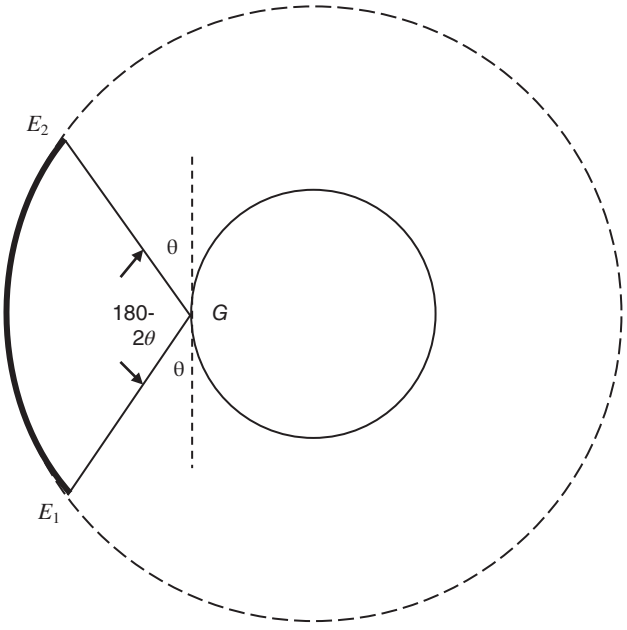


Figure 8.6 Observation arc of the satellite seen from position G as it moves from E_2 to E_1.

Hence

$$\alpha = 65.95^\circ \tag{8.13}$$

and

$$\beta = 180 - 100 - 65.95 = 14.05^\circ = 14.05 \times \frac{\pi}{180} = 0.245 \text{ rad} \tag{8.14}$$

which gives half of the observation arc as

$$r_s \times \beta_{\text{radians}} = 0.245 \times 6876 = 1685.11 \text{ km} \tag{8.15}$$

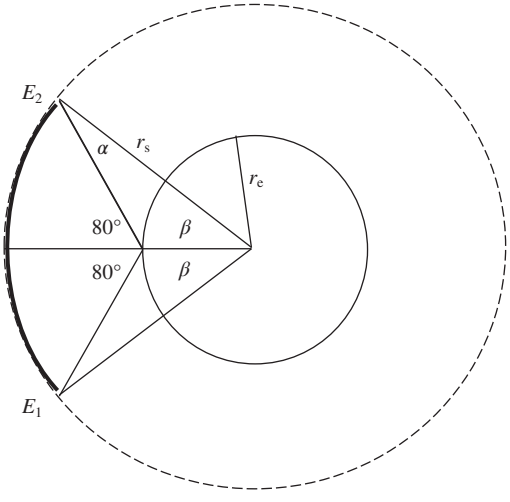


Figure 8.7 Calculation of the observation arc distance from E_2 to E_1.

and the total observation arc is therefore

$$1685.11 \times 2 = 3370.2 \text{ km} \tag{8.16}$$

If the satellite is traveling at 7.613 km/s, the maximum observation time is

$$\frac{3370.22}{7.613} = 442.6928 \text{ seconds} \approx 7.4 \text{ minutes} \tag{8.17}$$

This is the maximum time for which the satellite can be observed when the satellite passes directly overhead. In general, satellites have trajectories that are seen as arcs in the sky with maximum elevation angles between 90° and a predetermined minimum elevation angle. The duration of these passes for this example is always less than 7.4 minutes and will be shorter when the elevation of the pass approaches its minimum value.

An observer on the ground therefore does not have a lot of time to communicate with a LEO at an altitude of 500 km. If continuous communications need to be established with a constellation of such LEO satellites, then more satellites will need to follow in the same plane. If the maximum observational arc is 3370.22 km, another satellite must be in view before contact is lost with the first satellite. The circumference of the satellite's orbit is given by

$$2\pi \times r_s = 2\pi \times 6878 = 43,216 \text{ km} \tag{8.18}$$

The minimum number of satellites needed in a constellation in the same plane to provide continuous coverage will be

$$\frac{43216}{3370.22} = 12.8229 \Rightarrow 13 \text{ satellites} \tag{8.19}$$

Iridium chose to have 11 satellites in each of the six orbital planes (albeit at a higher altitude than in this example) so that there was always good coverage. Having the additional satellites also gave flexibility to provide coverage should one satellite fail. Smallsats are generally in LEO, although two were headed to Mars in June 2018 along with the Insight obiter and lander (NASA Mars 2018). Given the small size of the spacecraft and the many missions they are being used for, it is instructive to calculate some link budgets in a variety of missions.

8.2.8 Link Budgets for Various Smallsat Missions

Chapter 4 sets out the link budget calculating process for any given satellite and earth station. Key components are calculating the FSPL between the transmitter and receiver, the transmitter effective isotropically radiated power (EIRP), and the receiving antenna gain.

8.2.8.1 Earth Orbit Missions

In Example 8.1, we saw that the FSPL for a satellite at an orbital altitude of 1000 km was 31 dB higher for a transmitting frequency of 1.5 GHz when compared with the FSPL of a 40 MHz link. In this link budget example we will assume the following parameters for the VHF and L-Band communications links in Table 8.1. Tx stands for transmitter and Rx for receiver. The satellite antennas in both cases are pointed vertically down toward the earth. Both earth station antennas can track to within their 1 dB beamwidth (see Table 8.3).

Table 8.3 Parameters of the satellite links

	VHF satellite	L-band satellite
Frequency	40 MHz	1.5 GHz
Tx power	10 W	20 W
Tx antenna gain	12.7 dB	15 dB
Tx antenna beamwidth	40°	30°
Orbital altitude	750 km	750 km

Example 8.2

Question: The VHF earth station antenna has a gain of 6 dB and the L-band earth station antenna has a diameter of 3.3 m and an aperture efficiency of 55%. What are the received powers in each case with the earth stations at the nadir point of the satellites?

Answer
From Chapter 4 we saw that the received power, P_r, can be expressed in the link equation, with the parameters in decibel form, as

$$P_r = P_t + G_t + G_r - L_p - L_{misc} \text{ dBW} \tag{8.20}$$

where P_t is the transit power, G_t and G_r are the gains of the transmit and receive antennas respectively, L_p is path loss, and L_{misc} accounts for miscellaneous losses such as gaseous loss along the path.

40 MHz case
We have been given the transmit antenna gain as 12.7 dB and the receiving antenna gain as 6 dB. Before calculating the path loss, we need to find the wavelength for the 40 MHz signal. From Eq. (8.5), the wavelength is can be found by inverting the equation to give

$$\lambda = \frac{c}{f} = \frac{3 \times 10^8}{40 \times 10^6} = 7.5 \text{ m} \tag{8.21}$$

The path loss can be found using Eq. (8.3), which is repeated below as Eq. (8.22)

$$\text{FSPL} = \left(\frac{4\pi d}{\lambda}\right)^2 = \left(\frac{4 \times \pi \times 750000}{7.5}\right)^2 = 1.5791^{12} \text{ or } 122.0 \text{ dB} \tag{8.22}$$

From Eq. (8.20), assuming the miscellaneous losses are 0.1 dB, the received power, P_r, is

$$P_r = 10 + 12.7 + 6 - 122 - 0.1 = -93.4 \text{ dBW}$$

1.5 GHz case
In this case we need to calculate the wavelength and then the receive antenna gain.

Again, inverting Eq. (8.5) we have

$$\lambda = \frac{c}{f} = \frac{3 \times 10^8}{1.5 \times 10^9} = 0.2 \text{ m} \tag{8.23}$$

We have been given the diameter of the antenna ($D = 3.3$ m) and the efficiency ($\eta = 0.55$), so the receive gain of the antenna, G_r, can be found from

$$G_r = \left(\eta \times \left(\frac{\pi D}{\lambda}\right)^2\right) = \left(0.55 \times \left(\frac{\pi \times 3.3}{0.2}\right)^2\right)$$
$$= (0.55 \times 2687) = 1477.85 \Rightarrow 31.7 \text{ dB} \quad (8.24)$$

Before calculating the received power from Eq. (8.20), we need to calculate the FSPL using Eq. (8.3), which is repeated below as Eq. (8.25)

$$\text{FSPL} = \left(\frac{4\pi d}{\lambda}\right)^2 = \left(\frac{4 \times \pi \times 750000}{0.2}\right)^2 = 2.2207^{15} \Rightarrow 153.5 \text{ dB} \quad (8.25)$$

From Eq. (8.20), assuming the miscellaneous losses are 0.1 dB, the received power, P_r, is

$$P_r = 13 + 15 + 31.7 - 153.5 - 0.1 = -93.9 \text{ dBW} \quad (8.26)$$

In this example, the received power at VHF (−93.4 dBW) is almost equal to the L-band case (−93.9 dBW), despite the much higher gain of the L-band antenna (31.7 dB vs. 12.7 dB).

Example 8.3 Lunar Missions

The average distance from the moon to the earth is 384 400 km. A spacecraft with an S-band transmitter is located on the moon and operates at 2295 MHz, a frequency assigned to space to earth links for space research. The transmitter output power is 10 W and a steerable reflector antenna with a diameter of 1.0 m and aperture efficiency 60% on the spacecraft points toward earth whenever the earth is visible. A receiving antenna on earth with a system noise temperature of 25 K is used to receive the spacecraft transmissions. Low-density parity check (LDPC) coding of the data transmitted from the spacecraft allows the threshold carrier to noise ratio (CNR) to be 6.0 dB.

Question: Set out a link budget for the link from the moon to earth and create a table of receiving antenna diameters for an aperture efficiency of 60% with symbol rates of 10 kbps through 10 Mbps. Which combination would you recommend for this project?

Answer

The known and unknown (shown as *TBD*) parameters of this question are tabulated below (see Table 8.4).

The known and unknown parameters, and combinations of them, can be used to find a solution to this question. Specifically, the gain of the receiving antenna on earth and the symbol rate of the transmissions are not given in the question. All calculations are made in decibel units because only addition and subtraction are needed rather than multiplication and division.

We will begin by using the parameters we know to calculate the additional parameters we will need to determine the link budget. The received power, P_r, is found from the link equation (see Eq. (8.20), which is repeated below)

$$P_r = P_t + G_t + G_r - L_p - L_{misc} \text{ dBW} \quad (8.27)$$

where P_t is the transit power in dB watts (dBW), which is $10\log_{10}$ (P in watts), G_t and G_r are the gains of the transmit and receive antennas respectively, L_p is path loss, and L_{misc}

Table 8.4 Lunar mission parameters

Spacecraft antenna beam on-axis gain	G_r	25.4 dB
Path loss at 12.5 GHz, 384 400 km path	L_p	−211.3 dB
Receiving antenna gain, on axis	G_r	*TBD*
Clear sky atmospheric loss	L_a	0.1 dB
Miscellaneous losses	L_{misc}	0.2 dB
Received power, *C*	P_r	G_r − 176.2 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz
System noise temperature, clear sky, 25 K	T_s	14 dBK
Receiver noise bandwidth	B_n	*TBD*
Receive noise power	$N = k\ T_s\ B_n$	B_n − 214.6 dBW
Minimum CNR in receiver		6 dB

accounts for miscellaneous losses such as gaseous loss along the path. From Eq. (8.24), Antenna *G* is given by

$$G = \left(\eta \times \left(\frac{\pi D}{\lambda}\right)^2\right) \Rightarrow 10\log_{10}\left(\eta \times \left(\frac{\pi D}{\lambda}\right)^2\right) \text{dB} \tag{8.28}$$

where η is the aperture efficiency and λ is the wavelength in meters.

The wavelength for a frequency *f* in Hertz can be found from Eq. (8.23) as

$$\lambda = \frac{c}{f} \tag{8.29}$$

The FSPL can be found from Eq. (8.25) as

$$\text{FSPL} = \left(\frac{4\pi d}{\lambda}\right)^2 \tag{8.30}$$

where *d* is the distance from the transmitter to the receiver in meters.

The noise power *N* is calculated from

$$N = k + T + B_n \text{ dBW} \tag{8.31}$$

where *k* is Boltzmann's constant of 1.38×10^{23} J/K, (=−228.6 dB in decibel units), *T* is the system noise temperature of the receiver in dBK, and B_n is the noise bandwidth of the receiver in dBHz, which is set equal to the symbol rate on the link.

Finally, the CNR in the receiver is given by

$$CNR = P_r - B_n \text{ dB} \tag{8.32}$$

To enable us to find out the *TBD* values in the table, it is necessary to begin with some preliminary calculations as follows. Starting with the values given in the problem statement, we can find the gain of the transmitting antenna and path loss for the distance from the moon to the earth. First, however, we begin with calculating the wavelength for a frequency of 2295 MHz. From Eq. (8.29) we have

$$\lambda = \frac{c}{f} = \frac{3 \times 10^8}{2295 \times 10^6} = 0.1307 \text{ m} \tag{8.33}$$

The spacecraft transmitting antenna has a diameter of 1 m and an aperture efficiency of 60%. Using Eq. (8.28), we can calculate the transmit gain of the antenna, G_t, as

$$G = \left(\eta \times \left(\frac{\pi D}{\lambda}\right)^2\right) \Rightarrow 10\log_{10}\left(\eta \times \left(\frac{\pi D}{\lambda}\right)^2\right) = 10\log_{10}\left(0.6 \times \left(\frac{\pi D}{\lambda}\right)^2\right)$$
$$= 10\log_{10}\left(0.6 \times \left(\frac{\pi \times 1}{0.1307}\right)^2\right) = 10\log_{10}(346.6569) \Rightarrow 25.4 \text{ dB} \quad (8.34)$$

The FSPL can be calculated using Eq. (8.30), thus

$$\text{FSPL} = \left(\frac{4\pi d}{\lambda}\right)^2 = \left(\frac{4 \times \pi \times 384400000}{0.1307}\right)^2 = (1.3660 \times 10^{21}) \Rightarrow 211.4 \text{ dB} \quad (8.35)$$

Now using Eq. (8.27), we can find the received power with an earth station antenna gain G_r

$$P_r = P_t + G_t + G_r - L_p - L_{misc} \text{ dBW} \quad (8.36)$$

From Table 8.1, clear sky atmospheric losses are 0.1 dB for the atmosphere at a frequency of 2295 MHz and a miscellaneous factor of 0.2 dB (estimate to cover unknown losses and provide a margin of error). Hence

$$P_r = \quad 10.0 + 25.4 + G_r - 211.4 - 0.3 = G_r - 176.2 \text{ dB} \quad (8.37)$$

The receiver noise power, N, is given by

$$N = k + T + B_n = -228.6 + 14.0 + B_n = B_n - 214.6 \text{ dBW} \quad (8.38)$$

We know we must achieve a CNR of at least 6.0 dB, so we can write

$$CNR = 6.0 = P_r - N = G_r - 176.2 - B_n + 214.6 \text{ dBW}$$
$$= G_r - B_n + 38.4 \quad (8.39)$$

Hence any combination that satisfies $G_r = B_n - 32.4$ dB will work.

We can now look at some possible data rates and calculate the required receiving earth station gain. We will use a noise bandwidth numerically equal to the symbol rate, corresponding to the use of quadrature phase shift keying (QPSK) modulation with half rate forward error correction encoding. We can do the calculation for a variety of data rates (e.g., 10 ksps, 100 ksps, 1 Msps, and 10 Msps). In a like manner, we could have a receiving antenna gain of 7.6, 17.6, 27.6, and 37.6 dB. The calculation process is the same for each of the data rates in this example: we will select 1 Mbps.

Previously, we calculated a gain of 25.4 dB for a 1 m antenna at a frequency of 2295 MHz, so to obtain a gain of 27.6 dB we need an additional 2.2 dB of gain, a factor of 1.66. Gain is proportional to antenna diameter squared, so we need an antenna with a diameter of 1.28 m.

Example 8.4 Interplanetary Missions

Question

These questions are left as exercises for the reader.

a. The same spacecraft in the previous question for a lunar spacecraft is sent to the surface of Mars. The only parameter that has changed in the calculation of earth station antenna size is the distance from the spacecraft on Mars to earth: in this case the

communications distances are much larger. The minimum distance between earth and Mars is 54.6 million km (5.46×10^{10} m), but at its most distant it is 401 million km. Use this distance to calculate the smallest earth station antenna we can use for a Mars to earth link. You can ignore the likelihood of dust storms occurring on Mars and assume that the atmospheric losses are similar to the earth's atmosphere.

b. The distance between earth and Mars varies from a minimum of 54.6 million km to a maximum of 401 million km. The average is 228 million km. Repeat the calculations for the average distance and the maximum distance of Mars from the earth.
c. What is the maximum transmission time for this link?

A large number of spacecraft have been launched on various missions around the earth, to several of the planets in the solar system, and one has been to the Kuiper belt beyond the orbits of Uranus and Pluto. Whether the spacecraft is large or small, the primary mission usually requires that the payload be able to point in a specific direction that requires accurate orbital control. This can be particularly challenging for small satellites that do not have a large payload margin to accommodate the complex orbital guidance systems that are used in many of the larger satellites.

8.2.9 Orbital Control of Small Satellites

The first artificial earth satellites did not have the means to accurately point toward any specific point on earth. While Explorer 1 spun on its longitudinal axis, and so had some stability along one axis, it could not change the direction of its spin axis. Sputnik 1 just tumbled in orbit. For this reason, both spacecraft used *omnidirectional* antennas for communications. That is, the antennas could transmit (and receive) in all directions. While this significantly lowered the radiated (and received) power in the desired direction, it meant the satellite did not have to employ any form of orbital control to point the antenna in a specific direction and so the spacecraft was a relatively simple device to construct. However, the need to increase the *EIRP* – the Equivalent Isotropically Radiated Power – in a given direction, required an antenna with a higher forward gain than an isotropic radiator (see Appendix B). In order to have a high gain antenna point in a given direction, it is necessary to have actively controlled antenna beam pointing, or spacecraft attitude, or both.

Spacecraft attitude control systems for large satellites are described in Chapter 3. Momentum wheels are not a feasible option for a smallsat due to the space and mass they require. By the same token, standard liquid propellant thrusters, whether they use mono- or bi-propellant fuels, exceed both the available space and available mass of a typical smallsat. A thruster utilizes Newton's second law of motion to propel or stabilize a satellite: that is the force P exerted on the spacecraft equals the mass M of the thruster fluid exiting the nozzle of the thruster multiplied by the acceleration a of the fluid. As an equation

$$P = M \times a \tag{8.40}$$

The key point here for smallsats is the relation between the mass and the acceleration of the thruster fluid. If M must be small to allow the use of that thruster on a smallsat, then a must be large. The solution is to use a form of electric propulsion (Haque et al. 2013). If the thruster fuel can be ionized, that is develop a charge, it can be accelerated and controlled using electrostatic and magnetic fields. This is the principle of a Hall

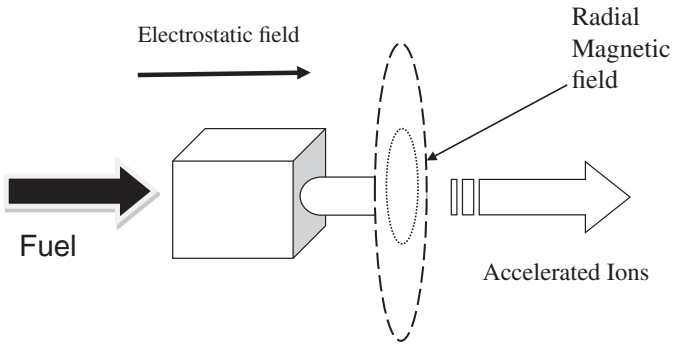


Figure 8.8 Schematic of a Hall effect thruster. The lightweight fuel, either in solid or liquid form, is excited by an electrostatic field and produces ions. The ions are accelerated by a radial magnetic field to velocities in excess of several km/s. An additional electrode (not shown here) neutralizes the ions as they escape to prevent a buildup of charge on the spacecraft structure.

Effect thruster, sometimes generically referred to as an ion thruster (Hall-Effect 2017). This is illustrated schematically in Figure 8.8.

The ideal thruster fuel needs to be a heavy molecule that is easy to ionize. For this reason, initial research concentrated on Mercury as the fuel, but Mercury is a toxic substance (e.g., see Beattie and Matossian 1989). Currently, xenon is the default thruster fuel (Tsay et al. 2016), but it needs to be stored under high pressure (~2000 psi). Research at Busek has identified iodine as a strong contender for an ion thruster (Tsay et al. 2016). Their research showed iodine can be stored as a solid (in any shape), sublimes with minimal heat input, and provides almost the same performance as a xenon thruster. It appears to be an ideal fuel to propel smallsats, as well as provide stable attitude control. The attitude of a smallsat is determined precisely using micro-electro-mechanical systems (MEMS) technology. Like everything to do with electronics, MEMS technology has become very small. Current MEMS attitude sensors incorporate three-axis accelerometers, three-axis gyros, and three-axis magnetometers within a chip approximately $25 \times 25 \times 3$ mm (e.g., Vectornav 2017). Typical power drain is 45 mA @ 3.3 V with a mass of 3.5 g (Vectornav 2017). This size, mass, and power consumption is well within most smallsat margins. Such devices are also utilized in aerial drones of many sizes, providing a number of attitude and position control options (Rotordronemag 2018). Other thrusters for use in smallsats include a plasma thruster (AW&ST 2018b) and one that uses water as a fuel (Satellitetoday 2018h).

8.2.10 Antenna Systems for Small Satellites

The antenna type that provides the highest gain on its main axis for a given aperture size is one that uses a reflector that is a paraboloid of revolution. A parabola has two focal points: one close to the antenna and the other at infinity. For this reason, energy that radiates from one of the focal points is directed toward the other focal point. Figure 8.9 illustrates a simplified cross-section of such an antenna. Appendix B provides more information on directive antennas. The larger the gain of an antenna, G, the smaller the beamwidth, θ, becomes, necessitating increasingly tighter antenna pointing tolerances. The beamwidth of an antenna is generally symmetrical about the electrical

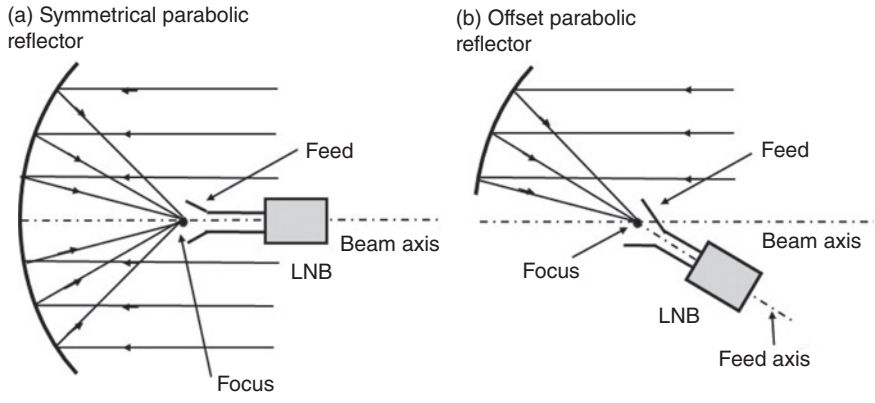


Figure 8.9 Schematic of a two front fed parabolic antennas in receive mode. The left hand antenna has a symmetrical configuration with the feed in the center of the aperture. The right hand configuration uses an offset reflector with the feed below the signal path.

axis, referred to as the boresight of the antenna. Depending on the use of the antenna, two standard beamwidths are quoted. Tracking antennas used for such purposes as missile defense try to keep the target within much less than the 1 dB beamwidth using difference beams. The 1 dB beamwidth is approximately half of the 3 dB beamwidth. Difference beams can achieve pointing accuracies of about one hundredth of the 3 dB beamwidth in a device called a *monopulse antenna* (Monopulse 2018). Communications antennas are normally designed to operate within a 3 dB beamwidth. The 3 dB beamwidth describes the total angle either side of the boresight where the power has decreased by half, that is, $\theta_{3\text{dB}}$. The antenna diameter, D, of a given parabolic antenna, the wavelength of the signal, λ, and the 3 dB beamwidth of that antenna, are typically related as shown in Eq. (8.41).

$$\theta_{3\text{dB}} = 1.3 \times \left(\frac{\lambda}{D}\right) \text{rad} \tag{8.41}$$

The diameter and wavelength must be quoted in the same units, usually meters. In this equation, the distribution of the energy across the aperture from the feed is nonuniform, following an approximately (cosine)^2 aperture distribution with the highest energy directed at the center of the antenna. It is common to simplify Eq. (8.41) into a form that quotes the beamwidth in degrees rather than radians, and good approximations for Eq. (8.41) are:

$$\theta_B = \left(\frac{75\lambda}{D}\right) \text{degrees} \tag{8.42}$$

and

$$(\theta_B)^2 = \left(\frac{30000}{G}\right) \text{degrees}^2 \tag{8.43}$$

Note that, in Eq. (8.43), the gain of the antenna, G, is a linear value, not in decibels. That is, if the gain of the antenna is quoted as 30 dB, the value of G used in Eq. (8.35) would be $10^{(G/10)}$ with G in dB, yielding a linear value of G of 1000.

As can be seen in almost all villages, towns, and cities across the globe, most of the small antennas used for receiving television signals or streaming video from a geostationary satellite are not axially fed, but have the feed horn (or horns) offset from the main axis. In a symmetrical antenna with the feed at the center of the aperture, the feed horn and its supporting structure cause blocking of the radiated (or received) energy. This lowers the gain of the antenna and raises the unwanted energy distribution either side of the main beam, referred to as sidelobes. An off-axis configuration removes the feed horn system from the antenna aperture and thus avoids these problems. In addition to obtaining better off-axis performance, or simply to make the structure of the antenna shorter from back to front, dual-reflector configurations can be used. The two principal cases are the Cassegrain antenna and the Gregorian antenna. These antennas are discussed in Appendix B.

Knowing the frequency of operation of a communications system, f, in Hz, the wavelength, λ, in meters, can be found from the standard equation that relates the velocity of light, c, to the frequency, f, and the wavelength, that is:

$$c = f \times \lambda \tag{8.44}$$

It is instructive to calculate the beamwidth and/or antenna diameter required for a given frequency and compare these to the anticipated dimensions of a smallsat. We see this in Example 8.5.

Example 8.5

Question
What is the 3 dB beamwidth of the following antennas

a. A 10 GHz, circularly symmetric, parabolic antenna with an aperture diameter, D, of 1 m
b. A parabolic antenna with a gain of 28 dB.

Answer
a. Using Eq. (8.44):

$$c = f\lambda \tag{8.45}$$

Given the frequency of operation, f, is 10 GHz, we have

$$\lambda = \left(\frac{c}{f}\right) = \left(\frac{3 \times 10^8}{10 \times 10^9}\right) = 0.03 \text{ m} \tag{8.46}$$

And so from Eq. (8.42), the 3 dB beamwidth is

$$\theta_B = \left(\frac{75\lambda}{D}\right) = \left(\frac{75 \times 0.03}{1}\right) = 2.25° \tag{8.47}$$

b. In this part of the question, we are only given the gain of the antenna (28 dB) with no information on the diameter or the frequency. We therefore must use Eq. (8.43). Converting 28 dB to a linear value gives the gain as 630.9573. It is not normal to use

four places of decimals in an answer, particularly when decibels are being used. However, this is only an interim calculation, so we will keep the four places of decimals for now. From Eq. (8.43) we have

$$(\theta_B) = \left(\frac{30000}{G}\right)^{\frac{1}{2}} = \left(\frac{30000}{630.9573}\right)^{\frac{1}{2}} = (47.5468)^{\frac{1}{2}} = 6.8954^\circ \approx 6.9^\circ \tag{8.48}$$

Note, if we had only wanted the gain of the antenna, we would not normally quote the answer to four decimal places, so 28 dB would convert to a gain of 630.9 (or even 631), giving:

$$(\theta_B) = \left(\frac{30000}{G}\right)^{\frac{1}{2}} = \left(\frac{30000}{631}\right)^{\frac{1}{2}} = (47.5436)^{\frac{1}{2}} = 6.8952 \approx 6.9^\circ \tag{8.49}$$

As we can see, using just one place of decimals, or even none, when using decibels does not change the answer significantly.

In Example 8.5a, the antenna diameter is given as 1 m. A standard 1 U smallsat is a cube with 10 cm sides. Figure 8.10 shows schematically how a 1 m diameter antenna would appear when attached to a 1 U smallsat. It can be seen in Figure 8.10 that the 1 m communications antenna completely dominates the 1 U smallsat to which it is attached. A 1 U smallsat would fit into many small LEO launch vehicles or most adaptor sections of larger rockets, but the size of the 1 m antenna, if it was rigidly mounted onto the smallsat, would preclude many of those rockets being available to launch this spacecraft. If a 1 m diameter antenna is required for this particular mission to achieve the transmission rate required (i.e., the data rate cannot be slowed down to compensate for the smaller gain of an antenna that would fit inside the payload bay with a 1 U smallsat), then the antenna will have to be able to collapse into a smaller volume for launch. This implies the antenna structure must be flexible so as to be able to fold into the available launch space. One proposal, this one operating at S-band (2–4 GHz), envisaged an inflatable antenna

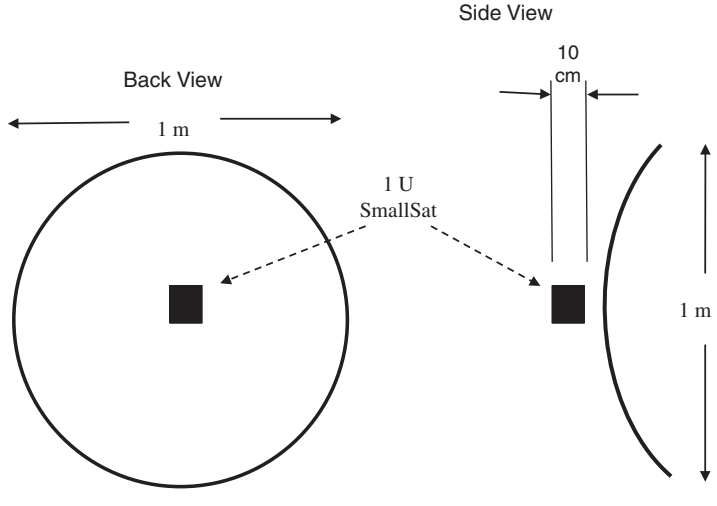


Figure 8.10 Schematic of a smallsat with a 1 m diameter parabolic antenna attached. Note that the 1 U smallsat is totally dominated by the 1 m antenna.

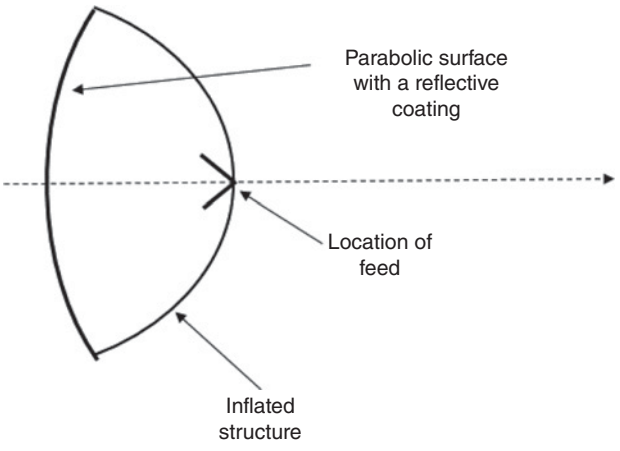


Figure 8.11 Illustration of an inflatable antenna. Not depicted are the supporting feed structure or the feed, which illuminates the inside surface of the inflatable antenna that has been coated with reflective material. The radiated and received energy is along the axis of the inflatable antenna. The transmit direction is indicated by the dotted line.

(Babuscia et al. 2014) with one inner surface of the balloon-like structure coated with reflective material. This concept is illustrated in Figure 8.11.

The stowed volume of this inflatable antenna was 0.6 U (Babuscia et al. 2014). Two other proposals that incorporated solar cells within the panels of a direct radiating array, termed an integrated solar array reflectarray (ISARA), were designed to operate at Ka-band (20–30 GHz) (Hodges et al. 2015) and S-band (Warren et al. 2015). The ability of the individual panels to fold up meant they could fit into relatively small volumes. The Ka-band antenna was developed for a 3 U cubesat (Hodges et al. 2015) and the S-band antenna for a 6 U cubesat (Warren et al. 2015), but the latter was able to fold into a 2 U space inside the spacecraft. Another concept for an inflatable antenna is shown in Figure 8.12.

All of the antenna concepts discussed above could operate in any earth orbit, or indeed on any cis-lunar or interplanetary mission (Hodges et al. 2017). The limited size of smallsats means that, despite their amazing individual capabilities, many missions will require the deployment of several such spacecraft to complete the mission successfully. Such multiple-mission smallsats could be deployed in a variety of ways, but in each deployment, the satellites must be able to communicate with each other. In addition, at least one of them must be able to communicate with an earth-based or space-based control station. One popular concept for the deployment of multiple smallsats is in a cluster.

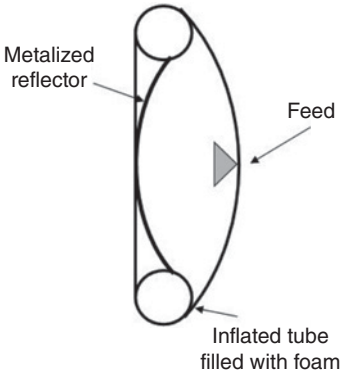


Figure 8.12 Illustration of another inflatable antenna concept. It looks like a rubber dingy with a reflector tied into the bottom and a cover over the top with the feed. The inflator material is a foaming plastic like the stuff used to seal gaps in houses. It sets hard so the shape of the structure is retained.

8.2.11 Cluster Operations

A cluster of smallsats implies a number of such spacecraft "clustered together," orbiting relatively close to each other. Two critical elements determine the success of cluster operations: (i) each satellite must know its position in three-dimensional space precisely, and (ii) each satellite must be able to communicate with any, or all, of the other spacecraft. Position location for LEO satellites is relatively simple using onboard GPS receivers. Once the three-dimensional position location is known, MEMS attitude sensors (Vectornav 2017) provide accurate, real-time, attitude knowledge to allow the electrostatic thrusters (Haque et al. 2013) to orient the spacecraft in all three axes. Using this position control, communications can be established between the cluster of smallsats via the first three layers of the Open Systems Interconnection (OSI) model, that is, physical, data link, and network layer. The physical layer can be a simple Bluetooth channel, or if longer transmission distances are needed, a variety of intersatellite links (ISLs) can be employed (Radhakrishnan et al. 2016; Tiainen 2017), one of which could use lasers (NASA 2017a). The use of lasers for communicating between spacecraft, or for links between a satellite and the earth, has always been an ambition for spacecraft system designers, but it was not until around the second decade of the twenty-first century that both the power and the anticipated operational lifetime of lasers both met the minimum requirements of about 100 mW mean output power and 100 000 hours operational lifetime, particularly the latter. Laser communications, unlike microwave communications, can provide 10 Gbps links with 2 kg terminals (Satellitetoday 2018d). A laser communications company, SpaceDataHighway, reports 10 000 successful laser communications connections between GEO and LEO satellites (Satellitetoday 2018e). Distinctly not low throughput links, SpaceDataHighway noted the 10 000 connections transferred a total of more than 500 TB of data. The very small size of laser systems potentially makes them a good choice for cluster operations. However, space-to-earth communications using lasers will have propagation problems through clouds, and will fail completely if a rain storm crosses the path.

One of the first attempts at cluster operations for smallsats was the Edison Demonstration of SmallSat networks (EDSN) using a cluster of eight smallsats (NASA 2017b). Unfortunately, the first attempt at launching the eight satellites ended in failure in November 2015. The follow-on network demonstration, called NODES (**N**etwork & **O**peration **De**monstration **S**atellite), began successfully in May 2017 with the launch of a pair of satellites (NASA 2017c). These are part of the NASA Pathfinder Technology Demonstration program (NASA 2017d) that seeks to lower the cost and technical risks for future smallsat operations as they move toward more sophisticated missions that will require a degree of autonomous control.

An autonomous control system for a smallsat has, of necessity, to be very small, in keeping with the size of the host spacecraft. NASA began investigating the incorporation of such minute control systems by turning to the mobile cell phone industry. This idea came to fruition in April 2013 with the successful launch of three phonesats from Wallops Island, Virginia (NASA 2017e). Each satellite was a 1 U cube with a mass of about 1 kg, and used a UHF link to transmit data and images for about one week to the control stations on earth before the satellites re-entered the atmosphere. The success of the initial Android cell phone to act as the system controller for the three satellites led to successively more and more advanced cell phone controllers to be developed until (in April 2017) PhoneSat-2-5 was readied for launch (Phonesat 2017). These successes

paved the way for NASA to propose more advanced concepts (NASA 2017f). In the same way that microminiaturization of all facets of spacecraft design and implementation has been a disruptive technology for many satellite applications, the burgeoning concepts and mission plans for smallsats has led to an overhaul of launch vehicle concepts. No longer are huge rockets needed for the majority of the satellite missions (SatelliteToday 2017a), nor indeed do these launches have to take place from large facilities like those at Cape Canaveral and Vandenberg AFB. From being just a relay post in space to forward communications, smallsats have started to dominate most operational applications in space. While at present (April 2018) high throughput geostationary communications satellites provide substantially higher revenues than satellites in other orbits, the balance in revenues between GEO and NGSO satellite streams is expected to narrow. Only time will tell if it will invert.

8.3 Operational Use of SmallSats

8.3.1 Educational

Smallsats were first conceived as a learning tool in 1999 by Robert Twiggs of Stanford University (SatelliteToday 2017b). Working with Jordi Puig-Suari of California Polytechnic State University, they devised a standardized cubic dimension for smallsats, that then were interchangeably called cubesats or smallsats (SatelliteToday 2017b). Instead of just staring at slides projected on a screen or reading a textbook, students could now build a satellite in the laboratory and have a realistic chance of seeing it launched into LEO. To quote from reference (SmallSat 2017): "SmallSats are poised to change the way we do science from space." As part of an initiative to encourage students to move into and, just as importantly to stay in, fields of science and engineering, NASA has an Educational Launch of Nano-satellites (ELaNa) program (NASA 2017g). Competitive launch opportunities are regularly posted and any university in the United States can submit a proposal (NASA 2017g). Three examples of student proposals in 2017 were (Spaceflight 2017):

a. A light-sail project from Cal Poly and Georgia Tech, with the Planetary Society, that will attempt to demonstrate that solar pressure can be used to propel a SmallSat.
b. A Cyclone Global Navigation Satellite System from the University of Michigan that will use a cluster of eight SmallSats to improve extreme weather forecasting.
c. A group of SmallSats aimed at measuring the reflected heat energy from the earth with more accuracy than a single large satellite is a proposal from MIT.

All of the proposals have one thing in common: they relay data toward the earth, where it is collected and sent to a master control station for analysis. Not all of the smallsats will have a visible station on the earth to which data can be sent, so most of them will have to store their data and only transmit it when an earth station is in view. This process was one of the first applications of LEO satellites: store-and-forward data gathering.

8.3.2 Store-and-Forward Data Gathering

Store-and-forward data gathering is rather like preparing for an exam. With luck, a student can store enough data gleaned from her or his lecture notes, before gathering it

together and successfully reproducing it at a later date in an exam. The *storing* of the data in this example is carried out in a different place than the *gathering* of the data. And so it is with LEO store-and-forward satellites. Data, particularly meteorological and oceanic data, are gathered by many measuring instruments, sometimes housed in *beehives*: the small, rectangular, boxes with slatted sides in which a plethora of measuring instruments are housed. Weather data do not usually change rapidly: the temperature may go up by 5° an hour on a hot day; the average wind direction will usually not change by more than 10° an hour; the average wind speed will likewise not change rapidly; and the pressure may change by a few millibars an hour. (The average sea level pressure is 14.7 psi, which in metric units is equivalent to 1013.25 mm Hg, usually referred to as millibars). Using the modern unit of Pascals to measure pressure results in similar numbers since 1 mbar is taken as being equal to 1 hPa (hectopascal).

The relatively slow moving weather data in terms of rate of change permitted strip charts to be used to record the measurements. The beehives were manually inspected every six hours or so, and the paper records taken for analysis to the nearest weather station. This labor intensive process does not work well for remote sites that are largely inaccessible by road, and especially if the weather instruments are located on a buoy at sea. The solution is to have the data stored digitally at the measuring site, rather than in analog form on paper, and then, periodically, to transmit the data to the weather station. However, many of the measuring sites were out of range of most low power wireless transmission systems. The solution was to incorporate a wireless system at the measurement sites that would transmit *upward* to a LEO satellite. Other applications are the tracking of animals that have GPS receivers and a satellite transmitter attached to their collars (GPS 2018, Wildlife 2018). The first satellite system to provide a store-and-forward commercial service was Orbcomm (2017a). The Orbcomm satellites are small, even the second generation has a launch mass of less than 175 kg. The smaller, first generation satellites, were sent into orbit using an air-launched rocket called *Pegasus* that was carried under the wing of a converted Lockheed TriStar passenger aircraft. More recent launches, with the heavier second generation satellites, have been on larger launchers, one being a Falcon 9 rocket that placed 11 Orbcomm satellites into orbit. The Orbcomm satellite constellation, which was probably the first operational LEO system with a revenue stream, has significantly broadened its mission to include tracking of mobile vehicles for trucking companies to optimize vehicle utilization and permit scheduling of routine maintenance. With the rapid increase in M2M (machine to machine) communications, this type of data collection from orbit is likely to increase. But perhaps one area where Orbcomm satellites, and their smallsat companions, will play a big part is in search and rescue operations.

8.3.3 Search and Rescue

The installation of emergency locator transmitters (ELTs) on military and civilian aircraft began in 1970 (Sarsat 2017a). They operated at a frequency of either 121.5 MHz or its second harmonic, 243 MHz. These devices were fairly crude, in that they transmitted no source identification and relied on detection by aircraft overflying the location of the beacon. The need for an Automatic Identification System for such beacons was quickly recognized (Auto ID 2017) and a program that went by the name COSPAS was proposed by Russia. The English translation of COSPAS is "space system for search of vessels in distress." It was quickly recognized that a satellite-based search system would be much

more efficient than one that relied on aircraft being flown to a possible accident location, and a program called SARSAT was formed by the United States, Canada, and France in 1978 (Sarsat 2017a). SARSAT stands for Search and Rescue Satellite Aided Tracking. The original COSPAS program was melded into SARSAT and formed a combined COSPAS/SARSAT system (Sarsat 2017b). At the same time, an upgraded beacon was adopted that used a frequency of 406–406.1 MHz (Sarsat 2017b). The first rescue using COSPAS was in 1982 (Sarsat 2017c). The statistics of rescues by COSPAS/SARSAT since then are remarkable. Data extracted from reference (Sarsat 2017c) show through June 2017:

- Worldwide 41 000+ people rescued since 1982
- United States 8151 since 1982

It is not known what the emergency situation was with each rescue, nor whether they were people lost on land or at sea, or even how they arrived at the rescue location (e.g., crashed light aircraft). Nevertheless, a significant number of people probably owe their lives to being found quickly and rescued. From the initial single LEO system in 1982, COSPAS/SARSAT now incorporates a fleet of five LEO satellites in polar orbit and 5 GEO satellites, with 46 LEO tracking stations and 15 GEO tracking stations that collect emergency beacon signals, plus 29 Mission Control Centers that distribute the COSPAS/SARSAT alert data (Sarsat 2017c). It is very likely that emergency beacon transponders will be located in a variety of LEO and medium earth orbit (MEO) systems (Sarsat 2017c) that will enhance the global reach of COSPAS/SARSAT. The beacon transponders could possibly be incorporated into earth observation satellites as part of a shared payload.

8.3.4 Earth Observation

An earth observation satellite does exactly that: it observes the earth below it. The first earth observation satellite was TIROS 1 (Television Infrared Observation Satellite – 1), which was successfully launched on 1 April 1960 (TIROS 2017). TIROS satellites orbited in a sun-synchronous polar orbit (see Section 9.2.6) that enabled them to have the sun behind them when they were photographing the earth below them, and more importantly the weather patterns. Weather forecasting was never to be the same again. The TIROS satellites initially used optical frequencies, and returned many pictures of the weather below it. TIROS was quickly followed by a more advanced LEO weather satellite Nimbus (Tepper 1961), by multi-spectral observation satellites (Multispectral 2017), and then by geostationary weather satellites called GOES (GOES 2017a). All of these satellites used *passive* observational techniques; either photographing what was below them with cameras or using *radiometers*. An example of such observation is shown in Figure 8.13.

A radiometer detects the emission temperature of what it is looking at. When mounted on a satellite, the radiometer antenna will be directed downward, capturing the emission temperature of cloud formations below. Using different observational frequencies can provide information on not just the type and distribution of weather systems being observed, but their height above the surface of the earth. Low frequencies (1–2 GHz) are not absorbed by clouds unless they are really laden with water, such as in tall cumulonimbus clouds that contain violent rain cells within their core. These frequencies will therefore generally only detect clouds close to the ground. As the satellite

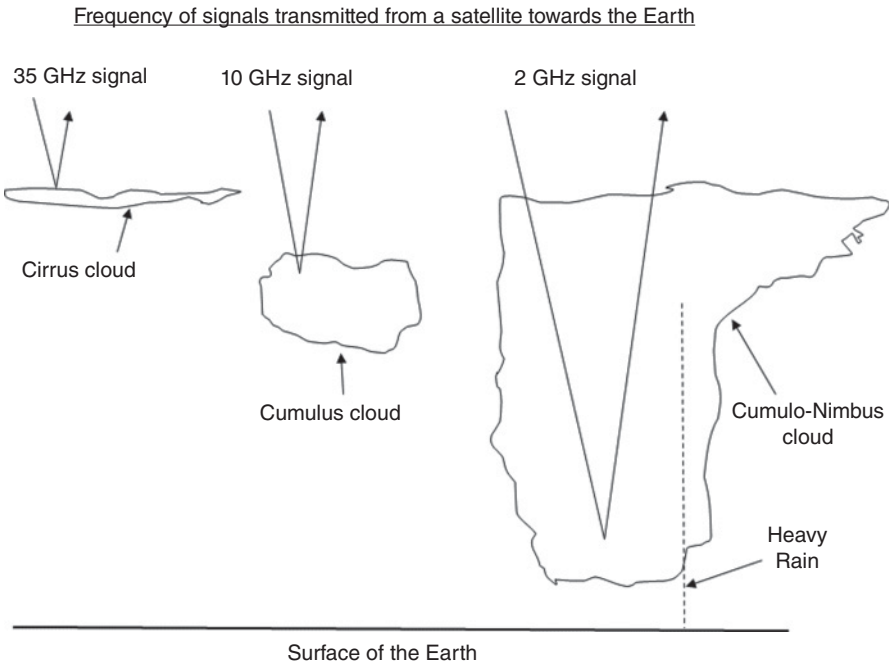


Figure 8.13 Schematic of the penetration depth of different downlink observation frequencies through different cloud formations. The high level cirrus clouds reflect SHFs (super-high frequencies) like 35 GHz. Ku-Band frequencies like 10 GHz will penetrate through cirrus clouds, but not heavy cumulus clouds. UHF frequencies like 2 GHz will penetrate though all but the heaviest rain.

observation frequencies go up, so the penetration through cloud cover decreases until, at around a frequency of 30–35 GHz, only the high level (cirrus) clouds will be accurately detected since these high frequencies will not be able to penetrate through them. Figure 8.13 depicts this schematically.

Cloud formations are usually not static and pictures of cloud cover do not tell the whole story. Clearly, images from a geostationary weather satellite of a hurricane (northern hemisphere) or a cyclone (southern hemisphere) will provide significant information on current location, movement of the eye, and extent of the severe weather. Most weather patterns are not as severe as hurricanes, and most people are only looking to find out whether it is likely to rain or snow on a particular day. The further out in time the predicted weather is, the less accurate it seems to be. One of the factors that leads to inaccurate weather forecasts, and possible the most important one for longer term forecasts, is knowing what the wind is doing. As anyone who has looked up at a cloud formation has often seen, the clouds seem to be going in more than one direction. Clouds are often contained within what is called a *weather front* (usually either a cold front or a warm front). Thus there can be at least two directions for the clouds: one that is moving along the edge of the front, and one that is being carried forward by the front. Knowing the actual speed of the wind, the height of the wind, and the direction of the wind are crucial inputs into forecasting the weather. Finally, if all goes well, by early 2019, weather forecasters will receive data from a satellite called *Aeolus* (AW&ST 2018a). While most weather patterns around the world develop over the oceans, and

sea surface temperature is a good indicator of long term weather, accurate data of the wind will greatly help forecasters. Aeolus will measure 24 horizontal layers of the atmosphere, with thicknesses varying from 0.25 to 2 km (ESA 2018). The strip of atmosphere measured will be 90 km long, with measurements taken by a *lidar* (laser radar) every 0.1 seconds from the *sunset/sunrise* sun synchronous orbit altitude of 320 km. Once each orbit, Aeolus will transmit down data that have already had the Doppler information pre-processed. Aeolus is not designed to observe the surface of the earth. To develop a picture of the terrain on the surface of the earth, a synthetic aperture radar is required.

In a synthetic aperture radar, the frequencies used, combined with the motion of the satellite, enable a three-dimensional picture to be built up of the terrain below with the appropriate digital synthesis. The lower the altitude of the satellite, the better is the three-dimensional picture. In particular, having many synthetic aperture radars operating in sequence will give an enhanced description of the features below (Ravindra et al. 2017). While the geostationary weather satellites can provide 24/7 coverage of a given third of the globe, and so provide accurate pictures of storm movement and characteristics in almost real-time, they cost well in excess of US$1B per spacecraft (GOES 2017b). For this reason, the use of smallsats for earth observation has become popular. Some of the proposals use a pair of sun-synchronous satellites orbiting between 500 and 900 km above the earth (Kuwahara et al. 2013) while others utilize a large constellation of smallsats to provide full earth coverage. One of these (Planet Lab) utilizes 88 smallsats (AW&ST 2017a) the first constellation of which was launched by an Indian rocket into polar orbit in 2017. These spacecraft will be able to provide a new set of worldwide imagery every day. The problem with having so much data, it can be very difficult – and usually extremely time consuming – to manually separate the wheat from the chaff. This is not just a meteorological problem but, in many similar cases, a military intelligence nightmare. To help solve the problem with its own NGSO observation satellites, the US Defense Department has begun an AI program code named *Project Maven* to help pre-sort the data prior to human analysis. Not to be outdone, China is reported to be intending to spend US$150B on AI through 2030 (Seligman 2018).

Another constellation called NanoRacks-Planet Labs-Dove (NASA 2017h) concentrates its initial fleet of 28 smallsats within 52° of the equator. This was a similar orbit inclination to the approach adopted by Globalstar (2017) but completely different to that chosen for the Iridium constellation (Iridiumnext 2017a). The Globalstar satellites orbited at an altitude of 1500 km, while Iridium satellites were in almost polar orbits (inclinations close to 90°) at an altitude of 780 km. Both were low throughput mobile communications systems targeted at voice communications. It is worth noting that LEO satellites provide an excellent platform to observe many aspects of the earth, not least of which is the movement of military hardware. Many military driven applications have been developed and continue to influence some of the designs of LEO smallsats (Military Aerospace 2018).

8.4 Low Throughput Mobile Communications Satellite Systems

We will look at three low throughput satellite systems in this section, Globalstar, Orbcomm, and Iridium. Of these, Orbcomm was the first satellite system designed

for M2M communications. It was also the first LEO communications satellite to generate a revenue stream. However, before discussing the architecture of these satellite constellations, it is instructive to remember the geopolitical situation in which they were conceived and the priorities that drove the system design.

Globalstar was a joint venture between Loral Corporation and Qualcom in the late 1980s. At this point, Loral was a major provider of geostationary communications satellites to, amongst others, INTELSAT and had deep ties to most US telecommunication companies. It therefore sought, in its system design, not to bypass these entities, and so elected to have all of its international traffic bundled together to be carried over submarine or land fiber optic cables. The mobile user would contact a nearby satellite, but the traffic, once sent back down to the local earth station, would be carried to its endpoint via cable. The designers also sought to capture the major traffic users and so elected to use an orbital inclination of 52°. A total of 48 satellites were launched in the original constellation, but in the second generation constellation, 24 satellites were orbited. Iridium took a very different approach.

Iridium, it is believed, was conceived as part of a military observation and communications system in the cold war and so the system design sought to make sure there was no possibility its signals would pass through any communist held territory. This led to the need for ISLs since the constellation was to be in LEO. Laser communications were not feasible at that time with the power and length of operational lifetime needed for such a system, and so Iridium satellites incorporated microwave ISLs. The size of the antennas (2 m in diameter) required the intersatellite antennas (there were four on each satellite: one forward, one backward, and one on each side) not move substantially when operating so as to preserve spacecraft pointing as far as possible, and this led to the adoption of a polar orbit. The original constellation had 77 satellites, but this was later reduced to 66. The material with an atomic number of 77 is iridium, hence the name, but it was facetiously stated that large deposits of iridium were laid down in the same era the dinosaurs were made extinct, and that iridium would make geostationary satellites extinct. Table 9.3 presents an interesting comparison between some of the high capacity NGSO systems and the three low throughput systems discussed in this chapter, Globalstar, Orbcomm, and Iridium. After a review of the Globalstar, Orbcomm, and Iridium systems, Table 8.5 lists some details of each of these systems.

8.4.1 Globalstar

The Globalstar satellites orbit approximately 1400 km above the earth at an inclination of 52°. The initial constellation consisted of 48 satellites, although there will be only 24 satellites in the Globalstar-2 series of LEO spacecraft. Each Globalstar-2 satellite has an expected lifetime of 15 years in orbit. Globalstar is essentially a bent-pipe system: what is received at the satellite is transmitted back down to earth with only a change in frequency to avoid interference between uplink and downlink. In a similar fashion to terrestrial mobile systems, where each user communicates through a cell tower that is connected to a switching center and a locator registry that has information on each user's location and access protocols – there is no direct user to user wireless connection like a CB radio system (Citizens Band 2018). A Globalstar user does not send or receive her/his own signal directly to or from the other user. All links are passed through one of the 24 major terrestrial base stations. The base station detects who is being called and who is calling, and, through its user locator registries, transfers the signal appropriately

to the location of the end user. If the end user being called is in the same satellite coverage region, then the return signal is routed upward from the base station into the same coverage region via the Globalstar satellite. If the end user being called is in a different coverage region, the return signal is not initially routed up through the satellite, but over a fiber optic cable, either to a satellite that covers the region the called party is in, or through a land-line to the end user. The 16 beams of Globalstar satellites allow user signals to pass from beam to beam coverage on the ground as the satellite passes over the region without interference as a different Code Division Multiple Access (CDMA) code is used for each separate user. A signal in the CDMA hierarchy looks just like noise, hence the use of the term pseudo random noise (PRN) sequence for the code generators. Each user is allocated a unique Direct Sequence Spread Spectrum CDMA code. The overall system is designed to be able to transition through the various generations of mobile telephony through 5G and, if necessary, beyond. The mobile frequencies and base station frequencies of the Globalstar system are in relatively similar bands. The range of uplink frequencies is 1610–1621.35 MHz, while the downlink frequencies are between 2483.5 and 2498.5 MHz. Globalstar receivers can add together identical signals transmitted from two satellites, so when one satellite is about to go below the horizon and has a weak signal, another satellite that has just risen above the opposite horizon is transmitting the same signal. The receiver adds the signals at intermediate frequency (IF), which improves the CNR by 3 dB.

Most calls in the initial operating period of Globalstar were statistically not mobile to mobile, but mobile to land line, or vice versa. This statistic has changed significantly with the huge rise in the use of mobile handsets, and has forced almost all providers of mobile telephony to change from a four-wire system to essentially a two-wire system. A two-wire system means that any user speaking into his/her handset cannot hear what the called party says until the called party stops talking, and vice versa.

8.4.2 Iridium

Iridium adopted a digital approach since onboard processing was required to be able to switch the user between the required downlink or ISL, as required. A user opens a channel to a nearby satellite, and this off-hook notification is sent directly to the controlling base station if it is in the same coverage region, or routed over ISLs to a satellite that is in the coverage region of the base station. The location of the called party is determined at the base station and the communications pathway calculated. If this requires a number of different satellites to be used, the signaling information is included in the header, and this is stripped off at each part of the link, until the final satellite in the chain delivers the signal to the end user.

Iridium satellites have 48 individual beams pointed toward the earth, each having about a 30 mi diameter footprint on the ground with a frequency reuse pattern that can provide a total of 1628 cells giving a total of up to 283 272 channels worldwide (Iridiumnext 2017b). Two things should be noted with this calculation of the total number of communications channels worldwide. First, most of the coverages are over oceans and so there will probably be few users in those locations. Second, at any given time, a significant proportion of the satellites are over the two poles, again with few users likely to want to access the satellites. Indeed, Iridium controllers switch out some satellites as they reach the poles in order to conserve power and prevent too much coverage overlap. As a user passes out of coverage of a given beam, the signal is switched to an appropriate

beam in a similar fashion to a mobile earth station on the ground. The appropriate new beam could be on the same satellite, or on another Iridium satellite that is following behind or is to one side or the other of the original satellite. The transmission system used is a combination of frequency division multiple access (FDMA)/time division multiple access (TDMA), with a TDMA frame length of 90 ms, which is divided between the go and return channel since the same frequency is used for the go and return signals. The mobile user will therefore send a transmission for 45 ms and then her/his transmitter will turn off to receive a return signal in the remaining 45 ms slot. This is referred to as *time division duplexing*, or TDD. The Iridium Next series of satellites offer 2.4 kbps for handheld user terminals. For larger, relatively fixed terminals on the surface of the earth, the rate can go up to 128 kbps (Iridiumnext 2017b). The user mobile frequencies are in L-band between 1616 and 1626.5 MHz, with the earth station feeder links in Ka-band between 19.4 and 19.6 GHz and the ISLs also in Ka-band between 23.18 and 23.38 GHz (Iridiumnext 2017b). The crosslinks operate at 25 Mbps.

Ka-band links to the earth stations can be severely affected by radiowave propagation factors: even light rain can cause an outage unless there is a very large link margin (see Chapter 7). To achieve the necessary link margin and high availability required of the earth station feeder links, two earth stations are operated together in a site diversity mode with an Iridium satellite. To allow for continuous operation at handover between satellites, four earth stations are used, two operating in site diversity mode with one satellite, while the other pair of site diversity earth stations prepare to operate with the next satellite coming into view. These are referred to as the Iridium Gateway Stations. At present (April 2018) there are two commercial gateways, one in Arizona, United States and the other in Fucino, Italy. Military gateways for the US government are in Hawaii (Gateway 2018), in Virginia, and elsewhere.

The communication protocol for Iridium is somewhat complicated. A typical Iridium path must go through a gateway station at each end of the link. A handset to handset link goes up to a satellite, down to a Gateway, back to a satellite (or two), down to a gateway, up to a satellite, and down to the other handset. That is six earth-space paths at 15 ms each = 90 ms. Add a frame time of 90 ms, plus TDD delay, gives a total delay for a call of at least 270 ms. Back in the 1990s, Iridium heavily advertised lower delay times than GEO, but never delivered.

Table 9.3 details the significant differences between Globalstar and Iridium on the one hand and two of the high throughput NGSO satellite constellations, O3B and OneWeb, on the other. It is interesting to note the low throughput capabilities of Globalstar and Iridium (less than 2400 equivalent voice channels per satellite) and O3B and OneWeb (up to 12 Gbps). Both the low throughput mobile communications satellites and the proposed smallsat systems are in relatively low orbit. This has significant implications with regard to the observation time available to a user on the ground, which we discussed in Section 8.2.7.

8.4.3 Orbcomm

Orbcomm was the first commercial satellite network dedicated solely to M2M operations (Orbcomm 2018c). The VHF frequencies selected for both uplink and downlink communications ensured that even thick tree cover would not inhibit reception. They also are largely unaffected by even severe weather systems. The Orbcomm system in many respects is like the early communications network setup for the Mercury and

Gemini missions. NASA did not want the manned spacecraft to be out of contact at any time with a control station on the surface of the earth. For this reason, NASA set up a chain of earth stations around the world that could be linked to any manned mission, 24/7. For their part, Orbcomm has set up a chain of 16 Gateway earth stations in 13 countries around the world to track and establish two-way communications with the Orbcomm satellites (Orbcomm 2018c). The network control station is in Sterling, Virginia. By the end of 2017, there were more than two million billable subscribers in the Orbcomm network (Orbcomm 2018d).

The first generation Orbcomm satellites were very small when compared with the other two low throughput satellite systems, Globalstar and Iridium. They weighed less than 45 kg (100 lb) and used a gravity gradient stabilization scheme to orient their satellites. The disc-shaped satellites, once in orbit, lowered a long, coiled, VHF antenna structure such that, due to the gravity gradient along the structure, the end of the antenna always pointed toward the earth. A pair of solar cell panels on the other end of the satellite were able to rotate on the long axis of the satellite and provide power whenever the sun was visible. The low orbit (between 700 and 750 km, with four orbital planes between 47° and 52°) meant that the satellites sometimes could not see the sun, and so the onboard batteries were used to maintain communications. The downlink, sometimes referred to as the *backhaul*, used a TDMA access scheme at 56 kbps (Orbcomm 2018d). Degradations and failures in the initial Orbcomm satellite constellation meant that the response time for users fell from the 1 minute objective to around 4 minutes, but 98% had a response time of less than 15 minutes. The second generation of Orbcomm satellites, known as OG2, are much larger than the first generation of OG1 spacecraft, weighing close to 180 kg (400 lb) (Orbcomm 2018e). In their stowed configuration, the spacecraft body measures 1 by 1 on a side, with a depth of 0.5 m. When deployed from one of the sides, the extended solar cell panels increase that dimension to 13 m, giving an overall spacecraft size of 13 by 1 by 0.5 m deep.

Table 8.5 compares the three LEO systems, Orbcomm, Globalstar, and Iridium. It is interesting to note that both systems are now operating their second generation spacecraft.

All of the three low throughput LEO systems above were operational at the end of 2018, and will probably be operational for at least the next decade. However, they were not the first low throughput satellite systems with a global reach. The first such system was not in LEO but GEO and went under the generic title of VSAT systems.

8.5 VSAT Systems

The acronym VSAT stands for very small aperture terminal and, like many technical terms, it has changed its precise meaning over the years. The first earth station antennas used in commercial satellite communications systems were very large and expensive (Rees 1989), with typical aperture diameters of 30 m. These large antennas operated in C-band (6/4 GHz). Details of the earth station standards used in the INTELSAT network can be found in reference (INTELSAT 2017). With the rapid expansion of satellite telecommunications worldwide, there was a need to make access to the satellite more affordable. This came about in two ways: a significant increase in the transmit power capabilities of satellites and the move to frequency bands above C-band. Both led to a rapid decrease in the size and cost of earth station antennas.

Table 8.5 Comparison of LEO low throughput satellite systems operational in 2018

Parameter	Orbcomm OG2 series	Globalstar second generation	Iridium next
Orbital height and incl.	715-750 km and 47-52°	1414 km and 52°	650-680 km and 86.4°
No. of active satellites	35 active	24 active	66 active
Mass at launch	172 kg (379 lb)	700 kg (1500 lb)	689 kg (1519 lb)
Design lifetime	5 years	15 years	15 years
DC power	400 W at BOL	2.5 kW BOL 1.7 kW EOL	2 kW at BOL
Launcher and date(s)	2012–2016 and Falcon 9	2010–2013 and Soyuz	2017/2018 and Dnepr, Falcon9
Stabilization	Gravity Gradient Boom	3-axis	3-axis
Multiple access	FDMA 189 channels	CDMA	TDD 1100 voice channels
Modulation	DPSK with SRRC $\alpha = 0.4$	BPSK non-differentially encoded	DQPSK comms DBPSK acquisition
User frequencies	148.00–149.99 up 137.00–138.99 down MHz	L-band and S-band with 16 transponders	1616-1626.5 MHz
Gateway frequencies		S-band and C-band 5 Gbps	Ka-band
Satellite antennas	Global beam	16 spot beams	48 spot beams
No. of transponders	One	16	On board processing
Voice telephony	No	Yes	Yes (2.8 kbps)
Bit rate	2400 bps users 4800 bps gateway	9600 bps	128 kbps
Polarization	RHCP	LHCP and RHCP	RHCP
Transponder type	OBP	Bent pipe	OBP

8.5.1 Introduction

Most VSAT systems operate in Ku-band, with earth station antenna diameters of one to two meters and uplink transmitter powers of one or two watts. This will change as the NGSO systems come on line (see Chapter 9) where Ku-, Ka-, and V-band frequencies are proposed to be used. To be able to cope with tracking the fast moving satellites, the small earth stations proposed for these NGSO systems will probably need to use phased array antennas, unlike earth stations operating to GEO spacecraft that do not need to constantly track the satellites. VSAT systems operating with GEO satellites are usually organized into two different architectures: a star network, in which the VSATs

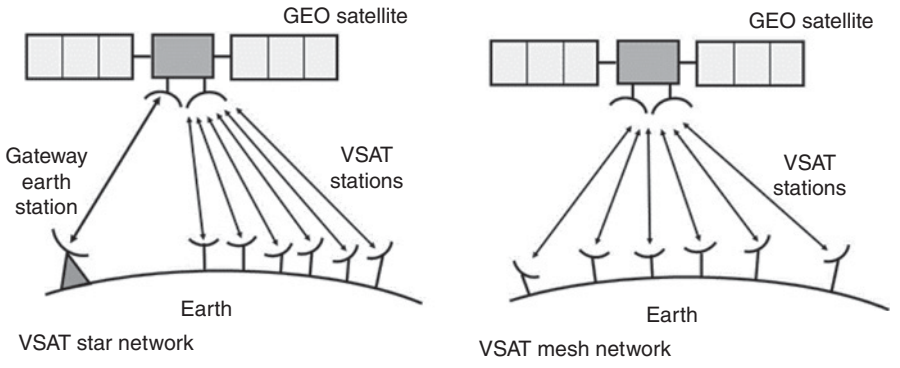


Figure 8.14 (a) Star network. All links go through the gateway earth station. Individual VSAT stations cannot connect directly to each other. This adds substantial delay to the link. (b) Mesh network. Individual VSAT stations can connect directly to each other. Larger terminals are needed than with a star network to compensate for the lack of a large gateway earth station.

connect to each other through a central hub station (the gateway earth station) via a GEO satellite, or a mesh network, where each of the VSATs operates directly through the GEO satellites to every other VSAT. In a star network, the gateway station acts as the network controller. In a mesh network, the network control can be allocated to one VSAT, or the control functions can be distributed amongst the VSATs. Data rates on the links are from a few thousand bits per second up to 256 kbps, depending on the traffic requirements. VSAT systems are used to link businesses and stores to a central computer system so that what are termed *point of sale* transactions can be completed more rapidly than by using a telephone line and modem, and so that a central office can rapidly distribute and collect information from a large number of locations in a region or country. Figure 8.14 illustrates the differences between a star network and a mesh network.

8.5.2 Growth of VSAT Systems

In the 1990s, there was a rapid growth of VSAT networks in the United States. Initially, the most common VSAT architectures were Star networks since the very low receive G/T of the VSATs, coupled with their limited transmit EIRP, was compensated for by using a large hub with high G/T and EIRP. Businesses adopted VSAT networks for the transmission of data as an alternative to using the terrestrial telephone and data systems then available. The next decade is expected to see growth of VSAT networks operating in Ku-, Ka-, and V-band as new GEO and NGSO constellations become available. These networks may operate directly to the home for internet connections and delivery of multimedia material, but in many of the new NGSO systems, a small gateway station will be used to connect the ground users via Wi-Fi or terrestrial cable links. Few VSAT systems are used just for voice traffic, although the data rates are well matched to digital voice bit rates. For this reason, voice over internet protocol (VOIP) became a growth segment in VSAT operations. The underlying concept behind most VSAT systems is to bring telecommunications service directly to the end user without a significant, or indeed any, intermediate distribution hierarchy.

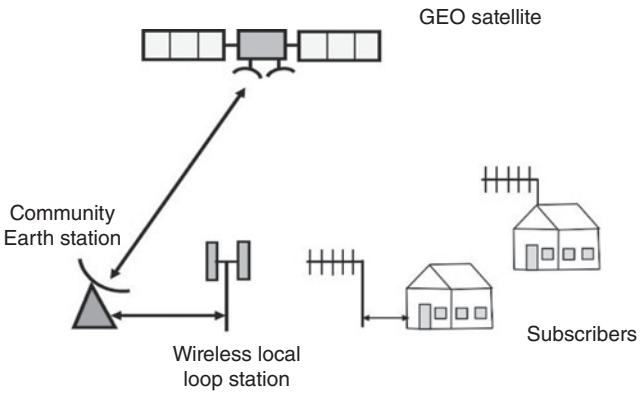


Figure 8.15 Schematic of a VSAT wireless local loop (WLL) concept. The local loop can provide telephony, video, and internet services.

Historically, terrestrial traffic from individual users was bundled together into ever-larger groups and carried over trunk transmission lines via terrestrial microwave systems, satellite systems, or optical fiber cables, before being divided up (demultiplexed) into smaller traffic streams and redistributed to the users at the far end. This is still the most economical transmission architecture for point-to-point communications when the services are being brought into areas with relatively high concentrations of users. Such conditions do not always apply, however, and VSAT networks take advantage of the wide area broadcast capabilities of GEO satellites (see Figure 9.29 that illustrates this potential VSAT service opportunity). The concept of geostationary satellites allied to microwave cellular technologies has been used to bypass completely the traditional expansion of analog telephony. One such solution is a wireless local loop (WLL) coupled with VSAT distribution architectures. Figure 8.15 illustrates the concept schematically (VSAT 2017).

The need to provide high receive gain and transmit EIRP meant the cost of the gateway in a star VSAT network was quite high and, at least for the smaller VSAT networks, somewhat prohibitive. This led to the concept of a shared hub, where several networks operated through one main hub, often referred to as a teleport. The difficulty with this approach for large countries with widely dispersed communities is that the host computers for the small VSAT networks are rarely close to the hub. A high speed terrestrial data link is required between the host computers of the networks and the hub, which increases the cost of the network. Rather than have one large hub for all of the VSAT networks sharing the same satellite, the overall network evolved to allow each sub-network to have its own hub as soon as the economics made it attractive. In this way, the host computer of each VSAT network was co-located with its own hub, thus eliminating the cost of the interconnection between the hub earth station and the computer controlling the service offered through the VSAT network. Whether the hub is shared or dedicated on the one hand or the VSAT is connected to a single user or a LAN with multiple users sharing access through an Ethernet connection on the other, in every case there will need to be an access control protocol.

8.5.3 Access Control Protocols

A satellite communications link occupies primarily the physical layer of the OSI/ISO model, which is where bits are carried between the terminals. ISO is the International

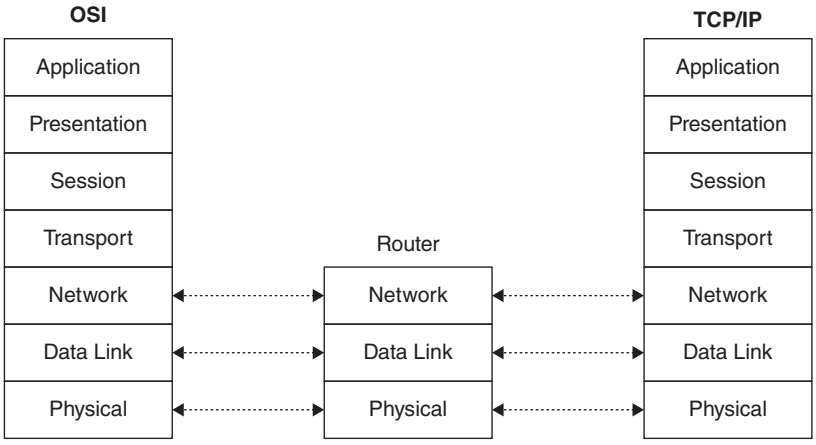


Figure 8.16 Schematic of the OSI/ISO and TCP/IP protocol stacks. The OSI model was developed when machines were connected via X.25 and X.75 digital links. With the introduction of internet connections via TCP/IP, the top three layers of the stack were merged. In a network that has many interconnection requirements, routers are used to pass information from user to user. A router occupies the bottom three layers of the protocol stack. The physical layer just carries the information on a copper wire, fiber optic cable, or radio wave. The data link is managed by headers on the blocks in the data link that describe the sender, the receiver, the amount of data being carried, and ends with a checksum (OSI Links 2018).

Standards Organization and OSI is the Open Systems Interconnection. A schematic of the OSI/ISO model layers is depicted in Figure 8.16 (OSI Links 2018). TCP/IP is the transmission control protocol/internet protocol used in internet traffic.

A VSAT network must have terminal controllers at each end of the link and these occupy the network and link layers, the two layers above the physical layer. The network control center typically controls the system and is responsible for the remaining layers. Unfortunately, few communications systems conform in an easily identifiable way to the seven layers of the ISO-OSI model. (For example, the IP protocol stack of five layers simply puts the first three layers of the ISO/OSI stack into one layer). It is, however, very useful as a conceptual model, which identifies functions that must be performed somewhere in every data communication network. Most data communication networks use some form of packet transmission, in which blocks of data are tagged with an address, error control parity bits, and other useful information before transmission. The receiving end of a link checks arriving packets for errors, and then sends an acknowledgment signal (ACK) that the packet was received correctly, or a not acknowledge signal (NAK) that tells the transmit end to resend a particular packet because the packet had an error. Some systems do not send acknowledgments, only NAK signals to request a retransmission of a packet with an error, since this speeds up data transmission. This is the error control method used in the internet protocol TCP/IP. Generically, such systems are known as automatic repeat request (ARQ). Chapter 5 discusses aspects of packet transmission systems and the problem of error detection and correction in packet networks using satellite links.

The ISO-OSI stack was initially developed for terrestrial communications systems. For this reason, the protocols that implement the functions of each layer were designed for use in terrestrial circuits with low-delay and low bit error rate (BER), that is, very high

performance levels. These are key points when trying to use such protocols over satellites, particularly those in GEO. Many of the early protocols had a connection time-out of a few milliseconds. If no reply was received from the recipient in this interval, transmissions ceased. Similarly, an errored signal received from the source or an intervening node would trigger an automatic error recovery sequence. For example, the X.25 and X.75 packet systems use an ARQ approach, which, on detecting an error in a packet, immediately requests a retransmission and halts further transmissions until the corrected packet is received. Frame relay and ATM (asynchronous transfer mode) systems flag the error but continue the flow of information (continuous transmission ARQ). In both cases, the errored transmission must be corrected and suitable buffers at the receiver end (or intermediate node) used to restore the packets in their original order. The more errors that occur in the link, necessitating many retransmissions of packets, the slower the effective data throughput rate of the link becomes. The potential for delay and (propagation induced) errors are therefore critical design elements in any digital network, particularly VSAT connections over GEO links with the potential for long delays between end users.

8.5.4 Delay Considerations

A typical slant range to a GEO satellite is 39 000 km. The one-way delay of t seconds over such a GEO link is:

$$t = 2 \times \frac{range}{velocity} = 2 \times \left(\frac{39000000}{3 \times 10^8} \right) = 0.26\text{s} = 260\text{ ms} \tag{8.50}$$

The one-way delay in a typical 4000 km transcontinental link via fiber optic cable is a little over 13 ms. Neither example includes processing delay (e.g., source coding and/or compression, channel coding, baseband processing in the switching elements, frame length), which can add several tens of milliseconds or even over a 100 ms. To maintain continuous, uninterrupted communications, it is essential for the link timing to remain within the protocol *window*. TCP/IP typically has a window of 60 ms. If there is a gap in the transmission of information between the end users that exceeds this window, a time out will occur and transmission ceases. As long as the window remains open, communications can continue without interruption. Figure 8.17 illustrates a continuous transmission ARQ system that has a 60 ms window with a 10 ms one-way delay, as found in a terrestrial communications system, and Figure 8.18 illustrates a link with a 60 ms window and a 260 ms delay, corresponding to a GEO satellite link.

Clearly, satellite systems have to operate satisfactorily, and seamlessly (i.e., the user has no idea whether the link is terrestrial or via a satellite), with existing terrestrial networks or their utility is severely compromised. This is particularly true for GEO systems and there are two ways to make terrestrial protocols work with a long delay satellite link. First, the protocols can be changed so that the time-out window is well in excess of 260 ms; second, the satellite element of the packet network can be configured to exist as a separate sub-network within the global packet network. In practice, both solutions are adopted. Figure 8.19 illustrates the concept (VSAT 2015). The VSAT and gateway protocol equipment act as processing buffers to separate the satellite (VSAT) network from the terrestrial network. This is sometimes known as spoofing because the terrestrial part of the system uses a conventional protocol and is unaware of the existence of the

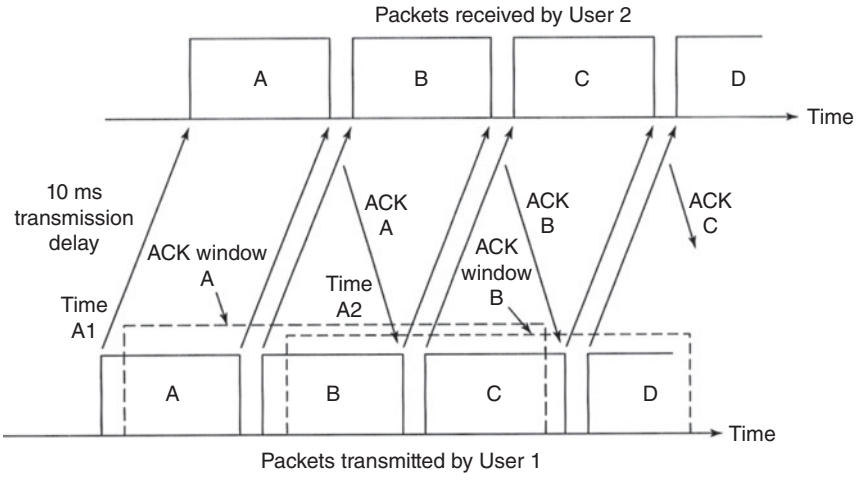


Figure 8.17 Illustration of a communications link with a 10 ms one-way delay and a 60 ms window. In this example, a packet or frame is sent at instant A1 from user 1 to user 2. User 2 receives the transmission without error and sends an acknowledgment back, which is received at instant A2, 20 ms after the initial transmission from user 1. This is well within the time window of 60 ms. The time window rolls forward after each successful acknowledgment. Thus the transmission from user 1 at instant B1 is received by user 2, and the acknowledgment received by user 2 at instant B2, within the new rolling time window of 60 ms. Each packet or frame is successfully received in this example.

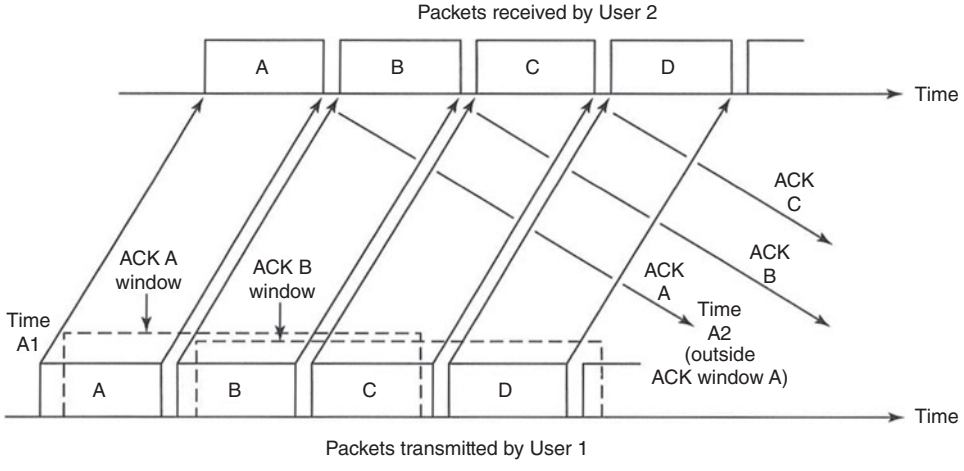


Figure 8.18 Illustration of a communications link with a 260 ms one-way delay and a 60 ms window. In this example, a packet or frame is sent at instant A1 from user 1 to user 2. User 2 receives the transmission without error and sends an acknowledgment back, which is received at instant A2, 260 ms after the initial transmission from user 1. Unfortunately, instant A2 is well after the rolling time window of 60 ms. Transmissions from user 1 are automatically shut down by the protocol when the time-out of 60 ms is exceeded. Ignoring processing delays in this example, user 1 is only transmitting for 60 ms in every 260 ms, thus drastically lowering the throughput. Again, no propagation errors are assumed to occur in the link.

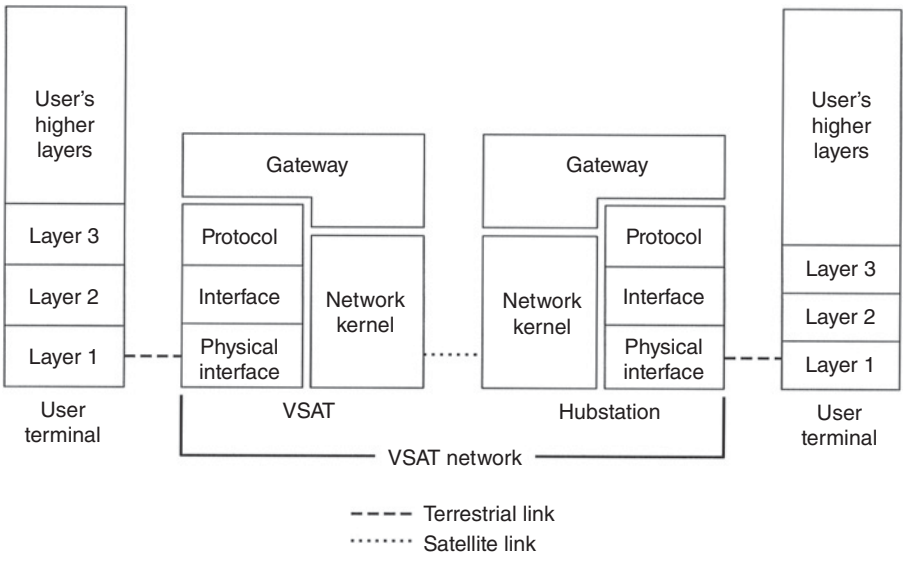


Figure 8.19 Protocol architecture of a Star VSAT network. Source: Figure 2.2.1 of ITU 1994. Reproduced with permission of ITU-R. VSAT networks are normally maintained as independent, private networks, with the packetization handled at the user interface units of the VSAT terminals. The satellite access protocol (with a larger time-out window) is handled in the VSAT/gateway network kernel, which also handles packet addressing, congestion control, packet routing and switching, and network management functions. Protocol conversion and, if necessary, emulation is handled by the gateway equipment.

VSAT network. The electronic processing and emulation permit traffic to flow seamlessly between two very different networks without operator intervention. In essence, this is the interface through which the VSAT user is connected to the VSAT network via the physical layer (see Figure 8.20). Once the user's traffic has moved from the terrestrial network through the interface and is inside the VSAT network, the packet header is reorganized, with the appropriate routing and address of the traffic attached, so that the information can pass successfully over the satellite network to the correct recipient. Network management of the VSAT system, which includes congestion control, is also carried out in this element of the VSAT network, termed the network kernel. In addition, all of the necessary protocol conversions are carried out so that they are mimicked in the VSAT interface/kernel.

In the modulo-8 operation shown in the top part of the figure, the VSAT network simply passes the traffic over the satellite without any change in the protocols in the link layer (layer 2). This results in extraordinarily low throughput for GEO systems. In the lower part of the figure, the bottom two layers of the ISO-OSI stack are formed inside the VSAT network and the modulo-8 operation is changed to modulo-128. The two-layer stack also emulates the other side of the VSAT network so that terrestrial network A believes it is linked directly with terrestrial network B. That is, both terrestrial networks' packets or frames can successfully pass over a satellite connection with a long delay.

A typical data link layer protocol (layer 2 in the ISO-OSI stack) that is used in a low delay, terrestrial link employs modulo-8 operation. That is, the protocol will transmit only seven unacknowledged frames before it stops transmissions; this leads to the

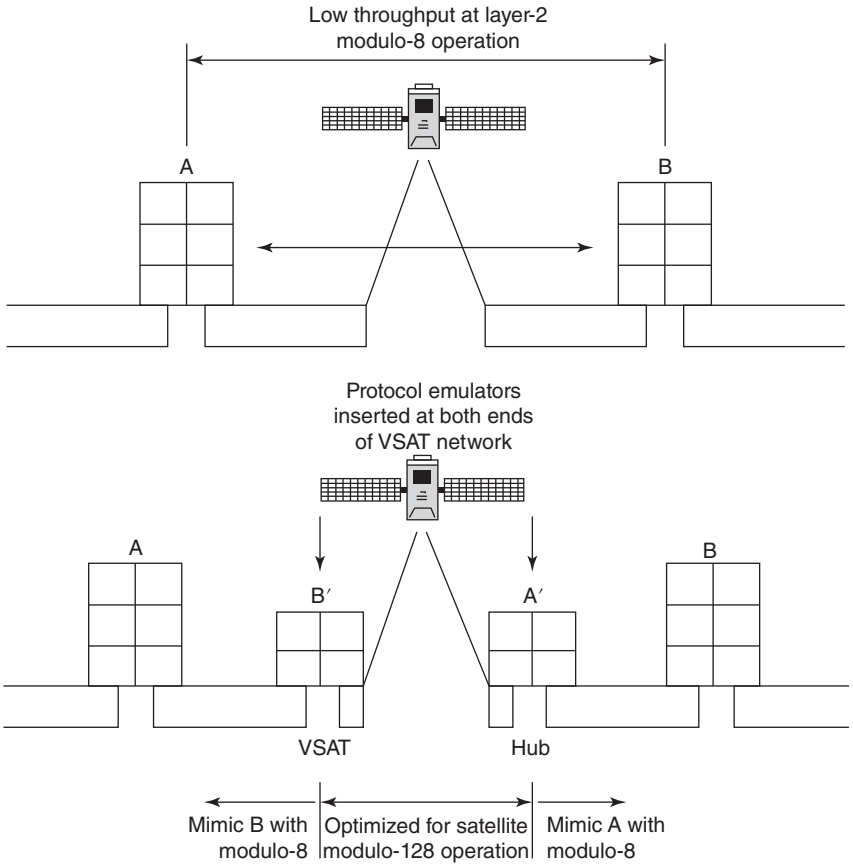


Figure 8.20 Schematic of protocol emulation to permit a VSAT network to cooperate seamlessly with a terrestrial network. Source: Figure 2.2.2 of ITU 1994. Reproduced with permission of ITU-R.

low throughput demonstrated in Figure 8.18, particularly for GEO satellite links. The high level data link control (HDLC) protocol used in layer 2 for satellite systems therefore usually employs a modulo-128 operation. That is, 127 frames may be sent without receiving any acknowledgments before the protocol shuts down transmissions. Moving from modulo-8 to modulo-128 operation significantly increases the window size permitted for the link layer control. The concept, called protocol emulation, is demonstrated in Figure 8.20 (VSAT 2015).

8.5.5 Polling

Another critical function performed in the VSAT interface/kernel sections is to respond to polling activity from the terrestrial packet networks. It is normal for packet networks to poll users to see if there are packets to be sent. The interface/kernel elements in the VSAT network respond to the polling signals of the terrestrial network immediately, thus avoiding the long delay that would occur if the polling signal had to be passed over the satellite link. Negative acknowledgments are made to the polling signals until a request to send data is received over the satellite link. Given that the correct protocols have been inserted at ISO-OSI layer 2 within the VSAT system, and the management

functions have been carried out (i.e., polling, switching, routing, addressing, and flow control) so that the link can operate successfully at a protocol level, there still remains the major part of the system design question to answer: how is the physical connection to be established over the satellite? To answer this question we must move from protocol design/emulation to transmission engineering. First, we will cover some of the basic techniques involved in developing a transmission design.

8.5.6 Multiple Access Selection

As set out in Chapter 6, there are three fundamental multiple access schemes: FDMA, TDMA, and CDMA. Within TDMA, there are two broad sub-divisions of access: those that are closely controlled in time and access ability and those (like ALOHA and other Ethernet-like connections) that are loosely controlled in time and access ability. Multiple access schemes that do not closely control time, frequency, and/or code are significantly less efficient than those that do (Raychaudhuri and Joseph 1988). Pure ALOHA, which is a random access scheme, has a maximum throughput of 18.4% (Raychaudhuri and Joseph 1988). By combining some aspects of TDMA with the random access of ALOHA, slot reservation ALOHA can have an efficiency exceeding 60% (Raychaudhuri and Joseph 1988; Abrahamson 1993). Slot reservation is akin to a controlled access TDMA scheme with a very large frame. The intended application and the potential interference environment often determine the choice between FDMA, TDMA, and CDMA for VSAT networks, with economics also playing a major part. FDMA generally offers the lowest costs for entry-level VSAT systems from the user's perspective since the receiver bandwidth and terminal transmit power required are the lowest.

These systems carry thin route traffic, typically the equivalent of one digital voice channel at 64 kbit per second. The occupied bandwidth, B, of an radio frequency (RF) channel carrying a digital signal with a symbol rate R_s and using error control coding with a code rate R_c is given by

$$B = R_s \left(\frac{1+\alpha}{R_c} \right) \tag{8.51}$$

where α is the roll off factor of the square root raised cosine (SRRC) filters in the link. For example, in a link using QPSK modulation where two bits of information are carried by each transmitted symbol, a message information rate of 64 kbit per second results in a transmission symbol rate of $R_s = 32$ ksps. If the message data bits are encoded with one-half rate forward error correcting (FEC), code rate $R_c = ½$, and the occupied bandwidth, B_{occ}, required for a 64 kbit per second signal is

$$B_{occ} = 32000 \times \left(\frac{1+\alpha}{1/2} \right) = 64000 \times (1+\alpha) \Rightarrow 64(1+\alpha) \text{ kHz} \tag{8.52}$$

Typical values of α for satellite links lie between 0.25 and 0.35, with the higher value being easier, and thus cheaper, to realize when conventional analog filters are used in the transmitter and receiver. If an $\alpha = 0.35$ SRRC filter is used, the occupied RF bandwidth of a 64 kbps QPSK signal with half rate FEC is

$$B_{occ} = 64 \times (1 + 0.35) = 86.4 \text{ kHz} \tag{8.53}$$

A VSAT that is required to transmit a 64 kbps stream of data using QPSK modulation, half rate FEC, and a SRRC filter with a roll-off factor of 0.35, therefore needs an RF channel bandwidth of 86.4 kHz and has a receiver noise bandwidth 64 kHz. Note that

the roll off of the SRRC filter, while adding additional spectrum requirements for the signal, does not alter the noise bandwidth; all RF and IF SRRC filters used in digital radio links have a noise bandwidth in hertz equal to the symbol rate in symbols per second. In practice, a guard band will have to be added between FDMA channels so that adjacent signals do not overlap in frequency at the satellite, and to allow the filters in the receiver that extract individual channels to roll off between channels. It is interesting to note that the DVB-S2 specification (2004) allows α to be 0.2, 0.25, or 0.35. An alpha value of 0.25 seems to be the most popular choice (see Chapter 5).

Most VSATs will operate unattended for most of the time and will be exposed to all weathers. The frequency of the transmitter and receiver RF local oscillators may therefore drift. For this reason, fairly large guard bands need to be designed in, which generally leads to a spectrum allocation at the satellite around 120 kHz for each 64 kbps (voice) channel of the type described above. Better stability in the oscillators allows closer spacing of the FDMA channels with guard bands around 10% of the channel width. For a channel that occupies 86.4 kHz, a guard band of 9 or 10 kHz would be typical, with a channel spacing of 96 kHz. This gets more channels per transponder bandwidth, which translates to additional revenue. This situation is illustrated in Figure 8.21. For a more detailed explanation of FDMA systems on satellite links, please see Section 6.5.

In Figure 8.21, the 64 kbps data stream, could be derived from a point-of-sale device, for example a credit card reader, an internet access request, and so on, all of which require onward transmission over a VSAT network. The 64 kbps equivalent voice channel shown in Figure 8.21 is the output of the terrestrial/satellite interface equipment, after the required emulations and protocol conversions have taken place prior to transmission over a satellite network. This channel from the VSAT to the network controller via the satellite is called the *inbound* or *inroute* channel.

8.5.6.1 FDMA

The RF transmission to the satellite from the VSAT will have a frequency that falls within the bandwidth of a specific transponder on the satellite. If the transponder operates in a bent pipe mode, with no onboard processing, the satellite will retransmit the multiple VSAT channels on the downlink with exactly the same channelization as on the uplink. Thus, in the example used in Figure 8.21, a transponder with 36 MHz bandwidth transmits 375 channels to the network controller or other VSATs in the network. If the VSAT network is being operated in a MESH mode, each VSAT must have a frequency synthesizer that allows it to select any of the 375 possible downlink channels, and the network must have a control channel that tells each terminal at which frequencies it should receive and transmit.

It is more usual, however, for an FDMA VSAT system to operate in a Star mode, illustrated in Figure 8.14a. The network controlling earth station is therefore designed to receive all 375 downlink channels. The digital signal in each channel is recovered and the address information read off so that the network controller can forward the information to the intended user. If the required end-user is external to the VSAT network (i.e., in the terrestrial network) the information is passed through the hub interface equipment and on to the public switched telephone network (PSTN). If the required end-user is within the VSAT network, or a response has been received through the interface equipment from the terrestrial PSTN, all of the information to be transmitted back to the various VSAT terminals is re-assembled at the control station into a return channel.

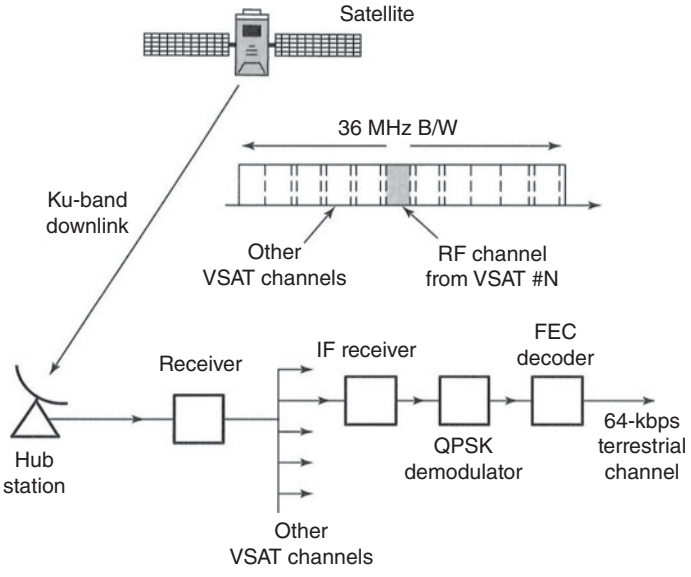


Figure 8.21 Schematic of 64 kbps equivalent voice channel accessing a satellite using FDMA. The 64 kbps equivalent voice channel is in a bandwidth of 86.4 kHz when transmitted to the satellite. The bandwidth of the satellite (from frequency f_1 to frequency f_2) is divided up, or channelized, into increments of 86.4 kHz so that a large number of VSATs can access the transponder at the same time. Each of the 86.4 kHz channels requires a certain amount of spectrum on either side to guard against drift in frequency, poor VSAT filtering, and so on. 86.4 kHz channels plus the guard bands on either side add up to a channel allocation of about 96 kHz per VSAT. From a spectrum allocation viewpoint, therefore, a typical 36 MHz satellite transponder would permit the simultaneous access of 375 VSATs, each of which is transmitting the equivalent of a 64 kbps voice channel. Because each VSAT uses a single channel continuously on the uplink, this is often referred to as single channel per carrier FDMA or SCPC-FDMA. Note: It is unlikely that a Star network would carry voice channels due to the double hop nature of the system; communications would be restricted to data or possibly a few voice channel to the gateway.

The return link from the control station to the satellite, and from thence to the individual VSAT terminals, is not normally sent as a multitude of narrowband FDMA channels. In most cases, the return channel from the hub to the VSAT terminals, called the *outbound* or *outroute* channel, is a single, wide-band stream in a time division multiplexing (TDM) format. In the TDM stream, the separate, low data rate, narrowband signals for the individual VSATs are assembled in a predetermined format so that each of the VSATs can extract the required information destined for that VSAT. Figure 8.22 illustrates the TDM downlink concept used in FDMA VSAT STAR networks. Note here the important difference between TDM and TDMA, which are often confused. TDM is not a multiple access technique. Digital signals from various sources are assembled into a single, high-speed data stream at one point, such as the controller of a VSAT network, and then transmitted as a single continuous stream. In TDMA, several sources, such as earth stations, transmit in coordinated time slots so that a sequence of RF signals is assembled at the satellite.

In the FDMA Star VSAT network examples shown in Figures 8.21 and 8.22, the VSAT network is quite large. Separate transponders are used for the inbound and the outbound channels. In many VSAT networks, the total instantaneous capacity required does not

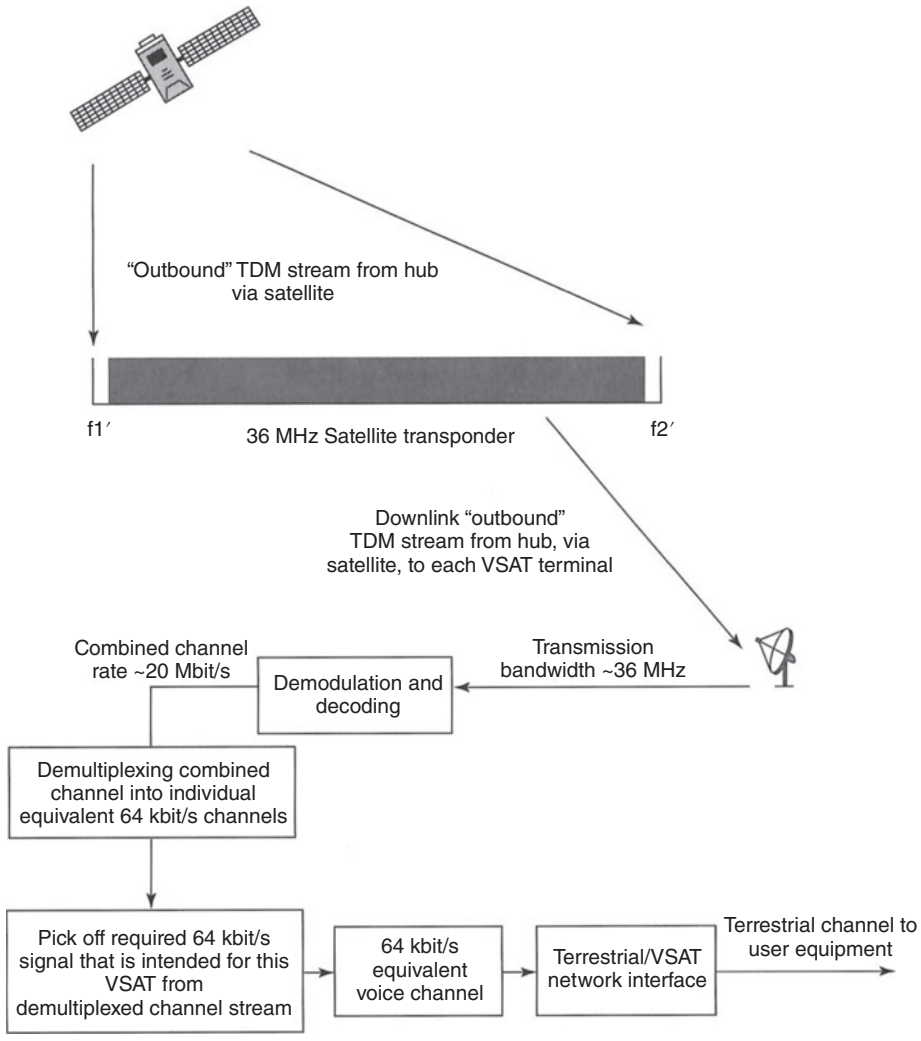


Figure 8.22 Schematic of the TDM downlink outbound channel from the control station, via the satellite, to the individual VSAT terminals. The 375 individual, narrowband, inbound channels received at the control earth station from the VSATs are sent back to the VSATs in a single, wideband, outbound TDM stream at a combined transmission rate of ~20 Mbps. Each VSAT receives the downlink TDM stream and then demodulates and decodes it (i.e., changes the modulated band-pass signal into a baseband line code and removes the FEC.) The line code is then passed through a demultiplexer, which is used to extract the required part of the stream that contains the 64 kbps data channel destined for that VSAT terminal. Carrier recovery and bit recovery circuits are used in the receiver in order to be able to identify the exact position of the required VSAT channel in time. The bandwidth of the satellite transponder (from frequency f_1' to frequency f_2') is fully occupied in this example.

justify two separate transponders for the inbound and outbound signals. Such a case, using a shared transponder, is illustrated in Figure 8.23.

In designing FDMA links, care should be taken to allocate the correct transmit power per channel in calculating the link budget so that the power spectral density is the same for every channel. For example, if a 54 MHz transponder is operated at an output power of 54 W, the power spectral density at the transponder output is 1 W per MHz. A single

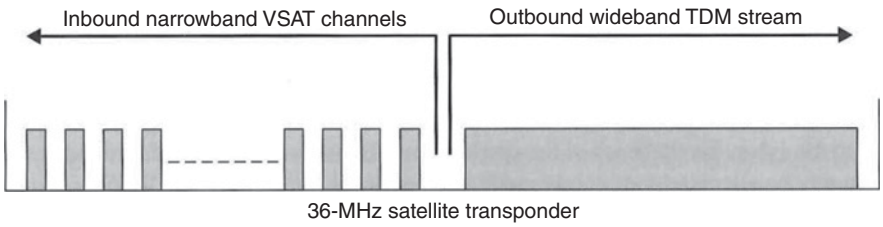


Figure 8.23 Illustration of a VSAT network frequency assignment in which the inbound and outbound channels share the same satellite transponder. In the example here, 18 MHz of spectrum is allocated to each side of the system connection. On the uplink to the satellite, the collection of FDMA narrowband channels transmitted by the VSATs coexists in the same transponder with the wideband TDM stream transmitted up by the control earth station. On the downlink from the satellite, the control earth station receives the collection of individual narrowband channels while the wideband TDM downlink stream is received by each VSAT. The precise frequency assignment can vary to suit the capacity of the VSAT network.

inbound 120 kHz downlink channel transmitted by the satellite will therefore have a transmit power level of

$$\left(\frac{120\text{ kHz}}{54\text{ MHz}}\right) \times 54\text{ W} = 120\text{ mW} \tag{8.54}$$

This transmit power level is multiplied by the gain of the antenna in the direction of the control station (less feed and other losses) to give the effective isotropic radiated power, or EIRP, per channel. Operating a transponder in an FDMA mode also requires careful power balancing to ensure linear operation, which requires the output amplifier to be backed off to obtain quasi-linear operation.

The non-linearity of the transponder at output levels close to saturation causes the generation of third order intermodulation products that degrade the CNR ratio in the single channel per carrier (SCPC) channels (see Chapter 6 for details). Amplifier output back off values of between 3 and 7 dB can be found in the VSAT literature. The value required in any particular case depends on the transponder non-linearity characteristic, the number of RF channels carried by the transponder, and the extent to which the power spectral density of each RF channel in the transponder is matched. Back off at the transponder output lowers the channel EIRP, and therefore degrades the downlink CNR. Automated transponder loading plans are used to optimize the back off at the transponder output such that the overall CNR of the link is maximized. The problem of optimizing transponder back off becomes particularly difficult when the bandwidth is split between the inbound and outbound directions. The end-to-end gain setting of the transponder controls the back off; it may be impossible to optimize the gain for both directions at the same time. Linearization of the transponder can be employed to decrease the back off required (see Chapter 10 for an example).

8.5.6.2 TDMA

A TDM downlink format is sometimes paired with a TDMA uplink plan, particularly for some advanced multimedia services that operate in Ka-band (30/20 GHz). Uplink FDMA formats are not as bandwidth efficient as TDMA. On the other hand, a VSAT that uses TDMA on the uplink is required to transmit at the full burst rate of the TDMA scheme, and must therefore have a much more powerful transmitter than an SCPC

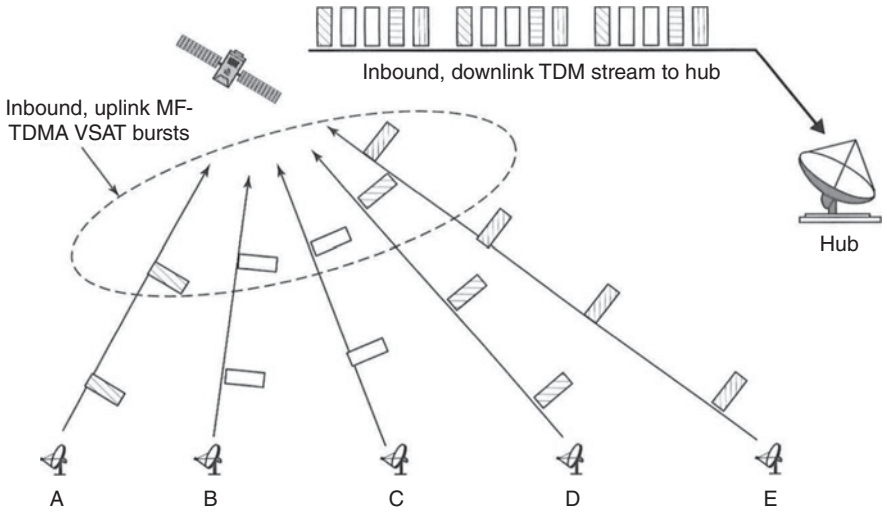


Figure 8.24 Example of a multifrequency TDMA (MF-TDMA) scheme. In this particular case, five VSAT terminals (A, B, C, D, and E) share the same frequency assignment; that is they all transmit at the same frequency. However, they each have a unique time slot in the TDMA frame when they transmit, so that they do not interfere with each other. The bursts from each VSAT are timed to arrive at the satellite in the correct sequence for onward transmission to the control earth station.

FDMA VSAT. If the average traffic for an individual VSAT is only one equivalent voice circuit (64 kbps), having to transmit at 5 Mbps, say, instead of 64 kbps can pose major difficulties. The VSAT transmit power must be increased by a factor of 5000/64 = 78 (or 18.9 dB) to maintain the same uplink CNR, since the earth station receiver must have a bandwidth that is wider by the same factor. This is not feasible because VSAT antennas have a broad beamwidth and are restricted in transmit power to avoid interference into adjacent satellites. VSAT economics and bandwidth efficiency tradeoffs have led to a hybrid TDMA-FDMA approach called MF-TDMA (multi-frequency time division multiple access), which is discussed in detail in Chapter 6. The concept of MF-TDMA is illustrated in Figure 8.24.

In the MF-TDMA example shown in Figure 8.24, each of the VSATs has to transmit at a burst rate that is approximately five times the normal single VSAT single-channel rate. If each VSAT transmits at a message data rate of 64 kbps and there are five VSATs sharing the same frequency, the minimum burst rate is 5×64 ps = 320 kbps. However, guard times have to be left in between each of the individual payloads within the TDMA frame to avoid overlaps due to incorrect clock timing. The satellite transponder plan looks very much like Figure 8.23 with the exception that each of the single channel, narrowband, inbound frequency slots would actually be allocated to a number of VSATs sharing a small TDMA frame transmitted at the same frequency. In the same way that the control earth station used in the FDMA scheme detects all of the individual inbound VSAT frequencies and then bundles the outbound return traffic into one wide-band TDM stream, the control earth station in the MF-TDMA scheme detects each of the inbound MF-TDMA VSAT signals and bundles the outbound traffic into a wide-band TDM stream. (See Examples 6.5.1 and 6.8.1 for a comparison of FDMA and TDMA in a VSAT network.)

8.5.6.3 CDMA

CDMA schemes were originally employed in VSAT systems for encryption purposes in military applications because they have a very low probability of intercept. However, unless there is a severe interference environment, as in some terrestrial microwave cellular radio systems, CDMA schemes are not normally selected because they are, in general, less bandwidth efficient than FDMA or TDMA. TDMA, in particular MF-TDMA for narrowband applications, is normally more bandwidth efficient than FDMA. Each VSAT operating in a CDMA mode transmits with the same frequency and at the same time, and relies on the orthogonal coding employed in a direct sequence, or frequency-hopping, spread spectrum application, to provide complete mutual separation of the individual communications signals. CDMA comes into its own for VSAT satellite applications when off-axis emissions from the earth terminals are likely to cause interference into another satellite. This is illustrated in Figure 8.25.

In Figure 8.25, if the wanted satellite (WSAT), unwanted satellite one (USAT 1), and unwanted satellite two (USAT 2), all use the same frequencies and polarizations, the use of CDMA prevents interference between the systems if orthogonal CDMA codes are used. There is, however, an increase in the noise received since each CDMA channel appears as a noise-like signal to every other CDMA channel. Thus, each additional CDMA signal will incrementally reduce the CNR of the channel. There is no hard and fast CNR threshold for CDMA links, as there is for TDMA and FDMA links, when the

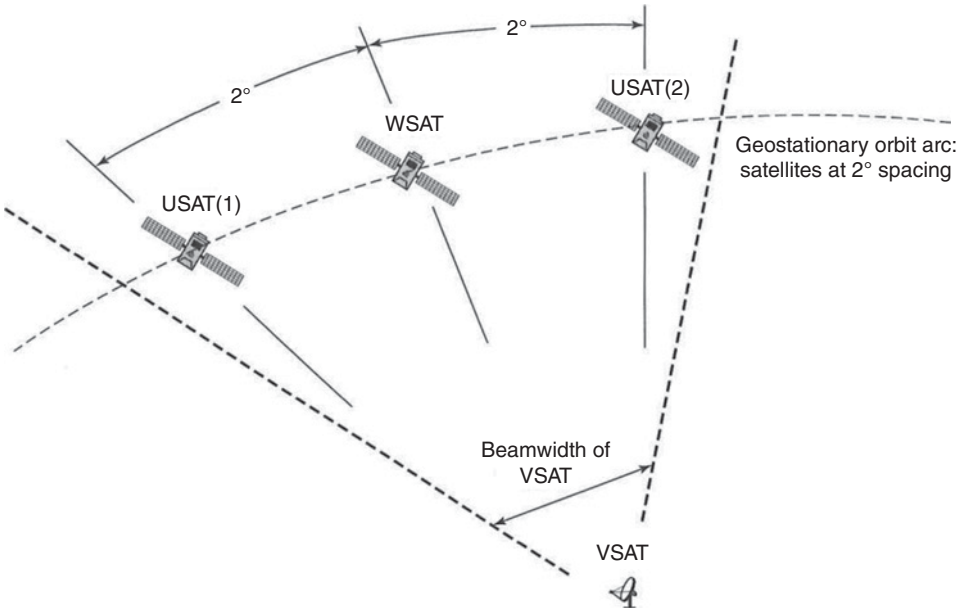


Figure 8.25 Illustration of how a VSAT can cause interference to other satellite systems. In this example, the VSAT is transmitting to a wanted satellite (WSAT) but, because the antenna of the VSAT is small, its beam will illuminate two other adjacent unwanted satellites (USATs) that are 2° away in the geostationary arc. In a like manner, signals from USAT (1) and USAT (2) can be received by the VSAT, thus causing the potential for interference if the frequencies and polarizations are the same. Off-axis emission is closely specified by the ITU-R and is a key element in uplink power control design. When LEO constellations are sharing the same frequency bands as GEO systems, the use of CDMA may confer some advantages for coordination purposes at the expense of system capacity.

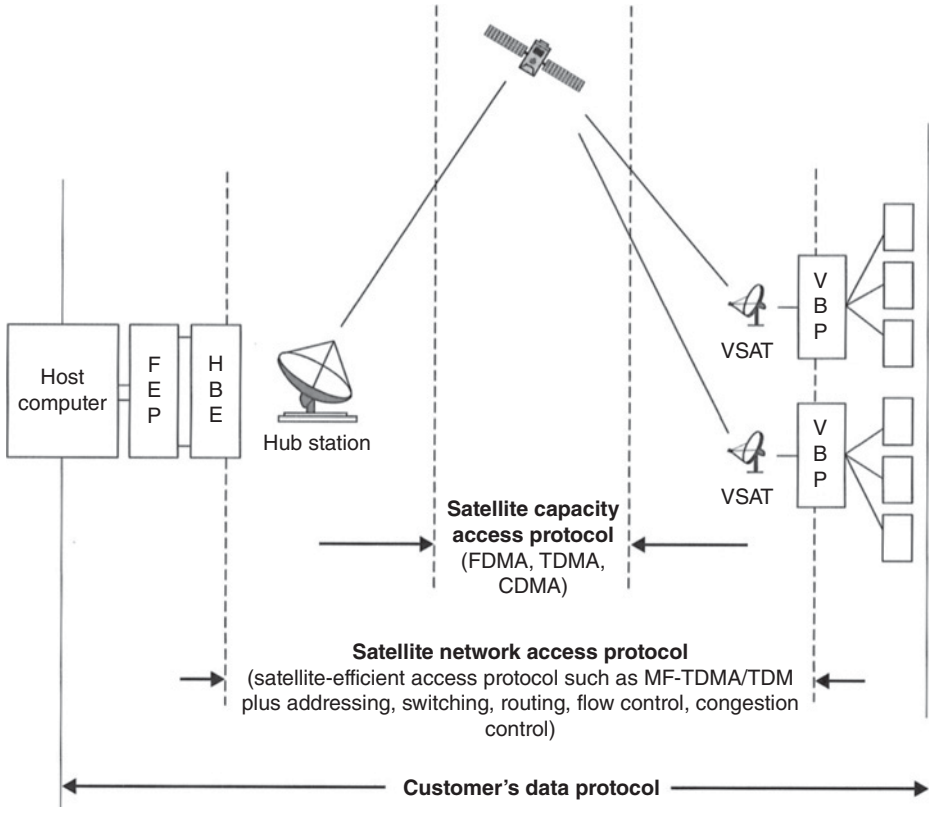


Figure 8.26 Illustration of the different layers of protocols used in VSAT networks (after Figure 3-1 in INTELSAT 1994). The host computer sends traffic for the VSAT network to the front-end processor (FEP) at the master control earth station of the VSAT network. The FEP passes the traffic to the hub baseband equipment (HBE) to be formatted for transfer over the satellite link via the selected satellite access protocol. The satellite then passes the outbound (or outroute) traffic on the downlink to the VSATs. The VSAT baseband processor (VBP) then extracts the relevant traffic for the user and forwards it after any necessary protocol conversion, etc. Source: Reproduced with permission of INTELSAT.

received signal characteristics become unusable. This soft threshold allows for some design flexibility, but care has to be taken to avoid undue errors induced by excess self-interference from the VSAT signals. Different satellite multiple access schemes can be used with a range of VSAT network access schemes, offering considerable flexibility to the system designer. Figure 8.26 illustrates this schematically.

In Figure 8.26, the satellite transponder is accessed using any of the three satellite multiple access modes: FDMA, TDMA, or CDMA. In addition to the multiple access scheme used, the VSAT network requires some form of Satellite Network Access Control. The Satellite Network Access Control ensures that the most efficient access protocol is used, for example, MF-FDMA/TDM or SCPC FDMA, for that particular satellite VSAT network. Not shown in this figure are the protocol emulators that act as the interface between the terrestrial and satellite networks, or the satellite access control that monitors switching, flow and congestion control, addressing, and so on (see Figure 8.20). Most of the signal formats for satellite multiple accesses are discussed in detail

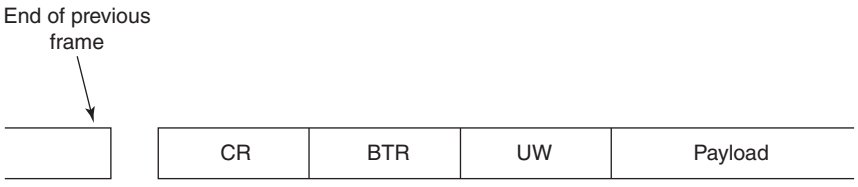


Figure 8.27 Generic sequence for the start of a burst from a VSAT inbound signal. When the burst is received at the control station, the first part of the packet enables the carrier recovery (CR) to occur, followed by the bit timing recovery (BTR). The unique word (UW) identifies the start of the payload in the new frame.

in Chapter 6. The subsection below reviews the generic case for digital satellite multiple access using TDMA schemes.

8.6 Signal Formats

The VSAT uplink signal, the inbound or inroute channel, in an MF-TDMA multiple access format must contain sufficient information for the intended receiver to acquire the carrier frequency, lock onto the incoming data so that the timing of the bit stream can be obtained, and then identify the start of the payload transmission. This generic procedure is shown in Figure 8.27. More details of various TDMA/packet signal formats can be found in Chapter 6.

8.6.1 Modulation, Coding, and Interference Issues

Modulation and channel coding are key considerations in determining the efficient and error-free transfer of information over a communications channel. They also have an impact on the potential for interference to another system and from another system. A modulation that has a large number of bits per symbol (e.g., 64 quadrature amplitude modulation [64 QAM]) will occupy a relatively small bandwidth but it will require relatively high amplifier linearity and a high CNR in the receiver. It is also more susceptible to interference than modulations with fewer bits per symbol. High-index modulations require significantly more margin than low-index modulations. In choosing the most appropriate modulation and channel coding for a VSAT system, ease of implementation is also a major factor since VSATs are very cost-sensitive. Faced with these trade-off decisions, the most common forms of modulation used in VSAT systems are QPSK and, when spectrum efficiency is less important, BPSK. In an ideal QPSK system with ideal SRRC filters and no channel coding, a value of E_b/N_0 of 10.6 dB will provide a BER of 10^{-6}, corresponding to a receiver overall CNR of 13.6 dB, ignoring any implementation margin. The CNR requirement can be significantly reduced if channel coding is applied. Coding aspects are discussed in Chapter 5. Only those aspects that touch on VSAT systems will be reviewed in the following section.

8.6.2 Channel Coding

Channel coding can take the form of a block code, a convolutional code, a turbo code, or LDPC code. Convolutional coding is a process where the encoding and decoding process

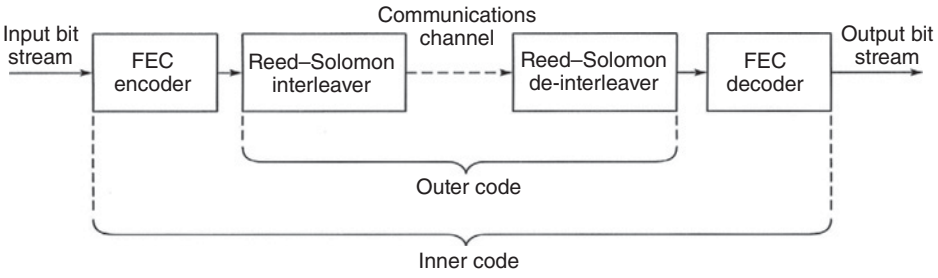


Figure 8.28 Schematic of the encoding and decoding process when an *inner* and *outer* code are applied to a telecommunications signal. The Reed-Solomon interleaved block code is applied after the FEC (either block or convolutional) on the encoding side. The reverse occurs on the decoding side. While it may look like the FEC code is outside the Reed-Solomon code, it is the time they are applied that counts. Since the FEC is applied first in the encoding process, it is then wrapped by the Reed-Solomon code, which becomes the outer wrapping of the doubly coded signal.

is applied to a group of bits in sequence rather than a bit at a time, as in a block code. The number of bits in the encoding sequence, *k*, is called the constraint length of the convolutional code. In the decoding process, *k* bits are used to evaluate the value of each bit transmitted. Since the encoding process is applied to the signal prior to transmission and is used to detect and correct for bit errors, it is called a forward error correction (FEC) code. In a like manner, a block FEC code is applied to the channel prior to transmission. Convolutional and block codes can be used together on a channel. One example is a channel that first has an inner convolution code applied to the bit sequence and then has an outer interleaved code such as a Reed-Solomon code applied. Reed-Solomon codes combine good error detection capability with high code rates. This form of *concatenated* coding is used extensively in many communications systems (see, for example, the coding used in DBS television discussed in Chapter 10), since the interleaved coding will counter burst errors while the convolutional FEC coding will counter individual bit errors. DBS television is one example of such a coding approach, and the recording of music on CDs is another. The encoding and decoding procedure is illustrated schematically in Figure 8.28. Details of turbo and LDPC codes are in Chapter 5.

For VSAT systems that have small traffic streams, excess processing delay can add significantly to the end-to-end link delay. This is very important for GEO systems and for LEO/MEO systems with satellites that have large onboard processing capabilities. The processing delay due to first interleaving a signal, then de-interleaving it adds a fixed amount of overhead, as well as requiring buffering at both ends of the transmission link. For this reason, Reed-Solomon outer codes are not normally added to signals that have information rates below about 256 kbps, even though the lower E_b/N_0value for a given BER performance is so significant (see Figure 10.9, repeated here as Figure 8.29). For links that have no real requirement for instant response times and multimedia interactivity, but require the best BER performance for a given E_b/N_o (typical of most internet links), Reed-Solomon codes are a very practical way of reducing the power requirements for a given link and BER specification.

8.6.3 Interference

Interference between systems operating with similar characteristics (i.e., frequency bands, polarizations, and services) is usually the subject of intense debate, particularly

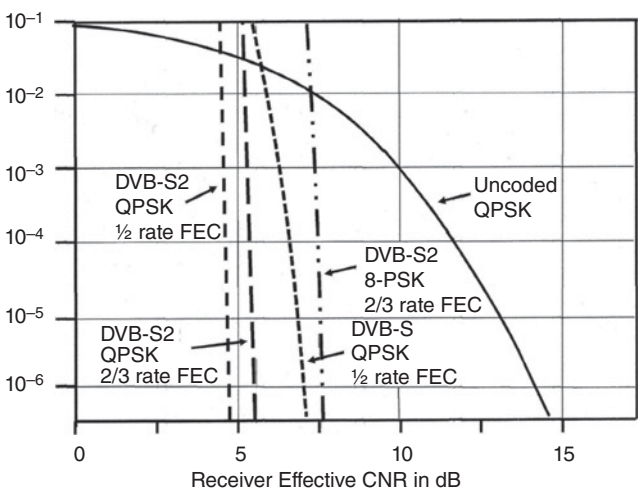


Figure 8.29 BER vs. E_b/N_o performance of coherent QPSK for various types of codes. A 1.8 dB implementation margin is included in these results. (Figure 10.9 in this text).

when a new system seeks to operate close to an existing system, in terms of orbital separation or antenna beam directions. The interaction between operators seeking to ensure that no harmful interference is caused by, or to, their respective systems is called coordination. The coordination process is the subject of extensive regulation by the ITU and national frequency management authorities (e.g., the FCC in the United States). The key aspect in such coordination exercises lies in determining the power radiated by the interfering station in the direction of the interfered with station. The calculation of the received interference power will have four elements:

- The output power of the interfering station's transmit amplifier
- The transmit gain of the interfering station's antenna in the direction of the interfered with station
- The receive gain of the interfered with station's antenna in the direction of the interfering transmissions
- The path loss between the two stations.

The interference geometry is illustrated in Figure 8.30.

Recommendation ITU-R S.728 (ITU-R 2017) mandates the maximum permissible levels of off-axis EIRP density from a VSAT transmitting in the 14 GHz band (Ku-band). The relevant part of the Recommendation is abstracted below.

"VSAT earth stations operating in the 14 GHz frequency band used by the fixed-satellite service must be designed in such a manner that at any angle φ specified below, off the main lobe axis of an earth station antenna, the maximum EIRP in any direction within 3° of the geostationary-satellite orbit should not exceed the values in Table 8.6: In addition, the cross-polarized component in any direction φ degrees from the antenna main-lobe axis should not exceed the values in Table 8.7."

There are two important notes contained in the footnote to the above Recommendation. Since the Recommendation was developed for 3° satellite spacing in GEO, the first note indicates that the off-axis limits may need to be reduced by up to 8 dB where the satellite spacing is 2°. The second note pertains to CDMA VSAT systems. When there are N VSATs expected to transmit simultaneously on the same frequency, the maximum permitted EIRP values should be decreased by 10 log N.

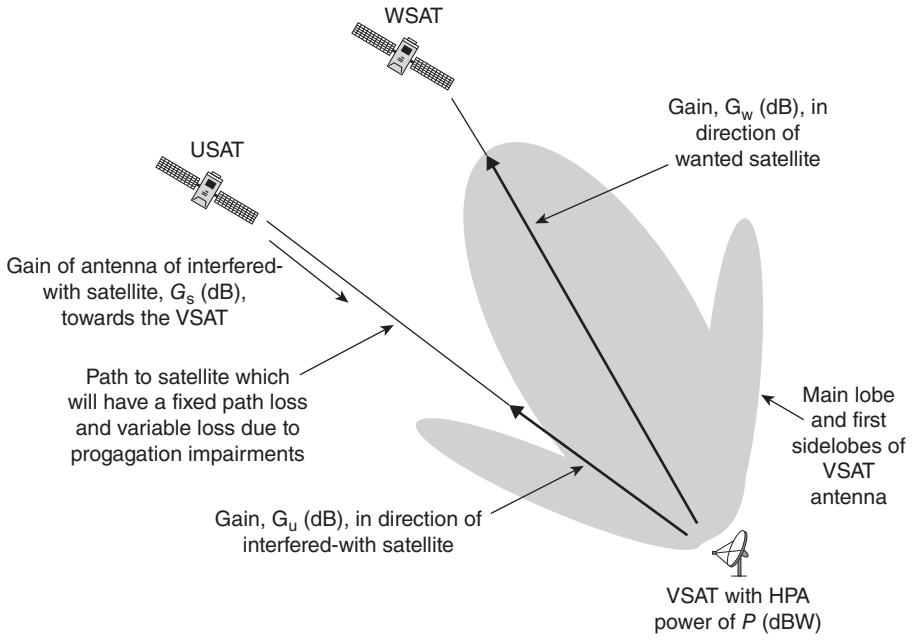


Figure 8.30 Illustration of the interference geometry between a VSAT and a satellite of another system. The EIRP of the VSAT toward the interfered with satellite [P (dBW) + G_u (dB)] is the interference power from the VSAT into the interfered with satellite. To develop the interference link budget, the gain of the interfered with satellite in the direction of the VSAT, G_s (dB), is used, plus any additional effects along the path (such as site shielding, if used, expected rain effects for the given time percentages, etc.).

The rapid increase expected in satellite delivery of internet-like traffic direct to homes and offices has led to a multitude of proposed new constellations of satellite systems, with the result that interference aspects have received a lot of study within the ITU and ETSI (European Telecommunications Standardization Institute). The off-axis limits have been tightened under a new proposal before ETSI, but which have yet to be incorporated in ITU-R S.728 above. The new proposals are directed toward the use of Ka-band satellites with VSAT apertures well below 1 m. The proposed ETSI limits are given below (see Table 8.8).

The maximum EIRP in any 40 kHz band within the nominal bandwidth of the co-polarized component in any direction φ degrees from the antenna main beam axis shall

Table 8.6 ITU(R) specification for VSAT transmitting stations in Ku-band co-polar

Angle off-axis	Maximum EIRP in any 40 kHz-band
$2.5° \le \varphi \le 7°$	$33 - 25 \log \varphi$ dBW
$7° < \varphi \le 9.2°$	12 dBW
$9.2° < \varphi \le 48°$	$36 - 25 \log \varphi$ dBW
$\Phi > 48°$	−6 dBW

Table 8.7 ITU(R) specification for VSAT transmitting stations in Ku-band, cross-polar

Angle off-axis	Maximum EIRP in any 40 kHz-band
$2.5° \leq \varphi \leq 7°$	$23 - 25 \log \varphi$ dBW
$7° < \varphi \leq 9.2°$	2 dBW

not exceed the limits given in Table 8.8 *under clear-sky conditions, within* $\pm 3°$ *of the geostationary orbit plane*.

For CDMA systems, N is the number of VSAT earth stations transmitting simultaneously on the same frequency. For TDMA and FDMA systems, N = 1. The proposed recommendation by ETSI is for satellites that are spaced 2° apart in GEO. The values in both cases (CDMA and TDMA/FDMA) may be relaxed for directions more than 3° away from the geostationary plane since VSAT antenna patterns are normally optimized for the GEO arc. During rain fade conditions, the values in the above equations may be exceeded through the application of uplink power control (ULPC) at affected stations, in order to overcome the rain attenuation. The effective off-axis emission levels received by adjacent satellite systems are not expected to vary substantially from that during the clear-sky condition if the ULPC system is designed and operated properly. ULPC systems have the potential to be very inaccurate (Allnutt 2011), particularly those that rely on open-loop control (Castanet et al. 1998; Vasseur et al. 1998). Carefully controlled experiments at Ku-band (Comsat Labs 1997) and Ka-band (Dissanayake 1997) have shown that there is an irreducible error on the order of ±1.2 dB for open-loop ULPC and, even with closed-loop ULPC, where the power level is measured at the satellite, there can be time-delay constraints that will limit the accuracy (Sweeney and Bostian 1999).

8.6.4 Transmitters and Receivers

Historically, large earth stations are assembled as discrete elements. On the receive side, the antenna and feed components are connected by waveguide to the front-end low noise amplifier (LNA). Behind the LNA, a mixer/down-converter changes the signal from RF to an IF. After filtering and amplification, the IF signal is demodulated, demultiplexed, and decoded, and the baseband signal forwarded to the user. The transmit side is the mirror image of the receive side with the signal input at baseband and the output at RF, with the LNA receiver replaced by a high power amplifier (HPA) transmitter.

Table 8.8 Limits on interference to adjacent satellites in GEO proposed by ETSI

Angle off-axis	Maximum EIRP in any 40 kHz-band
$1.8° \leq \varphi < 7.0°$	$19 - 25 \log \varphi - 10 \log N$ dBW
$7.0° < \varphi \leq 9.2°$	$-2 - 10 \log N$ dBW
$9.2° < \varphi \leq 48°$	$22 - 25 \log \varphi - 10 \log N$ dBW
$\Phi > 48°$	$-10 - 10 \log N$ dBW

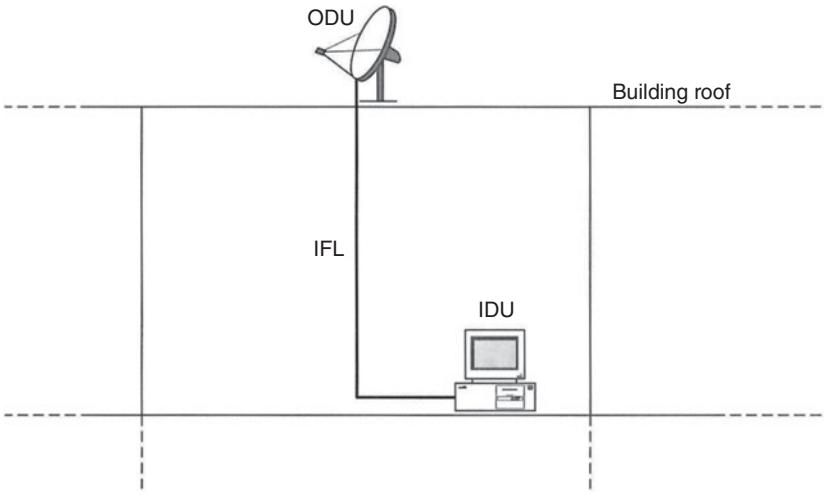


Figure 8.31 Schematic of the typical location of VSAT component parts. The VSAT outdoor unit (ODU) is located where it will have a clear line of sight to the satellite and is free from casual blockage by people and/or equipment moving in front of it. The interfacility link (IFL) carries the electronic signal between the ODU and the indoor unit (IDU) as well as power cables for the ODU and control signals from the IDU. The IDU is normally housed in a desktop computer at the user's workstation and consists of the baseband processor units and interface equipment (e.g., computer screen and keyboard). The IDU will also house the modem and multiplexer/demultiplexer (mux/demux) units if these are not already housed in the ODU.

This design of earth station is typical of a control earth station used in a VSAT network. Much of this discrete component design has changed with the introduction of digital receivers and the need to develop cheap, mass-produced VSAT terminals.

With the flat panel phased array designs being formulated for user terminals in NGSO systems, software radio concepts, first introduced into mobile radio systems, are being implemented in fixed earth station designs. As noted in Chapter 10, the receivers for DVBS2 are software controlled, by both the gateway station and the receive/transmit unit at the VSAT.

VSAT earth stations have historically consisted of two basic components: an outdoor unit (ODU) and an indoor unit (IDU). This is illustrated in Figure 8.31. The ODU and IDU are broken down further in Figure 8.32.

8.6.5 VSAT Link Design Example

The VSAT network discussed in Section 6.11 consisted of 198 Ku-band VSAT earth stations sharing one inbound and one outbound transponder on a GEO satellite. The transmit data bit rate for the VSAT stations was 64 kbps. The average outbound data rate to each VSAT station was 64 kbps. The inbound data link operated in FDMA-SCPC using demand access, QPSK modulation with $\alpha = 0.25$ SRRC filters, and half rate forward error correction. The outbound data link used a single continuous TDM stream, QPSK modulation, a 16 bit header and a 4 bit cyclic redundancy check (CRC) in each packet. Two additional channels were allocated as common signaling channels (CSCs) giving a total of 200 FDMA channels in the transponder. In this example, the number of VSAT

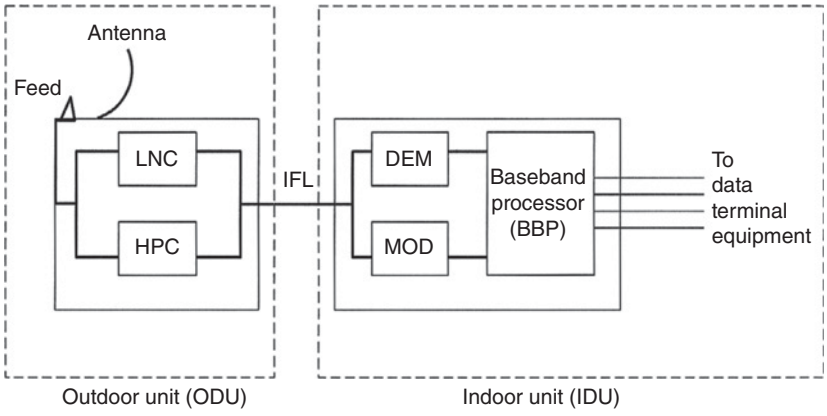


Figure 8.32 Schematic of the typical configuration of a VSAT earth station. Source: After Figure 4.1.1 of VSAT, ITU 1994. Reproduced with permission of ITU-R. The low noise converter (LNC) takes the received RF signal and, after amplification, mixes it down to IF for passing over the interfacility link (IFL) to the IDU. In the IDU, the demodulator extracts the information signal from the carrier and passes it at baseband to the baseband processor. The data terminal equipment then provides the application layer for the user to interact with the information input. On the transmit operation, the user inputs the data via the terminal equipment to the baseband processor and from there to the modulator. The modulator places the information on the carrier at IF and this is sent via the interfacility link to the high power converter (HPC) for upconversion to RF, amplification, and transmission via the antenna to the satellite.

stations is reduced to 98, but in all other respects the parameters of the VSAT network remain unchanged.

A VSAT network consists of 98 Ku-band VSAT earth stations sharing one inbound and one outbound transponder on a GEO satellite. The transmit data bit rate for the VSAT stations is 64 kbps. The average outbound data rate to each VSAT station is 64 kbps. The inbound data link operates in FDMA-SCPC using demand access, QPSK modulation with $\alpha = 0.25$ SRRC filters, and half rate forward error correction. The outbound data link uses a single TDM stream, QPSK modulation with half rate FEC, a 16 bit header and a 4 bit CRC in each packet. Two additional channels are allocated as common signaling channels giving a total of 100 FDMA channels in the inbound transponder. Data from the VSAT stations are sent in 10 ms packets that have 36 bit headers and a 4 bit CRC.

a. Determine the traffic bit rate on the inbound links using SCPC-FDMA.
   The outbound link is symmetrical and employs a single TDM bit stream that transmits packets with the same structure as used on inbound links and delivers traffic data to each VSAT at the same rate as the inbound link.
b. Find the inbound traffic data rate for one VSAT station and the TDM transmission rate from the gateway station.
c. What is the traffic data rate on the outbound link and the bandwidths of the VSAT and gateway station receivers?
d. Suggest a frame and packet size for the TDM link, a SCPC-FDMA frequency plan, and a demand access method.

8.6.5.1 Inbound Link: VSAT to Gateway

The traffic data transmitted from the VSAT station is a half rate FEC bit stream at 128 kbps carrying 64 kbps of data. With QPSK modulation the symbol rate is $R_s = 64$ ksps. The link has $\alpha = 0.25$ SRRC filters, so the occupied bandwidth of the QPSK signal is 80 kHz. The IF receivers in the gateway earth station have filters with a noise bandwidth of 64 kHz (equal to the symbol rate of the signal) and an occupied bandwidth of 80 kHz. Allowing a 20 kHz guard band between RF channels requires a carrier-to-carrier spacing of 100 kHz per channel. With a total of 100 channels, the bandwidth required in the transponder for the inbound channels is 10 MHz. Two of the channels are designated as CSCs, so there are 98 communication channels.

The bit rate on an inbound channel is 64 kbps with 10 ms packets. Each packet delivers 640 bits, of which 40 are header and CRC. Hence the traffic data in each packet is 600 bits and the traffic data rate is 60 kbps. The outbound link uses the same packet structure and must send packets to 98 stations as a continuous TDM stream. The minimum outbound transmission rate is 6.272 Mbps of which 5.88 Mbps is traffic data.

In many VSAT networks the individual stations do not have a continuous supply of data. For example, a company that owns hundreds of large stores uses a VSAT star network to connect each store to a gateway station at the company headquarters. Cash registers in each store connect via the VSAT network to the gateway station to process every transaction and record the details for accounting and stock control. At an individual store there may be quiet times when there are no transactions taking place, and the VSAT inbound link has no data to send. This scenario favors a SCPC-FDMA-DAMA scheme using a CSC to obtain access to the satellite when a VSAT station has data to send. If we assume that each VSAT has data to transmit once every 12 seconds, and it takes one second to establish the connection and one second to transmit the packets, a total of two seconds of satellite transmission is used every 12 seconds, giving a VSAT station loading of 16.6%. This means that (in theory) six VSAT stations could share the same RF channel. If lower latency is important, shorter packets containing fewer data bits can be used with more frequent transmissions to the satellite.

In practice, it is impossible to load the channels to 100% of their capacity because data arrives at random time intervals causing temporary overload when a large volume of data arrives at the same time. If we assume a 66% load factor for the inbound link, we can share one inbound channel between four VSAT stations. Demand access has clearly achieved considerable savings in bandwidth and power in this case.

8.6.5.2 Outbound Link: Gateway to VSAT

The gateway station transmits a TDM frame consisting of 98 sequential packets addressed to the VSAT stations with additional framing bits. Let's assume a one second frame with an additional 128 k framing bits, equivalent to two 64 kbps channels. If we apply half rate FEC to the outbound data stream and use QPSK modulation, we will transmit at 64 ksps per station. With 100 64 kbps equivalent channels, the average outbound symbol rate is 6.4 Msps, and the outbound data rate is 6.4 Mbps because we are using QPSK modulation with half rate FEC. The occupied bandwidth of the signal with $\alpha = 0.25$ SRRC filters is 8 MHz, and the VSAT receiver noise bandwidth will be 6.4 MHz.

If there are no data bits to be delivered to a given station, only the overhead portion of the packet, 40 bits, needs to be transmitted. That allows other stations to use the spare time in the frame to send additional data at a much higher than average rate.

Demand access is most valuable when the traffic mix changes a great deal. The multiple access system described here was designed to meet the needs of the average data rates transmitted on the inbound and outbound links. If many of the stations are inactive, the other stations can have increased data rates. For example, suppose only 49 of the VSAT stations are active. Each VSAT station can transmit two packets per frame doubling its traffic data rate to 120 kbps.

Alternatively, the VSAT station could transmit two carriers. The limitation on inbound data rate is likely to be VSAT EIRP and the resulting uplink CNR ratio in the transponder. SCPC-FDMA does not offer as much flexibility to change data rates as TDM.

TDM frames that offer variable packet length can easily accommodate a widely changing mixture of data rates delivered to each VSAT station. A field within the packet header tells the VSAT station how many data bits are in the packet, allowing great variability.

8.7 System Aspects

8.7.1 Visibility Arc

In Section 9.3 of the NGSO chapter, details are given of various orbit coverage capabilities. The coverage from an individual satellite in a given orbit will be determined by the minimum elevation angle and the orbital velocity of the satellite, both of which will determine the time over coverage. The lower the orbital altitude, the smaller the time any given satellite can operate to a given point on the earth's surface. This aspect is covered in some detail in Chapter 9 on NGSO systems, and an example calculation for the time available for a LEO satellite to communicate with a user on the surface of the earth is given in Section 8.2.7 using Figures 8.5, 8.6 and 8.7. Another question concerning the coverage of a LEO is given below in Example 8.6.

Example 8.6

Question

a. To be able to successfully receive a signal from a satellite, it is usual for the receiver on the ground to be within the 3 dB footprint of the satellite antenna. If a user on the surface of the earth is able to successfully receive a signal from a satellite that is in a circular orbit at an altitude of 1000 km above the surface of the earth when the elevation angle from the user to the satellite is 10°, what is the half-power beamwidth of the satellite transmitter?
b. If the same satellite above has a half-power beamwidth of 20°, what is the total coverage arc on the surface of the earth?

Answer

a. From Figure 8.5, distance SZ is the orbital altitude of the satellite = 1000 km. Angle θ is the elevation angle from the user to the satellite = 10°. Distance SC from the satellite to the center of the earth = 1000 + 6370 = 7370 km.

$$\text{angle } SEC = \text{elevation angle} + 90° = 10° + 90° = 100° \tag{8.55}$$

The angle $2 \times \delta$ is the half power beamwidth of the satellite transmitter. Using the law of sines

$$\left(\frac{\sin\delta}{r_e}\right) = \left(\frac{\sin 100}{SC}\right) \Rightarrow \sin\delta = 6370 \times \left(\frac{\sin 100}{7370}\right) = 0.8512 \tag{8.56}$$

and so

$$\delta = 58.3407^\circ \tag{8.57}$$

and the half power beamwidth is given by

$$2 \times \delta = 2 \times 58.3407^\circ = 116.681^\circ \approx 116.7^\circ \tag{8.58}$$

b. If the half-power beamwidth is reduced to 20°, then from Figure 8.7, the angle $\delta = 10^\circ$. Since δ is small, we can approximate the arc EZ by

$$\begin{aligned}\text{arcEZ} &= SZ \times \left(10 \times \left(\frac{\pi}{180}\right)\right) = 1000 \times \left(10 \times \left(\frac{\pi}{180}\right)\right) \\ &= 174.5329 \sim 174.54\,\text{km}\end{aligned} \tag{8.59}$$

The total coverage arc is therefore double this number = 349 km.

8.7.2 Transmit and Received Power Aspects

Equation (9.10) from Chapter 9 on NGSO systems is repeated below as Eq. (8.60) for a terminal with a receiving antenna gain $G_r = 1$. The received power at the mobile earth station is given by $P_r = F \times A$, where F is the flux density in W/m^2 and A is the coverage area of the satellite transmitting antenna in m^2, hence

$$P_r = (P_t G_t \lambda^2)/4\pi A \text{ W} \tag{8.60}$$

Thus the received power at the mobile with an omnidirectional antenna increases as the square of the wavelength, or decreases as the square of the frequency. The lower the RF frequency, the greater the received power for any given coverage zone. By reciprocity, the same result will apply when the mobile terminal transmits with an omnidirectional antenna. It therefore makes sense for mobile systems, which are forced to use omnidirectional antennas so as to avoid having to steer a directional antenna, to use the lowest possible RF frequency.

8.8 Time Over Coverage

This aspect was discussed in Sections 8.2.6 and 8.2.7, and it is clear that the lower the altitude of a satellite is, the less time it has either to observe what is below it or to communicate with a fixed terminal on the surface of the earth. The economic success of the DBS systems was that relatively cheap, non-steering terminals could be set up very simply. This will not be the case for the NGSO systems discussed in Chapter 9 as the high-throughput satellites in their various orbits will need to be accurately tracked. The ability to accurately track a satellite is also a problem for smallsat users who want to maximize the received signal from their spacecraft: an omnidirectional antenna can only receive relatively low data rates. Phased array antennas on the ground, if they can be built economically, are one approach to achieving higher data rates than an omnidirectional antenna. One phased array antenna proposal for Ka-band mobile

systems has been developed and tested but the design did not appear to be low cost (Herranz-Herruzo et al. 2018). The key for these flat panel antennas will be not be just their performance and cost, but their durability. Antennas operating with satellites are usually fully exposed to the weather 24/7 with little protection. A flexible, phased array antenna that can operate with a variety of spacecraft will be in demand and it is likely that, when a suitable antenna is developed, compatible to the many NGSO systems, economies of scale will significantly reduce the cost of the antenna to the end user. However, with the plethora of smallsat constellations launched, and proposed, there is an increasing risk of leaving a vast amount of orbital debris in LEO over the next decade.

8.9 Orbital Debris

The common perception is that the earth is surrounded by thousands of satellites that are creating a hazardous blanket for manned and unmanned spacecraft to navigate through. But this is not the actual case: operational satellites make up only 5% of what is orbiting the earth (Space Junk 2017). The rest – referred to as *orbital debris* – is made up of spent rocket stages, fragments of metal from ruptured fuel tanks, nose cone segments, and the like (Space Debris 2017). In an attempt to ameliorate the fragmenting aspect of LEO satellites, SpaceX is designing spacecraft with material that will burn up readily on re-entry, housing sensitive electronics so that minor collisions will not destroy the satellite's ability to respond to commands, and ensuring that the satellite has the capability to maneuver away from potential collisions with other satellites or large debris (AW&ST 2017b). The likelihood of two space craft colliding is greatly reduced by having Conjunction Data Messages (ISO 2018) issued by the Joint Space Operations Center (Joint Ops 2018). There are estimated to be 750 000 pieces of debris in space larger than a marble, the majority of it in LEO (Gugliotta 2017). Europe has proposed an *e.deorbit* mission to capture large satellites or pieces of debris that are no longer functioning and deorbit them (Gugliotta 2017). There have been other attempts to get a handle on the debris problem in space. A team of students from four universities in Virginia will have three smallsats ejected from the ISS (Satellitetoday 2018f). One the smallsats, called *Aeternitas* has a drag break and is intended to re-enter faster than the other two smallsats *Libertas* and *Ceres*, which are expected to be in orbit for two years. SpaceX launched a UK University of Surrey payload during its fourteenth ISS resupply mission (Satellitetoday 2018g; Space Junk 2018). Four experiments are planned. The first two are capture experiments. In the first, a net will be deployed to capture the target, while in the second, an attempt will be made to harpoon a target. In the third experiment, a small Cubesat will attempt to rendezvous with a target object using a lidar. Having rendezvoused with the target, the Cubesat will deploy a drag sail to cause a faster de-orbiting.

Most satellites in LEO and MEO, however, are not moved from their orbits at end-of-life. New satellites must therefore avoid the orbital debris, both when they are launched and while in orbit. To do this successfully requires an accurate knowledge of not only what is up there at any instant in time, but also the precise location and orbital ephemeris of every potentially hazardous piece of orbital debris. Orbital debris need to be tracked.

8.9.1 Tracking Aspects

The US Air Force operates a radar known as Space Fence (2017), which detects objects in orbit around the earth. It does not track them, merely detecting objects as they cross

an *electronic fence* and timing their orbit to determine if there are any changes, or new objects in orbit. As for the civilian space agency, NASA has an Orbital Debris Program Office (Orbital Debris 2017) that is becoming increasingly active. In the commercial world, a company called LeoLabs has developed a phased array system to track orbital debris (LeoLabs 2017). They plan to locate their arrays in many locations around the United States to provide detailed information on all orbiting artifacts.

8.9.2 End-Of-Operational-Life Considerations

Operators of geostationary communications satellites realized that they needed to move their spacecraft away from the geostationary orbit so that newer satellites could work safely. However, only about one in four satellites were safely moved through 2004 (Graveyard Orbit 2017), leading the FCC to mandate that any geostationary satellite providing services to the United States must be capable of being moved to a higher orbit at end-of-life. This orbit is about 300 km further out than the geostationary orbit, and because satellites moved there are essentially dead, it is called the graveyard orbit. For satellites in lower orbits than GEO, there is no really safe orbit that the satellite can be moved to at the close of operations. One proposal suggested three options for dealing with satellites that were no longer working (De-Orbit 2017):

"*Option 1*. Move the spacecraft to an orbit where drag will de-orbit it within a relatively short time (e.g., 25 years)
Option 2. Direct retrieval and de-orbiting
Option 3. Maneuver the spacecraft to an orbital region where it is less likely to interfere with future space operations"

Most LEO satellites in circular orbit close to the earth will naturally decay within less than 25 years, but even this may be too long. For this reasons, some satellite manufacturers are designing their satellites to have deployable panels or sails that will increase the drag when deployed and so speed up orbital decay.

8.10 Summary

Disruptive technologies occur all the time, and telecommunications is no exception. To stand still is to lose. From VHS to DVD took less than a generation: 21 years to be exact. Within 5 years, the ubiquitous thumb drive was here, and 10 years later could store as much as a DVD. The 8-in. floppy disk, the 5.25 storage disk, and the 3.5 in. storage medium went the way of the dodo in less than 30 years. Most lap tops sold in 2018 no longer contain an internal disc drive. Satellite communications also saw disruptive technologies, but initially they seemed to occur much more slowly. Geostationary satellites literally ruled the skies for more than a generation. The introduction of fiber optic cables across all three major oceans, and encircling most continents, saw a shift in the services provided from geostationary satellites from voice communications to direct television and internet backhaul. But the slow reaction time of geostationary satellites – the identification of a service, the design and building of the satellite, and then the launch – was just too long to resist the surge of new entrants into the market that had quicker reaction times.

Chapter 9 details the incursion of NGSO constellations into the geostationary satellite services, but perhaps the biggest revolution has been in the sheer *size* of the satellites now being designed and launched for new service offerings. The smallsat revolution has transformed not just the educational scene, but almost every other aspect of telecommunications. The time to design, build, and launch a small satellite can be timed in months and not years. The fact that a smallsat is relatively cheap to put into orbit means that the standard testing of the big GEO spacecraft (design review, critical design review, prototype, engineering model, vibration analysis, thermal vacuum runs, etc.) is not necessary: if you lose a smallsat, there are plenty more where that came from. And the launches are also relatively cheap, compared with the GEO spacecraft. The steady microminiaturization of not just components, but complete internal systems (such as attitude sensing), means that smallsats can take on ever-increasing roles, some of which require quite large throughput of signals. It may well be that NGSO constellations and smallsat services will merge with high throughput satellite systems so that a seamless service can be provided to the end user anywhere on the globe with any bandwidth, delay constraint, and pricing. These are exciting times (again!) for satellite communications.

Exercises

8.1 A cubesat satellite has an orbital altitude of 500 m and is used in store and forward mode to collect data from uplink terminals around the world and download the results to a receiving station in the United States whenever the satellite is within the visibility cone of the receiving station. The visibility cone for the receiving station is defined as a cone with its axis vertical and a maximum path length of 1000 km at the outer edge of the cone. The receiving earth station has an omnidirectional antenna with a gain of 0 dB at the edge of its visibility cone.
To begin with, assume that the satellite transmits a QPSK signal with half rate error correction coding at 1 ksps and an RF transmit power of 1 W. The earth station has a receiver with a noise bandwidth of 1 kHz and a system noise temperature of 300 K. The implementation margin of the link is 1.0 dB.

a. Using the geometry in Figure 8.5, calculate the coverage angle for the satellite's downlink antenna when the 3 dB beamwidth of antenna is set to have a maximum slant length of 1000 km (distance d in Figure 8.5). (Hint, use the cosine law for a triangle, or find a triangle solver on the web.)

b. Estimate the on-axis gain of the satellite antenna using the approximate formula (not in dB)

$$G = \frac{30{,}000}{3 \text{ dB } \textit{beamwidth squared}}$$

Hence determine the gain of the satellite antenna for a receiving station at a range of 1000 km.

c. Create link budgets for the downlink from the satellite to the earth station at RF frequencies of 144.0, 460.0, and 1550.0 MHz. Allow 0.5 dB for miscellaneous losses.

d. Calculate the CNR at the receiving earth station for each of the downlink frequencies.

e. Calculate the maximum bit rate on the downlink at each of the downlink RF frequencies if the threshold CNR is set at 6.0 dB.

f. A practical system needs a significant CNR margin above the threshold to ensure successful reception of data on every transmission from the satellite. If the margin is set to 7 dB above threshold, what is the practical downlink bit rate at each RF frequency?

8.2 Repeat the analysis in parts (c) through (f) of Question 8.1 for an earth station with a tracking antenna with an on-axis gain of 15 dB that can keep the satellite the satellite within the −1 dB contour of the earth station antenna pattern.

8.3 Calculate the orbital period of the satellite in Question 8.1. The satellite in an inclined prograde orbit at an inclination angle of 50°.

a. Calculate the distance that a point on the equator has moved due to the earth's rotation relative to the point at which the satellite crosses the equator in a south to north direction after one orbit of the satellite. Is it possible for an earth station on the equator to connect to the satellite on two successive passes? Give reasons for your answer.

b. If the satellite passes directly over the earth station located on the equator at 00 : 00 : 00 on Day 1, when will it next be visible to the earth station, within the station's visibility zone? Remember that the satellite can be seen on a N-S pass as well as on a S-N pass. Give your answer in day, hour, and minute format.

c. Is there a location in the continental United States (the lower 48 states) where the satellite could be seen on two successive passes?

8.4 The satellite in Question #1 has an orbital altitude of 500 km. An earth station can view the satellite whenever it has an elevation angle above 10°.

a. What is the longest time for which the satellite is in view of the earth station?

b. What is the shortest time for which the satellite is in view of the earth station?

c. If the satellite's transmitting antenna has a beamwidth 132°, what is the diameter of the circle on the surface of the earth that is illuminated by the antenna beam? Give your answer to the nearest 10 km.

d. Continuous coverage of the earth by multiple satellites requires a constellation of p satellites per orbital plane and q orbital planes. Find the minimum values of p and q that meet the coverage requirement. How many satellites would you propose to include in the constellation to provide a practical system with 24/7 capability?

8.5 GPS satellites are an example of a low throughput system. The data rate delivered to a user terminal is 50 bps, the rate of the navigation message transmitted by all GPS satellites. Although a C/A code GPS receiver is receiving DSSS chips at 1.023 Mcps, it takes 20 code sequences to deliver one navigation data bit. The L1 frequency of the C/A code is 1575.42 MHz, and the maximum distance between a GPS satellite and a user terminal is set at 26 800 km. The GPS satellite antenna gain is 13.0 dB for a user at maximum range and the user terminal antenna gain is 0 dB. The L1 signal is transmitted at a nominal power level of 10 W.

a. Create a link budget for a user terminal at maximum range from the GPS satellite and find the CNR in a receiver noise bandwidth of 50 Hz with 2.0 dB losses from all sources.
b. A user terminal has a system noise temperature of 300 K. The L1 GPS signal is BPSK modulated and not FEC encoded, giving a threshold CNR value of 10.6 dB. Find the link margin above threshold.

8.6 Why has the US FCC introduced a ruling that the owners of constellations of LEO satellites must include a provision for the satellites to be de-orbited at the end of their lifetime? Describe two ways in which this can be achieved for LEO satellites. Is one way better than another for a satellite in a 1200 km orbit?

References

Abrahamson, A. (ed.) (1993). *Multiple Access Communications: Foundations for Emerging Technologies*. New York, NY: IEEE Press.

Allnutt, J.E. (2011). *Satellite-to-Ground Radiowave Propagation*, 2e. London, UK: The Institution of Engineering and Technology.

Amateur Radio (2018). http://www.iaru.org/satellite.html (accessed 15 June 2018).

Amateur Radio UK (2018). www.amsat.org.uk/iaru (accessed 15 June 2018).

Auto ID (2017). https://en.wikipedia.org/wiki/Automatic_identification_system (accessed 16 June 2017).

AW&ST (2017a). Aviation Week and Space Technology, May 19–June 11, 2017, pp. 58–59.

AW&ST (2017b). Both Ways, a commentary in Aviation Week and Space Technology, October 30–November 12, 2017, page 19.

AW&ST (2018a). Aviation Week and Space Technology, June 11, 2018 https://outlook.office.com/owa/?realm=gmu&vd=mso365#x_3 (accessed 12 June 2018).

AW&ST (2018b). http://aviationweek.com/program-management-corner/startup-bringing-plasma-propulsion-technology-smallsat-crowd (accessed 2 March 2018).

Babuscia, A., Van de Loo, M., Wei, Q.J. et al. (2014). Inflatable antenna for CubeSat: Fabrication, deployment and results of experimental tests. In: *Proc. IEEE Aerospace Conference*, 1–12.

BBC (2017). http://www.bbc.com/news/magazine-22909590 (accessed 5 March 2017).

Beattie J.R. and Matossian, J.N. (1989). *Mercury Ion Thruster Technology*, Hughes Aircraft Company REF F3358; Hughes Research Laboratories 3011 Malibu Canyon Road, Malibu, California 90265. March 1989. NASA, NAS3-23775 Final Report 18 February 1983 through 18 October 1984.

Bluetooth (2017a). https://www8.cs.umu.se/kurser/TDBD16/VT07/Bluetooth-Tutorial-2001.pdf (accessed 29 March 2017).

Bluetooth (2017b). https://en.wikipedia.org/wiki/Bluetooth (accessed 12 April 2017).

Castanet, L., Lemorton, J. and Bousquet, M. (1998). Fade Mitigation Techniques for New SatCam Services at Ku-Band and Above: A Review. COST 255 *First International Workshop on Radiowave Propagation Modelling for SatCom Services at Ku-Band and Above*, WPP-146, pp. 243–251, October 1998.

Citizens Band (2018). https://en.wikipedia.org/wiki/Citizens_band_radio (accessed 27 April 2018).

Comsat Labs (1997). INTEL-1474 Final Report, *Demonstration of Advanced Networking Concepts*, COMSAT Laboratories, February 1997.
Cubesat (2017). https://www.nasa.gov/mission_pages/station/research/benefits/cubesat (accessed 5 May 2017).
De-Orbit (2017). http://www.dlr.de/Portaldata/55/Resources/dokumente/sart/dglr-2002-028.pdf (accessed 22 June 2017).
Digital (2017). http://www.arcelect.com/rs232.htm (accessed 9 April 2017).
Dissanayake, A.W. (1997). Application of open-loop uplink power control in Ka-band satellite links. *Proceedings of the IEEE* 85 (6): 959–969.
Drums (2017). https://en.wikipedia.org/wiki/Drums_in_communication (accessed 5 March 2017).
ESA (2017). http://www.esa.int/Our_Activities/Operations/ESA_s_Cluster_satellites_in_closest-ever_dance_in_space (accessed 12 April 2017).
ESA (2018). http://www.esa.int/Our_Activities/Observing_the_Earth/Aeolus/Facts_and_figures (accessed 21 June 2018).
Flaginstitute (2017) https://www.flaginstitute.org/pdfs/Barrie%20Kent.pdf (accessed 7 April 2017).
Gateway (2018). http://www.nalresearch.com/NetRef_Gateway.html (accessed 17 April 2018).
Globalstar (2017) http://www.globalstar.com/en (accessed 5 March 2017).
GOES (2017a). https://weather.msfc.nasa.gov/GOES (accessed 20 June 2017).
GOES (2017b). https://en.wikipedia.org/wiki/Geostationary_Operational_Environmental_Satellite (accessed 22 June 2017).
GPS (2018). https://en.wikipedia.org/wiki/GPS_wildlife_tracking (accessed 27 April 2018).
Graveyard Orbit (2017). https://en.wikipedia.org/wiki/Graveyard_orbit (accessed 22 June 2017).
Gugliotta, G. (2017). Earth, clean up your trash. *Air & Space*: 31–35.
Hall-Effect (2017). https://en.wikipedia.org/wiki/Hall-effect_thruster (accessed 7 May 2017).
Haque, S.E., Keidar, M. and Lee, T. (2013). Low-Thrust Orbital Maneuver Analysis for Cubesat Spacecraft with a Micro-Cathode Arc Thruster Subsystem, Paper IEPC-2013-365 Presented at the *33rd International Electric Propulsion Conference*, The George Washington University, Washington, DC 20052, USA, October 6–10, 2013.
Helvajian, H. and Janson, S.W. (2008). *Small Satellites: Past, Present, and Future, Aerospace Books*. The Aerospace Press/American Institute of Aeronautics and Astronautics.
Herranz-Herruzo, J.I., Valero-Nogueira, A., Ferrando-Rocherm, M. et al. (2018). Low-cost Ka-band switchable RHCP/LHCP antenna array for mobile Satcom terminal. *IEEE Transactions on Antennas and Propagation* 66 (5): 2661–2666.
Hodges, R.E., Radway, M.J., Toorian, A. et al. (2015). ISARA-integrated solar array and reflectarray CubeSat deployable Ka-band antenna. In: *Proc. IEEE International Symposium Antennas and Propagation*, 2141–2142.
Hodges, R.E., Chahat, N., Hoppe, D.J., and Vacchione, J.D. (2017). Deployable high-gain antenna bound for Mars. *IEEE Antennas and Propagation* 59 (2): 39–49.
INTELSAT (2017). INTELSAT Earth Station Standards (IESS) available on the web at http://www.intelsat.com/tools-resources/library/iess-documents (accessed 6 December 2017).
Iridiumnext (2017a). https://iridium.com/network/iridiumnext (accessed 5 March 2017).

Iridiumnext (2017b). http://www.decodesystems.com/iridium.html (accessed 10 August 2017).
ISO (2018). https://www.iso.org/standard/64784.html (accessed 21 June 2018).
ITU (1994). VSAT Systems and Earth Stations, Supplement No. 32 to the *Handbook on Satellite Communication*. International Telecommunications Union, Geneva 1994 (for updates on this handbook, please refer to www.itu.int).
ITU-R (1976). ITU-R Report 720–1, 1976, Radio emission from natural sources in the frequency range above 50 MHz, CCIR Vol. 5, *Propagation in Non-Ionized Media*, ITU, 2 Rue Varembé 1211, Geneva 20, Switzerland, 1976. ITU.
ITU-R (2017). Recommendation ITU-R S.728, Maximum Permissible Levels of Off-Axis EIRP Density for Aperture Terminals, (VSATs) https://www.itu.int/dms_pubrec/itu-r/rec/s/R-REC-S.728-1-199510-I!!PDF-E.pdf (accessed 6 December 2017).
Joint Ops (2018). https://en.wikipedia.org/wiki/Joint_Space_Operations_Center (accessed 21 June 2018).
Kuwahara, T., Yoshida, K., Sakamoto, Y., Tomiioka, Y., Fukuda, K., Sugimura, N., Kurihara, J. and Takahashi, Y. (2013). Constellation of Earth Observation Satellites with Multi-spectral High Resolution Telescopes, 27th annual *AIAA/USU Conference on Small Satellites*, paper SSC13-IV-7.
LeoLabs (2017). https://www.leolabs.space/posts/leolabs-unveils-dedicated-phased-array-radar-for-tracking-space-debris (accessed 22 June 2017).
Marconi (2017a). https://en.wikipedia.org/wiki/Guglielmo_Marconi (accessed 7 April 2017).
Marconi (2017b). http://www.history.com/this-day-in-history/marconi-sends-first-atlantic-wireless-transmission (accessed 7 April 2017).
Maritime flags (2017). https://en.wikipedia.org/wiki/Maritime_flag_signalling (accessed 5 March 2017).
Martello Towers (2017). https://en.wikipedia.org/wiki/Martello_tower (accessed 7 April 2017).
Military Aerospace (2018). https://www.militaryaerospace.com/articles/2018/04/military-satellites-secure-low-earth-orbit-leo.html (accessed June 18, 2018).
Mobile Phones (2017). https://en.wikipedia.org/wiki/History_of_mobile_phones (accessed 9 April 2017).
Monopulse (2018). https://www.microwaves101.com/encyclopedias/monopulse-antennas (accessed 26 April 2018).
Multispectral (2017). https://en.wikipedia.org/wiki/Multispectral_image (accessed 20 June 2017).
Nanoracks (2017a). http://nanoracks.com/products/smallsat-deployment (accessed 5 May 2017).
Nanoracks (2017b). https://www.nasa.gov/mission_pages/station/research/experiments/NanoRacksNCESSEAntares.html (accessed 30 October 2017).
Nanoracks (2017c). http://nanoracks.com/nanoracks-boeing-first-commercial-airlock-module-on-iss (accessed 7 May 2017).
NASA (2017a). https://www.nasa.gov/press-release/cubesat-to-demonstrate-miniature-laser-communications-in-orbit (accessed 30 May 2017).
NASA (2017b). https://www.nasa.gov/directorates/spacetech/small_spacecraft/edsn.html (accessed 30 May 2017).
NASA (2017c). https://www.nasa.gov/mission_pages/station/research/experiments/1828.html (accessed 30 May 2017).

NASA (2017d). https://www.nasa.gov/directorates/spacetech/small_spacecraft/feature/CubeSat_Technology_Missions (accessed 30 May 2017).
NASA (2017e). https://www.nasa.gov/phonesat (accessed 12 June 2017).
NASA (2017f). https://www.nasa.gov/directorates/spacetech/small_spacecraft/feature/university_partners_for_small_spacecraft_collaboration.html (accessed 12 June 2017).
NASA (2017g). https://www.nasa.gov/mission_pages/smallsats/elana/index.html (accessed 16 June 2017).
NASA (2017h). https://www.nasa.gov/mission_pages/station/research/experiments/1326.html (accessed 22 June 2017).
NASA Mars (2018). https://mars.nasa.gov (accessed 15 June 2018).
Orbcomm (2017a). https://www.orbcomm.com (accessed 5 March 2017).
Orbcomm (2017b). http://www.prnewswire.com/news-releases/crew-of-18-rescued-off-chilean-coast-using-orbcomm-satellite-network-73206812.html (accessed 9 April 2017).
Orbcomm (2018c). https://www.orbcomm.com/en/networks/satellite/orcomm-og2 (accessed 18 June 2018).
Orbcomm (2018d). https://en.wikipedia.org/wiki/Orbcomm (accessed 18 June 2018).
Orbcomm (2018e). http://spaceflight101.com/spacecraft/orbcomm-g2 (accessed 18 June 2018).
Orbital Debris (2017). https://www.orbitaldebris.jsc.nasa.gov (accessed 22 June 2017).
OSI Links (2018). https://cse.sc.edu/~wyxu/515Fall08/slides/OSI-Link.pdf (accessed 29 April 2018).
Phonesat (2017). https://directory.eoportal.org/web/eoportal/satellite-missions/p/phonesat-2-5 (accessed 12 June 2017).
Planet Labs (2017). https://www.extremetech.com/extreme/244486-india-sets-world-record-104-satellites-single-launch (accessed 20 July 2017) and https://en.wikipedia.org/wiki/Planet_Labs (accessed 20 July 2017).
Prisma (2006). Prisma – an autonomous formation flying mission. In: *ESA Small Satellite Systems and Services Symposium (4S)* (eds. S. Persson, P. Bodi, E. Gill, et al.). Sardinia, Italy. ESA Publications Division, Noordwijk, Netherlands.
Procurement (2018). https://www.forbes.com/sites/lorenthompson/2018/04/23/as-space-war-looms-air-forces-biggest-weakness-may-be-how-it-buys-space-systems/?ss=business#780ae97b78ab (accessed 25 April 2018).
Radhakrishnan, R., Edmonson, W.W., Afghah, F. et al. (2016). Survey of inter-satellite communication for small satellite systems: physical layer to network layer view. *IEEE Communications Surveys and Tutorials* 18 (4): 2442–2473, IEEE Communications Society. [cite as https://arxiv.org/abs/1609.08583].
Rahmat-Samli, Y., Monahar, V., and Kovitz, J.M. (2017). For satellites, think small, dream big. *IEEE Antennas and Propagation Magazine* 59 (2): 22–30.
Ravindra, V., Akbar, P.R., Zhan, M. et al. (2017). A dual-polarization X-band traveling-wave antenna panel for small-satellite synthetic aperture radar. *IEEE Transactions on Antennas and Propagation* 65 (5): 2144–2156.
Raychaudhuri, D. and Joseph, K. (1988). Channel access protocols for Ku-band VSAT networks: a comparative evaluation. *IEEE Communications Magazine* 26 (5): 34–44.
Rees, D.W.E. (1989). *Satellite Communications: The First Quarter Century of Service*. New York: Wiley.
Rotordronemag (2018). http://www.rotordronemag.com/flight-basics (accessed 26 April 2018).
Sarsat (2017a). http://www.sarsat.noaa.gov/emerbcns.html (accessed 16 June 2017).

Sarsat (2017b). http://www.sarsat.noaa.gov/faq%202.html (accessed 16 June 2017).
Sarsat (2017c). http://www.sarsat.noaa.gov (accessed 16 June 2017).
SatelliteToday (2017a). http://interactive.satellitetoday.com/via/april-2017/launch-overhaul-what-new-rockets-mean-for-the-next-decade (accessed 16 June 2017).
SatelliteToday (2017b). http://interactive.satellitetoday.com/via/march-2017/market-innovation-driving-cubesats-into-the-mainstream (accessed 16 June 2017).
Satellitetoday (2018a). https://www.satellitetoday.com/innovation/2017/09/05/design-delivery-build-cubesat-week/undefined (accessed 21 June 2017).
Satellitetoday (2018b). https://www.satellitetoday.com/innovation/2017/12/19/new-venture-manufacture-ultra-small-satellites-leo/unefined (accessed 3 January 2018).
Satellitetoday (2018c). https://www.satellitetoday.com/innovations/2018/03/27/astranis-targets-cellular-backhaul-with-geo-smallsats/undefined (accessed 1 April 2018).
Satellitetoday (2018d). https://www.satellitetoday.com/innovation/2018/05/21/bridgesat-laser-terminals-to-connect-iceyes-microsat-constellation/undefined (accessed 22 May 2018).
Satellitetoday (2018e). https://www.satellitetoday.com/innovation/2018/05/18/spacedatahighway-achieves-10000-successful-laser-co/nnections (accessed 23 May 2018).
Satellitetoday (2018f). https://www.satellitetoday.com/innovation/2018/01/31/students-launch-cubesats-iss-study-orbital-decay/undefined (accessed 1 February 2018).
Satellitetoday (2018g). https://www.satellitetoday.com/innovation/2018/04/04/removedebrismission-to-test-conceptsfor-cleaning-up-space (accessed 4 April 2018).
Satellitetoday (2018h). https://www.satellitetoday.com/innovation/2018/04/03/water-to-propel-blacksky-and-leostella-satellites (accessed 5 April 2018).
Seligman, L. (2018). How Shadowy 'Project Maven' uses AI to mine combat data. *Aviation Week and Space Technology*: 42–43.
Semaphore flags (2017). https://flagexpressions.wordpress.com/2010/03/23/history-behind-semaphore-flags (accessed 5 March 2017).
Semaphore line (2017). https://en.wikipedia.org/wiki/Semaphore_line (accessed 5 March 2017).
SmallSat (2017). http://theconversation.com/smallsat-revolution-tiny-satellites-poised-to-make-big-contributions-to-essential-science-71440 (accessed 16 June 2017).
Smoke Signals (2017). https://en.wikipedia.org/wiki/Smoke_signal (accessed 5 March 2017).
Software Defined Radio (2017). http://www.radio-electronics.com/info/rf-technology-design/sdr/software-defined-radios-tutorial.php (accessed 6 December 2017).
Space Debris (2017). https://phys.org/news/2014-02-space-debris-satellites.html (accessed 22 June 2017).
Space Fence (2017). http://www.lockheedmartin.com/us/products/space-fence.html (accessed 22 June 2017).
Space Junk (2017). https://www.cnet.com/pictures/space-junk-worse-than-you-think-pictures (accessed 22 June 2017).
Space Junk (2018). Fishing for Space Junk, *IEEE Spectrum*, June 2018, pp. 7–9.
Spaceacademy (2017). www.spaceacademy.net.au/spacelink/radiospace.htm (accessed 6 December 2017).
Spaceflight (2017). http://www.spaceflight.com/universities-space-seriously-higher-education (accessed 16 June 2017).

Sprite (2018). A Sprite-ly Spacecraft, *Discovermagazine*, November 2018. http://discovermagazine.com/2018/nov/a-sprite-ly-spacecraft (accessed 18 November 2018).

Sweeney, D.G. and Bostian, C.W. (1999). Implementation of adaptive power control as a 30/20 GHz fade countermeasure. *IEEE Transactions on Antennas and Propagation* 47 (1): 40–46.

Tepper, M. (1961). The report on the meteorological satellite program. *Weather Wise* 14 (4): 131–138.

Tiainen, A. (2017). *Inter-Satellite Link Antennas: Review and the Near Future*. Approved version, Master's Thesis, March 16, 2017, Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Sweden. https://ssel.montana.edu/edsn.html (accessed 26 May 2017).

TIROS (2017). https://www.lib.noaa.gov/collections/TIROS/tiros.html (accessed 16 June 2017).

Tsay, M, Frongillo, J., Model, J., Zwahlen, J. and Paritsky, L. (2016). Flight Development of Iodine BIT-3 RF Ion Propulsion System for SLS EM-1 CubeSats, Presented at *30th AIAA/USU Conference on Small Satellites*, 6–11 August 2016, North Logan, Utah Pre-Conference Workshop SSC16-WK-39.

Urban WiFi (2018). https://www.computerworld.com/article/2970867/wireless-networking/how-new-white-space-rules-could-lead-to-an-urban-super-wi-fi.html (accessed 14 June 2018).

Vasseur, H., Czarnecki, M., Castanet, L. and Bousquet, M. (1998). Performance Simulation of a Ka-Band VSAT Videoconferencing System, *COST 255 First International Workshop on Radiowave Propagation Modelling for SatCom Services at Ku-Band and Above*, WPP-146, October 1998, pp. 227–234.

Vectornav (2017). http://www.vectornav.com/products/vn-100 (accessed 15 May 2017).

VSAT (2015). VSAT systems and earth stations, Supplement No. 3. In: *Handbook on Satellite Communication*. Geneva, Switzerland: International Telecommunications Union, 1994. (Updated in 2015 at www.itu.int).

VSAT (2017). *INTELSAT VSAT Handbook* available at. 7900 Tyson's One Place, McLean, VA, USA: Applications Support & Training, INTELSAT.

Warren, P.A., Steinbeck, J.W., Minelli, V., and Mueller, R.J. (2015). Large, deployable, S-band antenna for a 6U CubeSat. In: *Proc. 29th. Annu. American Inst. Aeronautics and Astronautics/Utah State University Conf. Small Satellites*, 1–7.

Wildlife (2018). https://en.wikipedia.org/wiki/GPS_wildlife_tracking (accessed 27 April 2018).

9

NGSO Satellite Systems

Most people take for granted the movement of the moon around the earth and its changing phases, but it was not always so. Even though Johannes Kepler had explained the motion of a smaller world – a satellite – around a larger world more than a dozen years earlier, Galileo was still disbelieved when he spoke of the moons of Jupiter and that the earth itself revolved around the sun. Venturing outward is in the DNA of humankind. Over the last 50 years, all of the planets in our solar system have been explored by space probes, including one in the *Kuiper Belt*. The *New Horizon* probe gave us a first look at one such object in the *Kuiper Belt* called *Ultima Thule* that was at a distance of 1.6 billion kilometers beyond the orbit of Pluto. Closer to home, who has not gone outside of their home at dusk to see if they can spot the International Space Station? Watching it move across the sky at 7 km/s, it is possibly the ultimate non-geostationary satellite.

9.1 Introduction

9.1.1 The Beginning of the Space Age

The first successful earth satellite launched was Sputnik 1, on the night of 4/5 October 1957. It is not known how many launch failures preceded this success, but if the USSR failure rate was similar to that of the United States, possibly three or four. Paradoxically, it was probably this high launch failure rate that led to the proposal of COMSAT being accepted by the group of Signatories that was to become INTELSAT (the International Telecommunications Satellite Organization). The COMSAT Corporation (COMSAT 1994) was created by the Communications satellite Act of 1962 to manage the US participation in INTELSAT. A competing idea to initiate international satellite communications was submitted by the major telecoms companies in England (the British Post Office) and the United States (AT&T). That proposal was for a chain of medium earth orbit (MEO) satellites around the equator, rather than using the geostationary orbit, sometimes called the Clarke orbit (Clarke 1945). It is easy to understand why: at this point (1962) no satellite had been successfully put into a geostationary orbit. The 10, or so, MEO satellites needed to complete 24/7 worldwide communications would have required 30–40 rocket launches to successfully complete the initial system. The reasoning by COMSAT for proposing the use of the geostationary orbit was simple: one successful launch would provide 24/7 communications over a third of the world. Hence up to four launches (assuming one in four succeeded) would be needed rather than up

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

to 40 for 24/7 links over one of the three ocean regions (Atlantic, Indian, and Pacific). To complete worldwide coverage from geostationary orbit, up to 12 launches would be needed. And so it was that the geostationary orbit proposal by COMSAT was accepted, but the geostationary orbit was not an easy option at the time.

The first attempt to reach geostationary orbit seemed to fail when communications was lost with the spacecraft, although radar indicated a quasi-geostationary orbit was achieved (NASA 1994a). The second attempt placed the satellite into an inclined orbit at geostationary altitude, while the third was fully successful, achieving geostationary altitude and almost zero inclination (NASA 1994b). In April 1965, INTELSAT 1, also known as Early Bird (space skyrocket 2018a), ushered in commercial satellite communications and, by the time of the Apollo 11 mission in July 1969, INTELSAT had satellites over the three ocean regions. Thus, worldwide television coverage was available for the first step onto the moon that was relayed from the antenna at Honeysuckle Creek, near Canberra, in Australia. As Australia is a day ahead of the United States, the moon landing took place on 21 July 1969, in Australia, rather than 20 July 1969, when it is celebrated in the United States.

The Clarke orbit, more usually called the geostationary earth orbit (GEO) or geostationary satellite orbit (GSO), is a unique resource that has enabled the generation of many billions of dollars in revenues per year from communications satellites and the associated launchers. Communications satellites and their launchers were the only commercial space ventures in the twentieth century that had any significant return on investment. This may change in the twenty-first century with the new generation of non-geostationary satellite orbit (NGSO) satellite constellations currently under development, in deployment, and in operation for a variety of commercial ventures. The first ventures in the late twentieth century (e.g., Teledesic 2000 and Skybridge 1997) were, unfortunately, conspicuous failures. The initial Teledesic constellation was for 924 satellites orbiting at an altitude of 700 km, subsequently reduce to 288 orbiting at 1500 km: that of Skybridge was 64 satellites at 1500 km altitude. The relatively large satellite mass (800 kg) and number of launches required to complete the networks, ultimately could not provide a satisfactory return on investment. These big low earth orbit (LEO) systems, as they became known, were feasible with the technology available at the time but lacked the breakthroughs necessary to make them economically viable. Two of the key breakthroughs necessary were the micro-miniaturization of complete subsystems (e.g., micro-thrusters, three-axis determination, and antennas, which are discussed in Chapter 8) and launchers (see Chapter 2).

9.1.2 The Altitude Range of NGSO Satellites

There are two belts of ionized radiation surrounding the earth, as we shall see later in this chapter (Allnutt 2011). The two concentrations of ionized energy are called the van Allen radiation belts after their discoverer, professor van Allen. He designed the instrument on the first US satellite, Explorer 1, that was aimed at measuring the energetic energy of ionized radiation. His instrument, at times, could not measure accurately the energy and so van Allen determined, correctly, that his instrument had become saturated at certain parts of the orbit, which ranged from 354 to 2515 km above the earth (NASA 1994a). The terms LEO and MEO are generally used for specific orbit altitude ranges. LEO satellites are confined between an upper orbit altitude of about 1500 km by the lower ionospheric belt and a lower orbit altitude dictated by atmospheric drag (generally around 500 km).

MEO satellites have a lower orbit altitude of around 1500 km and generally an upper bound set by the GEO altitude of around 36 000 km. Most MEO systems orbit in the 10 000–15 000 km range, although there are notable exceptions (e.g., the Chandra X-Ray Observatory, which has an apogee of 134 527.6 km and a perigee of 14 307.9 km) (NASA 1994b).

LEO and MEO satellites – now generally referred to generically as NGSO satellites – have been used in a variety of roles. From an era in the late 1950s when every launch made front page news we have now become somewhat blasé about satellites: they have become part of everyday life, much like computers, iPhones, and the internet. NGSO satellites brought us the first voice broadcast from orbit (SCORE), the first pictures of our cloud cover for weather forecasting (TIROS), the first navigation aids in space (TRANSIT), the first live television pictures across oceans (TELSTAR), the first Geographic Information Systems pictures of the earth (SPOT), the first infrared, ultraviolet, and X-ray view of the universe from outside the earth's atmosphere and, of course, the first manned missions (Vostok and Mercury). Each of these missions has been succeeded by more complex satellites with more advanced capabilities, perhaps the most complex currently being the International Space Station – ISS: more volume inside than a Boeing 747 with 84 kW of power to run the dozens of experiments on board. As the satellite missions became more complex, the requirements for the specific orbits became more precise. Some satellites have to be very close to the earth, some in highly elliptical orbits (HEOs), and yet others in orbits with a plane that matches the view angle to the sun. This chapter reviews the different earth orbits available and what missions may use them to advantage, beginning first with the GSO.

9.1.3 The 50-Year Reign of GSO Satellites

The GSO has been the preferred orbit for satellite communication systems for about 50 years, although this dominance may cease by the mid-2020s as advanced NGSO Systems begin to mature. There was a consistent downturn in ordering GSO satellites in the second decade of the twenty-first century, from an average of 20 to 25 a year to only 17 in 2016 and 2017 (Satellite Today 2017a,b). The reason for the long dominance of geostationary satellites is simple: more bits can currently (2018) be sent per dollar of capital investment when a satellite is in a geostationary orbit than in any other orbit. This was realized quite early in the development of satellite communications, and Intelsat, which was the first provider of commercial satellite systems, developed a series of geostationary satellites, beginning in 1965 with Early Bird (renamed Intelsat I). International and domestic satellite systems followed in the 1970 and 1980s, all using GEO satellites. Direct to home (DTH) satellite television broadcasting, one of the most financially successful applications for satellite systems, also requires GEO satellites so that customers can use small, fixed dish antennas. In such a direct broadcasting satellite-TV system (DBS-TV), the major investment is in the earth stations; not in the satellites. Ten million earth stations bought for US$250 each, for example, cost US$2.5B, well in excess of the cost of a cluster of GEO DBS-TV satellites.

There are some specialized applications that require non-geostationary satellites. Surveillance of the earth's surface, using both optical (e.g., *Keyhole* satellites (Spaceflightnow 2015)) and radar techniques need low altitude orbits. While optical surveillance can be accomplished by a single satellite using simple technology, radar applications, particularly those that use synthetic aperture processes, are much more complicated

and expensive. A single, synthetic aperture radar has been found to offer less capabilities than a chain of smallsats, each with overlapping coverages (Spacenewsmag 2016). Satellites providing global navigation, such as the global positioning system (GPS) constellation, must utilize orbits that place the satellites in widely spaced positions in the sky, as seen by the receiver. Some of the satellites can be in GEO, but most must be in inclined orbits with an even distribution over the earth's surface. GPS uses 24 satellites in orbits with an altitude of around 20 000 km and an inclination of 55°.

Mobile satellite communication systems demand an earth station with a low gain antenna that has a near omnidirectional pattern. A GEO satellite used for communication with a satellite telephone that is handheld, like a cellular telephone, requires a very large antenna with hundreds of individual beams to achieve the necessary gain. A high gain satellite antenna is needed to compensate for the low gain of the antenna employed by the user's telephone handset. An alternative to a GEO satellite with a high gain antenna is a LEO or MEO satellite constellation with a smaller multibeam antenna. Because the satellite is not geostationary, a large number of satellites is required to maintain continuous coverage over any particular location on the earth. For example, the Iridium system uses 66 operational satellites in LEO to provide continuous global coverage.

Building, launching, and maintaining a constellation of communication satellites in LEO is expensive. When LEO satellite constellations were first proposed for mobile satellite services (MSSs), the satellites were envisaged to be small, simple, and low cost compared with GEO satellites. Early estimates for the cost of the Iridium system, for example, were between US$1B and US$2B. As the development of the LEO systems progressed, the satellites became more and more complex and their cost steadily increased, becoming comparable to the cost of GEO satellites. The satellites proposed for the ICO global system, for example, were actually modified versions of a large GEO satellite (Spaceflightnow 2001). Since any LEO or MEO system requires many more satellites than a GEO system serving the same region, the cost of a LEO or MEO system using such large satellites will exceed the cost of the equivalent GEO system.

The Iridium system eventually cost well over US$5B, compared with a typical cost of US$250M to launch and maintain a single, large GEO satellite. Iridium failed as a commercial venture because the final cost greatly exceeded initial projections, and the system was unable to attract a sufficient number of customers quickly enough. Debt repayments on the high capital cost of the system came due before the customer base had built to a large enough size to service the debt. Analysis of the cost per bit transmitted through an Iridium satellite shows that it is much higher than the cost per bit for a GEO satellite, and any LEO system must therefore be able to offer considerable advantages to its customers over that of an equivalent GEO system if it is to succeed commercially. While Iridium was an incredible technological success, it was not successful commercially. The fact it was ever funded led to speculation that it was part of the Strategic Defense Initiative (Cold War 2008), being part of the Brilliant Eyes (SDI 2018), and Brilliant Pebbles (New York Times 1989) of the Reagan administration. In a remarkable coincidence, 20 years after the launch of the first five Iridium satellites, in May 1997, the first 10 second generation Iridium satellites were launched into orbit by SpaceX on 25 June 2017. The second generation spacecraft in the Iridium constellation, however, has a key secondary payload: ADS-B (automatic dependent surveillance-broadcast). ADS-B technology permits the accurate tracking of aircraft, both en route and in the take-off and descent phases. New constellations of NGSO satellites in a variety of orbits will be launched over the next decade. Some of these are aimed at markets not yet well served

by GSO satellites; others hope to provide the necessary linkage between fixed satellite services (FSSs) and MSSs within the terrestrial network. Yet others specifically target broadband internet provisioning. We shall see that choosing the correct orbit (altitude, inclination, and eccentricity) are major factors in the design of NGSO systems.

This chapter discusses a number of applications and satellite systems that are not in GEO orbit, beginning with those in simple, circular, equatorial orbits; moving through simple inclined orbits to those with high eccentricity; and then reviewing those that take advantage of specific attributes of their orbit for observations (sun synchronous orbits) or the provisioning of navigation services through half-sidereal periodic orbits (GPS). The so-called inclined orbit GEO satellites are not discussed in this chapter. These satellites, once fully stationary GEO satellites, have had their in-orbit operational life extended by removing station keeping in the N–S direction while maintaining E–W station keeping so that the average subsatellite point remains nominally the same. Such inclined-orbit operation was first started after an unusual run of launch vehicle failures in the 1986 time frame when every single type of commercial satellite launcher failed (including the tragic loss of the space shuttle Challenger). The up to two-year hiatus in some satellite replenishment programs forced inclined orbit operation of GEO satellites on all service providers. Currently, such problems are avoided by designing the orbital maneuvering life (OML) to be many years longer than the orbital design life (ODL). This aspect is discussed in Chapter 2.

The first spacecraft launches relied on terrestrial radar tracking and guidance commands transmitted from the ground. This, and the relatively crude control capabilities of the rockets themselves, dictated relatively wide error bounds for the intended orbit. Indeed, achieving orbit in those early days – any orbit – was declared a success! Rapid advances in rocketry, which included the ability for multiple restarts of high-energy upper stage engines, and the inclusion of sophisticated onboard guidance computers, quickly enabled spacecraft mission planners to design with some confidence orbits that were mission-specific. That is, the mission could not be declared successful unless the designed orbit was achieved within the specified tolerance. In some cases, the mission was for a single spacecraft (such as a meteorological satellite) while, in others, a constellation of spacecraft would be required to achieve the mission goals. In all cases, careful analysis of the mission goals led to the selection of a particular orbit altitude, ellipticity, and inclination and system architecture (number of satellites, number of planes, spacing of satellites within the plane, connectivity, etc.). Quite often, tight launch windows were also dictated – specific time periods when the launches had to be executed.

In the sections that follow, we will examine the parameters that need to be determined in the selection of an orbit that will achieve given mission goals. Only earth orbit missions are considered; spacecraft missions requiring escape velocity from the earth are beyond the scope of this book.

9.2 Orbit Considerations

Once in orbit, the motion of a satellite is determined by orbital mechanics, as discussed in Chapter 2, Orbital Mechanics and Launchers. However, while the satellite moves in

such a way as to balance centrifugal and centripetal forces, the earth is also in motion beneath it. As well as rotating once a sidereal day, the earth also moves around the sun; and the solar system, with the sun at its center, is orbiting around the center of the home galaxy, the Milky Way. There is therefore a complex relationship between the various motions of the natural and artificial bodies. How many of these need to be considered simultaneously will depend on the design goals of the satellite system. A satellite designed to observe the earth's surface will not need to know where the stars are at any particular time, but the location of the local star, the sun, may be important if the satellite needs to use sunlight to illuminate its coverage region on the surface of the earth. On the other hand, a satellite designed to observe background thermal radiation levels of deep space in the infrared band will need to know the position of each of the neighboring planets. Should the telescope of the satellite inadvertently point toward one of these planets, the temperature viewed would not reflect that of the true background radiation level. In the sections that follow, we will review all of the different NGSO orbits that have been used for scientific, military, and commercial satellite missions. The simplest NGSO orbit is an equatorial orbit.

9.2.1 Equatorial Orbits

Equatorial orbits lie exactly in the plane of the geographical equator of the earth. That is, the orbital path lies directly above the equator at all times. In order to take advantage of the 0.45 km/s eastward rotational velocity of the earth, most satellites are launched toward the East into a prograde orbit. A westerly directed orbit is called a retrograde orbit. A satellite in an eastwardly directed equatorial orbit will have two periods: a real orbital period that is referenced to inertial space (the galactic background) and an apparent orbital period that is referenced to a stationary observer on the surface of the earth. The real orbital period, denoted here as T hours, is given by Eq. (2.6). The apparent orbital period to the observer on the equator will be P hours where

$$P = (24T)/(24 - T) \tag{9.1}$$

To be exact, 23.9344 hours, one sidereal day, should be used in place of 24 hours in Eq. (9.1). Table 9.1 illustrates the difference between P and T for a number of orbital altitudes and elevation angles. It also shows the time the satellite is visible to the observer, neglecting atmospheric refraction and assuming the satellite is in a circular equatorial orbit and passes directly over the observer. The observation time assumes that the satellite can be tracked down to 0°, that is, right down to the horizon, and represents the longest time for which a satellite at the given orbital altitude can be observed. Any pass that does not

Table 9.1 Orbital periods and observation time

Orbital height (km)	Orbital period (h)	Apparent period (h)	Observation time (h)
500	1.408	1.496	0.183
1469	1.921	2.089	0.387
5000	3.355	3.902	1.212
10 255	5.930	7.883	2.954
35 786	23.934	∞	∞

go directly over the observer will have a shorter observation time. Other implications of the observing time are considered in more detail in Section 9.3, Coverage and Frequency Considerations.

The plane of a satellite's orbit must be in the plane of the equator for the satellite to be in equatorial orbit. This can be achieved by launching the satellite in one of two ways. The first launch method is to locate the launch site on the equator and to launch the spacecraft toward the east along the equatorial plane. The second method is to launch the satellite into an inclined orbit and to execute a maneuver either during the launch trajectory or when the satellite is in an inclined orbit that changes the plane of the initial orbit so that the final orbit is in the plane of the equator. Removing the inclination from the orbit, so that the satellite orbits exactly over the equator, requires significant energy, particularly if the launch site is well removed from the equator. The first two sites from which orbital flights were made, Cape Canaveral in the United States and Baikonur in Kazakhstan, were not close to the equator (approximately 28°N and 46°N, respectively). In addition, the early launch vehicles lacked the ability to alter the trajectory significantly during launch. The first artificial earth satellites were therefore placed into inclined orbits, that is, the planes of the orbits were inclined to the equatorial plane.

9.2.2 Inclined Orbits

There are advantages and disadvantages to inclined orbits, depending on the mission goals and the data recovery requirements. The greater the inclination of the orbit is, the larger the surface area of the earth that the satellite will pass over at some time in its flight. Figure 9.1 illustrates this for a LEO satellite.

In Figure 9.1b, the inclined orbit will take the spacecraft, at one time or another, over the earth's entire surface that lies approximately between the latitudes given by ± (the orbital inclination). For example, an orbit with an inclination of 30° will cover all regions that lie approximately between latitudes 30° north and 30° south. The superior coverage of the earth with an inclined orbit satellite is counterbalanced by the disadvantage that the master control station (MCS) will not be able to communicate directly with the satellite on every orbit as with an equatorial orbit satellite. A LEO satellite orbits the earth with a period of about 90–100 minutes and, for an inclined orbit satellite, the earth will have rotated the MCS out of the path of the satellite on the next pass over the same side of the earth. Depending on the quantity of data that need to be passed to the MCS, or if real-time communications are required continuously, a system architecture that employs multiple satellites will need to be considered.

The simplest, and lowest cost, solution to pass data between an inclined orbit satellite and an MCS is to design the satellite to store the data acquired over many orbits (when it is out of sight of the MCS) and then, when it passes within radio range of the MCS, to dump the data rapidly to the MCS. This is called store-and-forward and it is one of the capabilities of some LEO systems, including Orbcomm satellites (Orbcomm 2018). It was also the technique used for the very first communications satellite, Project SCORE, in December 1958. In the Orbcomm system, if a user on the ground is unable to establish contact via an Orbcomm satellite to a gateway earth station (GES) in the Orbcomm system, a "GlobalGram©" may be left stored within the satellite for later transmission to the GES when it comes into view of the satellite. The downlink transmission rate must be high enough to enable all of the stored messages in a LEO satellite to be sent to the MCS in the period when it is within range of the satellite. If

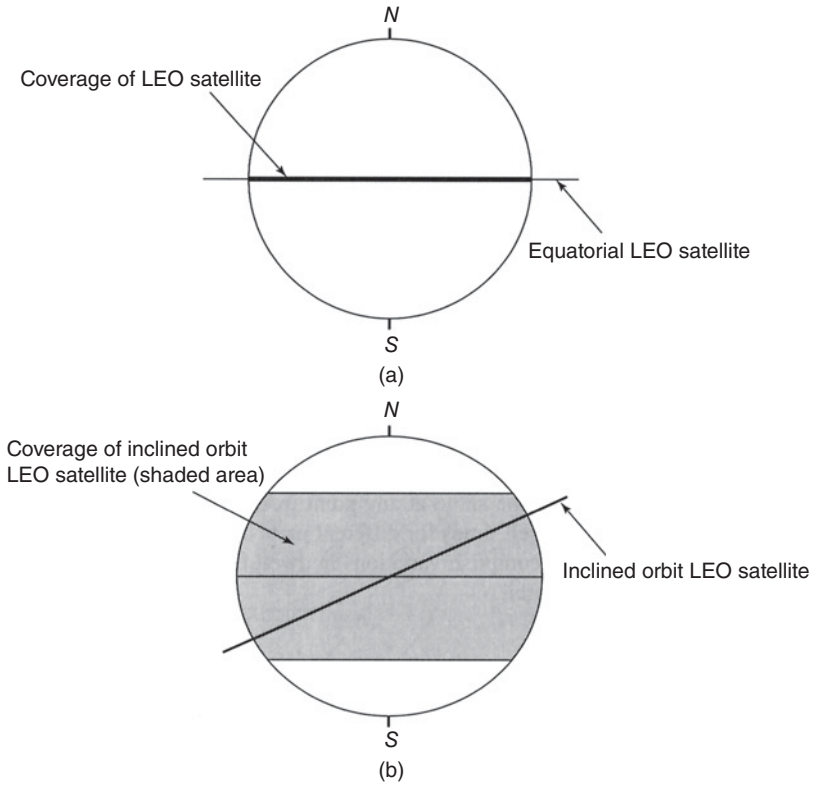


Figure 9.1 (a) Coverage of an equatorial orbit LEO satellite. The LEO satellite is in an equatorial LEO orbit and so it will only pass over the equator. The coverage of the equatorial LEO satellite will therefore be limited to a swathe of the earth close to the equator, determined by the height of the orbit and the beamwidth of the satellite's antenna. In this example, the orbit is assumed to be circular and the antennas beamwidth has been ignored. (b) Coverage of an inclined orbit LEO satellite. The LEO satellite is in an orbit that is inclined at approximately 40° to the equator. The satellite will therefore pass over, at one time or another, all regions of the earth between 40°N and 40°S of the equator. The coverage of the inclined orbit LEO satellite will therefore be a swathe of the earth between ±40° of the equator, determined by the height of the orbit and the beamwidth of the satellite's antenna. In this example, the orbit is assumed to be circular and the antenna beamwidth has been ignored. Note: The higher the orbit and the greater the inclination, the further the satellite's total coverage will reach.

a continuous, real-time connection is required between a LEO satellite and the MCS, there are only two approaches that can be used.

- The first approach is to locate control stations around the world so that the LEO satellite is never out of sight of at least one of the control stations. Terrestrial or GEO satellite connections are then established between the many control stations and the MCS to bring the LEO data back to the MCS in real time.
- The second approach is to establish intersatellite links (ISLs) to relay the LEO data traffic back to the MCS. The ISLs can either be set up amongst the LEO satellites in the constellation, so that the LEO data traffic is relayed between the LEO satellites in orbit via their ISLs, or the ISL link can be set up between the LEO satellite(s) and one or more GEO satellites. The GEO satellite relays the LEO data traffic back to the

MCS directly, if it is within sight of the MCS, or via another GEO satellite. Iridium (Iridium) adopted the former solution of LEO satellites interlinked via ISLs within their own constellation. Globalstar (Globalstar) chose a different approach. The reason for the different approaches of Iridium and Globalstar is historical. The builders of Globalstar had significant ties to the major telecommunications companies at the time and so sought a system architecture that included them in the distribution of the traffic. The builders of Iridium had no such ties, and if the rumor that the Iridium LEO system was inspired by military planners who did not want any of the traffic passing through countries with soviet background is true, it is clear that an ISL architecture was required. NASA used the tracking and data relay satellite system (TDRSS) satellites for shuttle missions (see sidebar) and geostationary relay satellites have also been used for military reconnaissance missions by at least the United States and Russia – transmitting onward data received from LEO observation satellites. Figure 9.2 illustrates the two concepts.

In both of the examples shown in Figure 9.2, the LEO satellite is in a circular orbit. A circular orbit gives a constant dwell time over a given coverage region since the angular velocity of the satellite is the same at any point in the orbit. In many cases, mission goals will dictate different dwell times for different parts of the orbit. A circular orbit will not achieve this result. To accomplish variations in dwell time around the orbit, the satellite must be in an elliptical orbit.

NASA built and operated a number of relay stations around the globe for the Mercury, Gemini, and Apollo programs. In none of these missions was the manned spacecraft ever out of real-time contact with the Manned Spaceflight Center in Houston, United States (which acted as the MCS in this case) except when the Apollo craft was behind the moon or in the re-entry phase where ionized plasma causes a radio blackout for all spacecraft.

Communication with the Space Shuttle was maintained using the Tracking and Data Relay Satellite System (TDRSS). This same approach is used for communications from Houston to the International Space Station. Several TDRSS satellites in geostationary orbit relay data from the ISS to earth stations around the world that then send the data to NASA's MCS for manned space flight in Houston, Texas.

9.2.3 Elliptical Orbits

As noted in Chapter 2, an elliptical orbit will have a non-zero eccentricity. The orbit eccentricity, e, is determined by the lengths of the semimajor axis, a, and the semiminor axis, b, of the orbit ellipse

$$e^2 = 1 - (b^2/a^2) \tag{9.2}$$

Alternatively, if R_a is the distance between the center of the earth and the apogee point of the orbit and R_p is the distance between the center of the earth and the perigee point, the eccentricity is

$$e = (R_a - R_p)/(R_a + R_p) \tag{9.3}$$

Figure 9.3 illustrates the geometry of Eq. (9.3).

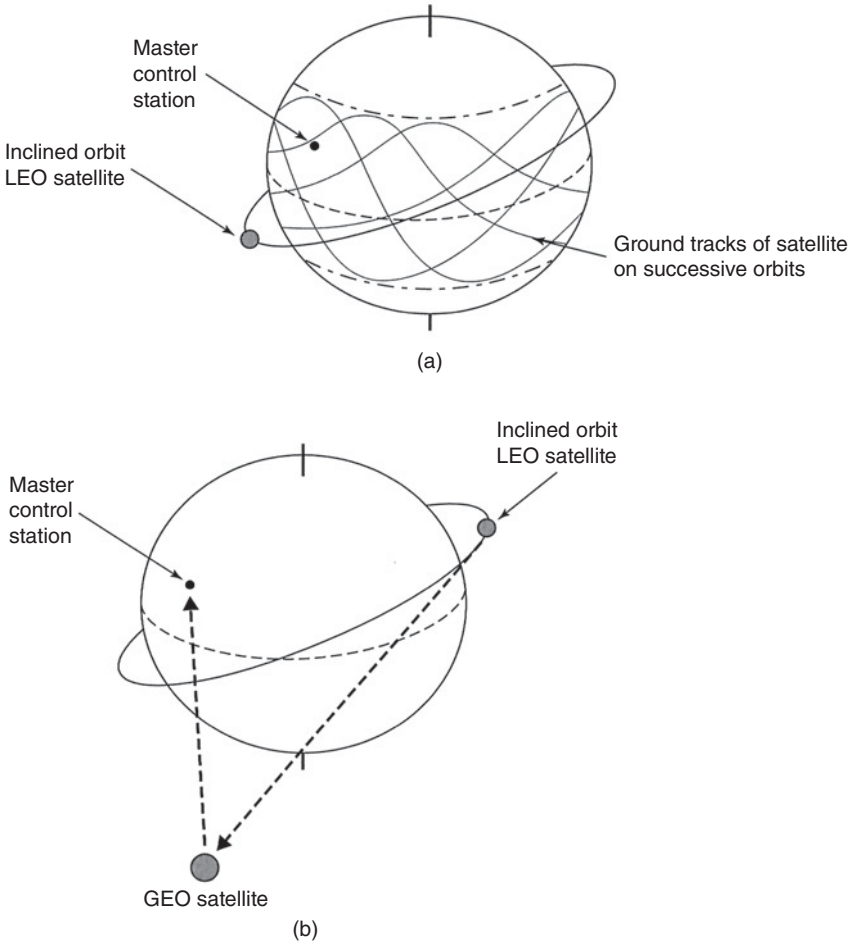


Figure 9.2 (a) Store-and-forward concept. In this LEO application, the satellite stores information it has gathered while orbiting the earth and, once within range of the master control station, it downloads the stored data. The (uplinked) data storage rate is usually low, a few kbps at most, while the download is at a much higher rate due to the small time the satellite has available when it is within range of the master control station. (b) Real-time data transfer via a GEO satellite. In this approach, the LEO satellite can transfer data in real time via the GEO satellite to a master control station whenever it can see the GEO satellite. If there were a number of GEO satellites equipped with intersatellite links (ISLs) distributed around the geostationary orbit, then the LEO satellite need never be out of real-time contact with the master control station.

In Eqs. (9.2) and (9.3), if the orbit is exactly circular, $a = b$ and $R_a = R_p$, and the eccentricity reduces to zero. In general, no orbit is truly circular for a variety of reasons, but eccentricity values of 10^{-3} or less can be considered to correspond to circular orbits for all practical purposes. The eccentricity is another way of describing the variation in the radius of the orbit. If R_{av} is the average radius of an orbit from the center of the earth, then the variation, Δ_R, in the orbital radius, is given by (Morgan and Morgan 1993)

$$\Delta_R = \pm eR_{av} \tag{9.4}$$

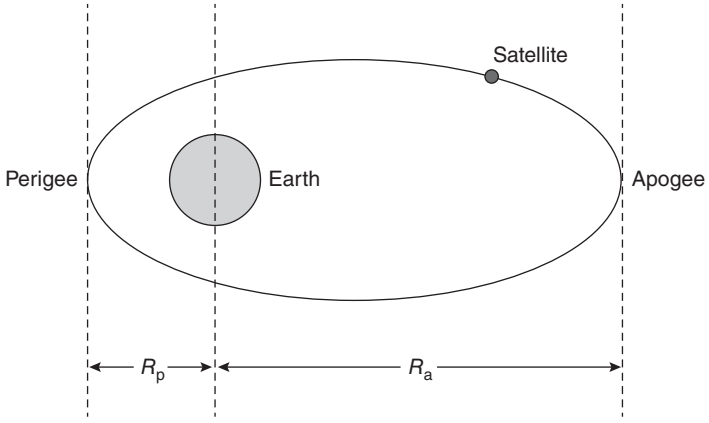


Figure 9.3 Schematic of an elliptical orbit illustrating eccentricity. The satellite orbits the earth with a perigee distance from the center of the earth given by R_p and an apogee distance from the center of the earth given by R_a. Note that the perigee and apogee are *always* exactly opposite each other in the orbit. This is true of any object in any orbit around any other body.

For a geostationary satellite ($R_{av} = 42\,164.17$ km) with an eccentricity of 10^{-4}, Δ_R will be ± 4.2 km. For a LEO constellation with a circular orbit of approximately 800 km above the earth, with each LEO satellite having an eccentricity of 10^{-4}, Δ_R will be ± 0.7178 km (assuming the earth mean radius is 6378.137 km). If the orbit becomes less circular and the eccentricity increases to 10^{-3}, Δ_R increases to ± 7.178 km. If the LEO satellites in the constellation pass over (under) each other, then the vertical separation must be sufficient to prevent any likelihood of a collision between satellites. The average orbital altitude and eccentricity of the orbit will determine the likelihood of a collision. One of the more famous orbits has an eccentricity ≈ 0.74. This is a special case of a HEO known as the Molniya orbit.

9.2.4 Molniya Orbit

The former Soviet Union had a difficult communications design problem. Much of the landmass is in far northern latitudes. Archangel, the port on the White Sea, is close to latitude 60°N; immense tracts of Siberia lie inside the Arctic Circle. To compound the problem further, the country was spread across 11 time zones: it was the largest country in the world (and Russia still is). The signals from a geostationary satellite can reach well inside the Arctic Circle if operations at elevation angles below 5° are permitted, but a single GEO satellite cannot reach that far north over 11 time zones simultaneously. A new type of orbit was required to provide good communications coverage over the former USSR. What transpired was the Molniya system.

The first Molniya satellite was launched in April 1965, becoming the second 24/7 domestic satellite system after Canada's Anik satellite, and it gave its name to both the system of satellites and to the unique orbit. The word Molniya means *flash of lightning* in Russian. Anik means *little brother* in the language of the first nations of Canada. The apogee of the Molniya orbit is at an altitude of 39 152 km and the perigee is at an altitude of 500 km. The orbital period is 11 hours and 38 minutes and the orbital inclination is 62.9°. This combination of apogee, perigee, and inclination ensures that the ground

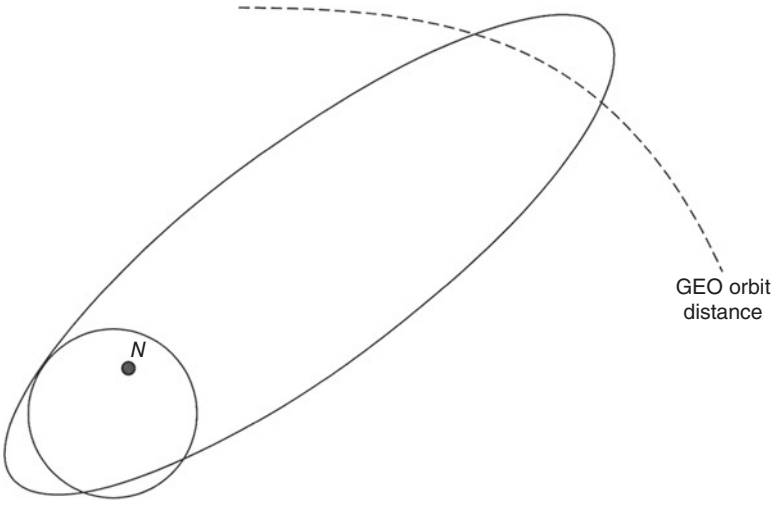


Figure 9.4 Schematic of a Molniya orbit. In this example, the trajectory is configured to have a large dwell time over the northern part of the orbit so that it can serve a country that has most of its landmass in this region. This was the design adopted for the original Molniya system of the former Soviet Union. Approximately 60% of the Molniya orbit, which stretches more than 3000 km beyond the height of a GEO orbit, has good look angles for latitudes between 30°N and 90°N. This translates to more than 6 hours of the 11 hours 38 minutes orbital period.

track of the Molniya orbit repeats every other orbit. That is, if the orbit passes exactly over Moscow on orbit one, it will do so again on orbit three, five, seven, nine, and so on. Figure 9.4 illustrates the orbit geometry.

Two Molniya orbits, with the planes of their orbits separated by 180°, will thus provide coverage over the extreme latitudes of Russia for 24 hours per day using two satellites, if correctly phased – one in each of the two Molniya orbits. When one of the satellites is at its apogee over Russia in Molniya orbit one, the other satellite will be at its perigee somewhere over the south Indian Ocean in Molniya orbit two. By the time the second satellite has moved to its apogee in Molniya orbit two, the earth will have rotated half a turn under it and Russia will again be spread beneath the satellite. Figure 9.5 illustrates the dual Molniya orbit concept.

The two-satellite, dual Molniya orbit requires that earth station operations be carried out at elevation angles well below 30° for full 24 hours-per-day coverage of one region. Note that if four satellites are used, two in each of the two Molniya orbits, diametrically across the orbit from each other, service could be provided at each of the apogee segments on the opposite sides of the Northern Hemisphere. By one of those strange cold war coincidences, the second coverage area would have been North America. If operations must always be above an elevation angle of 30°, then four Molniya orbits are required. The planes of the four Molniya orbits should be orthogonally distributed around the earth with one satellite in each Molniya orbit, correctly phased in its own orbit to provide coverage from the apogee sections of that orbit as the region rotates beneath the four satellite constellation. Up to eight Molniya satellites, in eight different Molniya orbital planes separated by 45° and suitably phased around their orbit, have been used to provide continuous coverage over Russia.

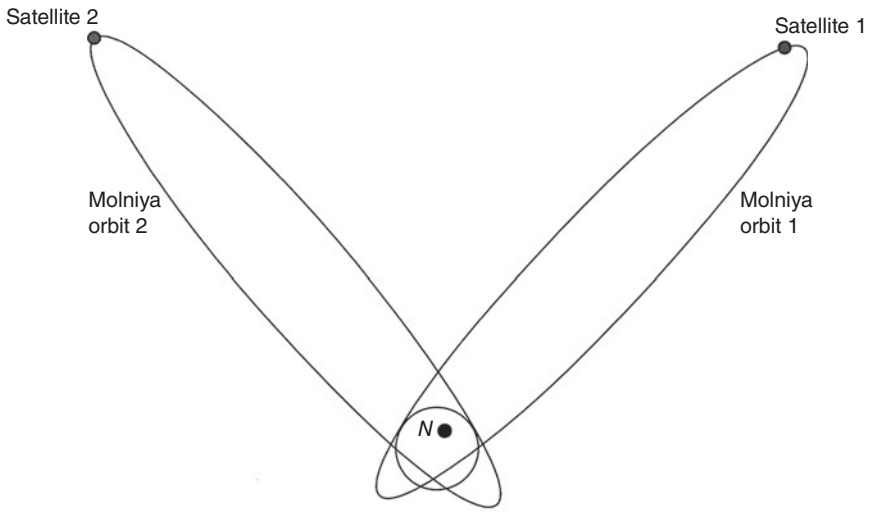
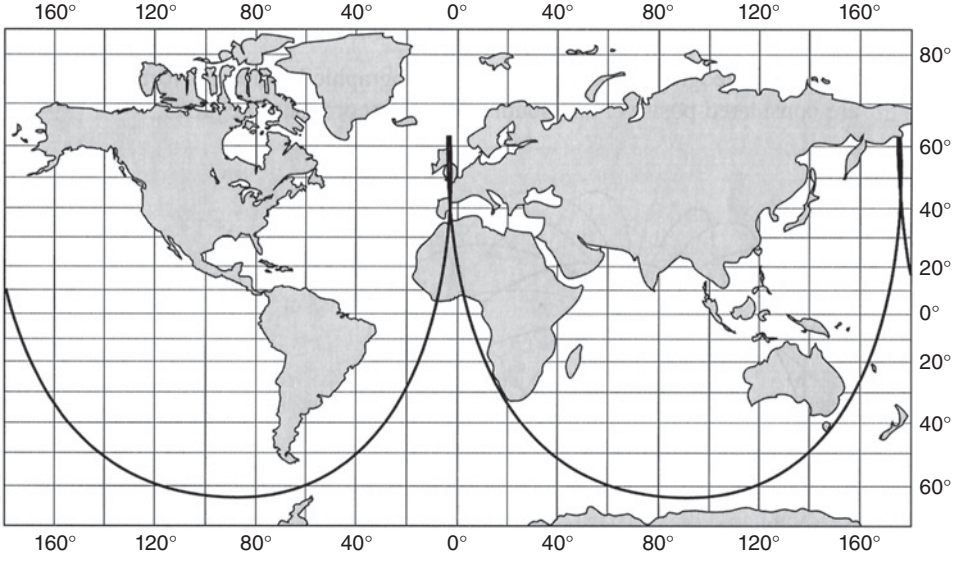


Figure 9.5 Schematic of an operational Molniya system. Satellite 1 in Molniya orbit 1 is providing service over Russia at close to its apogee while the second satellite is also close to its apogee in Molniya orbit 2. Molniya orbits 1 and 2 are separated by 180° in their orbital planes. By the time satellite 2 has moved around its orbit once and back to its apogee (a period of about 12 hours) the earth will have rotated by about 180° and the second satellite will be over Russia.

The Molniya orbit has another advantage for specific services intended for latitudes well away from the equator. The orbit takes the satellite so far away from the equatorial plane and the apogee is so distant from the earth that long dwell times with elevation angles close to 90° can be achieved at high latitudes on the earth. This fact was used in a proposal for a Molniya orbit to deliver MSS to automobiles. A view of the earth, with the plot of the orbit track, abstracted from this proposal is shown in Figure 9.6 (Watson private communication, 1996).

The Molniya orbit, in addition to the long delay time associated with the communications range when at apogee and the lack of continuous 24 hours contact with a single spacecraft from a fixed earth station also suffers from three drawbacks that increase the overall end-to-end costs. The first is the requirement to track the spacecraft over significant elevation and azimuth angles. The second is the need to switch communications to the other Molniya satellite – rather like a mobile radio handoff situation – when the first goes out of coverage as the other comes into coverage. Due to the wideband nature of the traffic and the large angular separation between successive Molniya satellites as seen from one earth station, this requires two, fairly large, reflector antennas at each site. In the early twenty-first century, phased array antennas still cannot provide accurate coincident tracking of both transmit and receive beams simultaneously well away from the (unsteered) electrical boresight over bandwidths that exceed a few percent of the carrier frequency and at a cost that commercial systems can accept. The third drawback to a Molniya orbit is the radiation environment that the satellite has to pass through four times a day – twice on ascent and twice on descent. While the first two drawbacks may be less of an inhibition to commercial success in the long-term with direct-to-home services when relatively cheap – and efficient – phased array antennas are available that track over the required range of look angles, the third drawback will always be a major factor.

(a)



(b)

Figure 9.6 View from above the Molniya orbit apogee showing the ground track (Watson private communication, 1996). (a) View from the apogee point of a Molniya orbit positioned at almost 0° longitude when at apogee. (b) Ground track of the Molniya orbit shown in Figure 9.6*a*. Note the two apogees in the orbit, one over close to 0° longitude and the other close to 180°. The apogee occurs at a high latitude, from which the elevation angles are well above 70° over quite a large region. With these high elevation angles, blockage of buildings would be minimized and thus allow relatively high availability for an MSS system operating to automobiles in most cities. This proposal was for a European MSS, but the apogee could occur at any longitude so that cities at high latitudes, but arbitrary longitude, could operate to an MSS satellite in a Molniya orbit.

9.2.5 Radiation Effects

The effect of radiation on electronics in space is generally separated out into two main aspects (Benedetto 1998): *Total Dose* and *Single-Event Upsets*. The total dose is simply the cumulative effect of radiation over the lifetime of the electronics in space and is mainly due to trapped electrons and protons in the Van Allen belts. (The Van Allen radiation belts are discussed later in this section). Eventually, the cumulative effect of radiation will degrade the performance of the transistor junction/chip such that it cannot be relied upon to generate the correct responses, and so on. This is particularly harmful in the computer elements that control the operation of the satellite and the payload. Single-event upsets are caused by heavy ions ejected from the sun, usually protons, impacting the circuitry at a critical point such that they deposit enough charge to induce an energy (bit) flip – that is, change an open-circuit to a closed circuit, create a logical one instead of a logical zero, etc. These single-event upsets are more critical if the bit flip is permanent, that is latch up occurs in a set position from which it cannot be changed.

The Van Allen radiation belts were named after the discoverer of these belts, James A. Van Allen (see Section 9.1.2). The radiation belts consist of high-intensity protons and electrons that are temporarily trapped in the earth's magnetic field. While the trapped electrons can have energies up to 7 MeV (7 million electron volts), the trapped protons can have energies up to 500 MeV (Benedetto 1998).

The relative motion between the liquid core of the earth and the solid mantle and outer crust above generates the earth's magnetic field. The magnetic field lines stretch out around the earth as shown schematically in Figure 9.7. While generally symmetrical close to the earth, the magnetic field lines of the earth become distorted further out from the earth due to interaction with the energy flowing toward the earth from the sun. The boundary where the solar atmosphere and the earth's magnetic field meet far out in space is called the bow shock, much like the pressure waves concentrating in front of the wing of an aircraft. Since the earth's magnetic and geographic poles are not coincident, the magnetic equator (and magnetic latitudes) will be different from the geographic equator (and geographic latitudes). The geomagnetic latitude Φ can be computed from (ITU-R 1986)

$$\phi = \arcsin[\sin\alpha \sin 78.5^{\circ} + \cos\alpha \cos 78.5^{\circ}\cos(69^{\circ} + \beta)] \tag{9.5}$$

where α = geographic latitude and β = geographic longitude. North and east coordinates are considered positive, and south and west coordinates negative.

The electrons and protons become ensnared in the earth's magnetic field when their kinetic energy cannot overcome the trapping effect of magnetic lines of force at the given point of encounter in space. Since the magnetic field strength decreases with increase in altitude on a given radial from the center of the earth, only the electrons are trapped in the higher reaches of the earth's environs (>10 000 km altitude above the earth) since the field forces are relatively low at these altitudes. Both electrons and the higher energy protons are trapped lower down in the earth's atmosphere ~200–10 000 km (Benedetto 1998), where the field is relatively more intense. The radiation levels induced by the electrons and protons fluctuate wildly with latitude, longitude, altitude, and with the sunspot cycle.

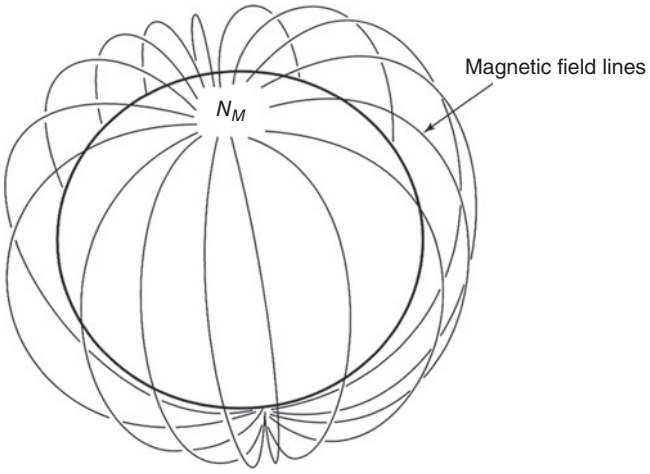


Figure 9.7 Representation of the magnetic field lines that flow between the north and south magnetic poles of the earth. The earth has a strong magnetic field due to having a liquid core that is spinning at a different rate than the solidified outer shell. The magnetic poles, however, are not coincident with the geomagnetic poles and so the magnetic equator is not located in the same position as the geographical equator. Sometimes the geomagnetic latitudes are referred to as *dip latitudes* since they will correspond to the dip in the magnetic field at that point. N_M is the north magnetic pole in the above figure.

Sunspots are disturbances on the surface of the sun. Sunspots appear to generate huge outflows of energy from the sun and the amount of energy closely follows the number of sunspots – or rather groups of sunspots – which can be counted on the surface of the sun. The sunspot count, and hence the level of energy, varies with a mean period of about 11 years, although the actual cycle spans a 22-year Hale cycle as the magnetic field lines associated with the sunspot activity on the sun's surface reverse every 11 years. The 11-year sunspot cycle period is not constant. The period has been as short as 9.5 years and as long as 12.5 years (Mursala and Ulich 1998). The first cycle that has been given an official number is the 1755–1766 period. The last full solar cycle of the twentieth century (1986.8–1996.4) was labeled Cycle #22. The first full solar cycle of the twenty-first century was Cycle #24, which started in December 2008, peaked in April 2014, and will end in 2020. A schematic of Cycle #22 is given in Figure 9.8. Cycles #23, #24, and #25 will have increasingly fewer sunspots, but they are expected to show the same large variation in sunspot count as shown in Figure 9.8.

The variability of the sunspot cycle leads to large fluctuations in the radiation environment in space. While there are large variations in the radiation environment with latitude, height above the earth, and with orbital inclination, it is normally considered that there are two main Van Allen radiation belts where the effect is more concentrated. The center of the first belt is at a height of about 1500 km above the earth and the second at around 15 000 km, measured around the equator, although these distances are somewhat arbitrary and there is some evidence that the outer belt may actually be two merged belts. The belts can be considered as doughnut-shaped, with the energy at its highest toward the center of the given belt. Figure 9.9 illustrates the concept. The trapped electrons and protons travel northward and southward along the magnetic field lines shown

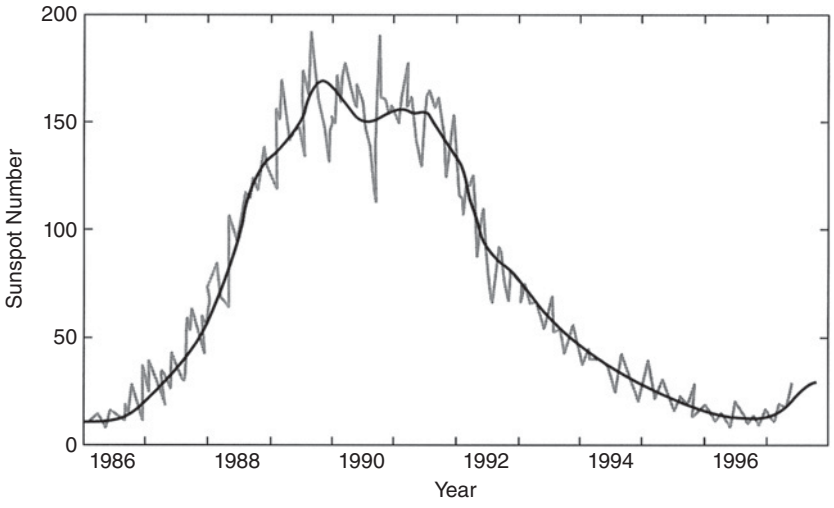


Figure 9.8 The general variation of the sunspot number over solar cycle 22. The smoothed sunspot number is averaged over several months. The fluctuations in the actual sunspot number are shown about the smoothed average. Not only does the sunspot count vary widely from month to month, it does so also from day to day. The higher the average sunspot number is, the larger the variation in actual sunspot number count is in general. Note the rapid rise then decline in the average sunspot number count and the fairly long period when the sunspot activity was very high. Because of the flat nature of the sunspot maximum period (up to four years) it is usual to determine the sunspot period from their minima.

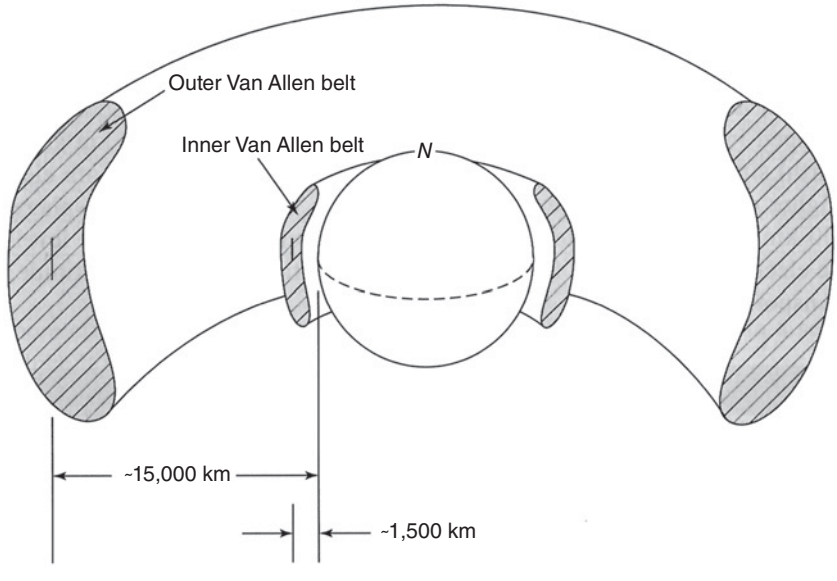


Figure 9.9 Pictorial representation of the two Van Allen radiation belts. The above schematic is a vertical slice through the radiation belts that exist around the equator. The shaded areas are the regions in the two belts where the radiation is a maximum. The two principal NGSO regions lie under the first Van Allen radiation belt – the low earth orbit or LEO region – and between the two radiation belts – the medium earth orbit or MEO region – so as to avoid the highest radiation doses. However, radiation never falls to zero and exists in all areas (see Benedetto 1998).

Table 9.2 Typical total radiation doses for various orbits

Orbital altitude (km)	800	1100	2000
Polar orbit (90°)	30 krad(Si)	100 krad(Si)	>500 krad(Si)
Equatorial orbit (0°)			>2000 krad(Si)

The data are based on a 10-year mission life using silica-based electronics in a satellite with a 2.5 mm thick aluminum skin.
Source: Data extracted from the text of ANSI/IEEE, 1992.

in Figure 9.7. They are reflected when they are close to the magnetic poles (Benedetto 1998) and so, statistically, spend more of their time closer to the equator than the poles, hence the Van Allen belts are positioned around the geomagnetic equator. The closer to the center of the radiation belts a satellite is positioned and the longer it is in space, the higher the total radiation dose becomes.

Total dosage for semiconductors that are fabricated using silicon is measured with a unit called the krad(Si). A rad(Si) is a unit of energy absorbed by silicon from radiation and it is equivalent to 0.01 J/kg (Benedetto 1998). Radiation in near-earth space is highly variable. It changes both with height above the earth and with the inclination of the orbit with respect to the equatorial plane. Since the radiation is concentrated at the equator, satellites that are in equatorial orbits will receive a higher dosage than those that are in polar orbits. In a like manner, as the orbital height moves from very close to the earth (300 km) outward for the first few thousand kilometers, the radiation dose will increase. Table 9.2 gives some typical examples of total radiation dosage for a LEO satellite designed for a 10-year operational lifetime.

Choosing an orbit that has a reduced level of radiation can therefore reduce the potential for radiation damage. Where this is not possible, then either radiation hardened (rad-hard) devices must be selected for the satellite or suitable shielding employed. Both are expensive options, the former because of the fabrication costs and the latter because radiation shields can be heavy and are non-productive elements of the payload. Developing electronic devices that can withstand total radiation doses of 1 Mrad(Si) is possible with rad-hard technologies but newer techniques for approaching these levels of radiation hardening of devices will be needed for constellations of dozens of satellites. Relatively low cost production processes have been shown to provide consistent shielding to 100 krad(Si) total dosage (Benedetto 1998) and it is likely that such techniques, plus local site shielding with aluminum strips, will be largely employed for the foreseeable future. The same approach is being used for protection against single-event upsets.

With the ever-smaller integrated circuits (ICs) being developed for flight operations, there is an increased likelihood that the linear energy transfer (LET) that is generated by the heavy ion collision will cause a single-event upset. The potential for an upset depends on the LET generated and the threshold level at which the device will incur a single-event upset. Many space-bound integrated circuits (ICs) have LET thresholds greater than 37 Mev-cm^2/mg (Benedetto 1998), which means that heavy ions with LETs of less than this amount will cause few single-event upsets. Fortunately, ICs can be manufactured relatively inexpensively with LET thresholds of 37 Mev-cm^2/mg and the incidence of heavy ions with a LET exceeding 37 Mev-cm^2/mg is very rare (Benedetto 1998). The potential for latch-up will also reduce as technology advances permit devices to be

used that employ a lower drive voltage. In addition, many of the silicon-on-insulator (SOI) and silicon-on-sapphire (SOS) technologies have been found to be immune to latch-up, as there are no parasitic paths (Benedetto 1998). The increasing density of integrated circuits has made the task of radiation hardening even more difficult (ESA 2011). The secondary particles that are given off when a high energy particle hits the aluminum shield around a device can lead to a cascading effect (ESA 2011). However, shielding technology has significantly improved, which is encouraging, as there are some satellites that will require a unique type of orbit that cannot be selected from a radiation dosage perspective. One such orbit is the sun synchronous orbit.

9.2.6 Sun Synchronous Orbit

A sun synchronous orbit is a special form of LEO where the plane of the orbit maintains a constant aspect angle with the direction to the sun. Some satellite missions require a specific orbit with such a constant relation to the direction of the sunlight. One example is an earth resources satellite that requires a large amount of direct sunlight to illuminate the region below the satellite so that photographs can be taken. This satellite would be in orbit B in Figure 9.10. Another example of a satellite needing this same orbit is a meteorological satellite, where images of the clouds and their directions of motion are critical in developing forecasts. While communications satellites have returned the greatest tangible investment returns for their owners, it is arguable that meteorological satellites have led directly to huge savings in human life, as well as to less property damage and farm animal destruction in extreme weather situations. The early meteorological satellites (e.g., TIROS) were in LEO sun synchronous orbits, but all recent meteorological satellites are in GEO orbits to provide more instantaneous and continuous coverage. Other satellites that employ sun synchronous orbits are surveillance satellites.

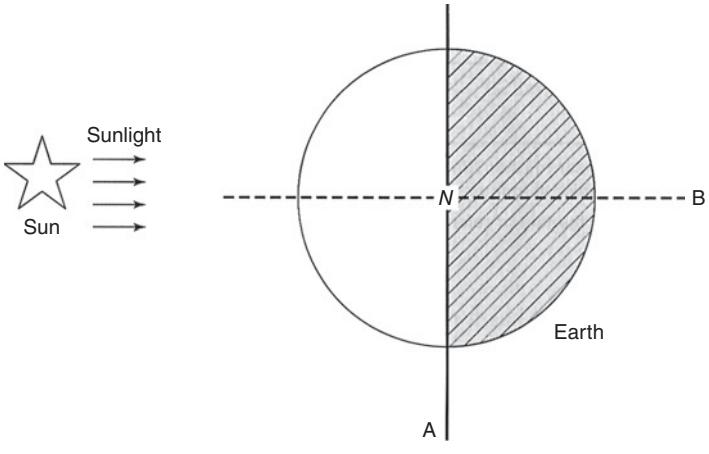


Figure 9.10 Examples of two sun synchronous orbits. In the illustration above, the earth is viewed from above the north pole, *N*, with the sunlight illuminating the left side of the earth. Two sun synchronous orbits are shown. Orbit B is designed so that it will always have one half of the orbit with the sun almost directly behind it; orbit A is designed to be always within sight of the sun – the so-called sunset-sunrise orbit since it will always be orbiting over the *terminator*. The terminator is the line that divides night from day in the figure above.

Some surveillance satellites use Orbit B of Figure 9.10, so that the maximum illumination is provided once per orbit. Others use orbit A of Figure 9.10. This particular sunset-sunrise orbit always has the satellite illuminated by the sun while the region below it has the sun at almost grazing incidence. There are two advantages in this orbit. First, the satellite need not have a large battery capacity for eclipse operations since it is always illuminated. Second, since the shadows are so long in the region being surveyed, changes in terrain or structures will be immediately obvious.

Before synthetic aperture radars were orbited, the sunset-sunrise orbit was used advantageously to detect changes in terrain following natural disasters such as earthquakes. The long shadow of a German V2 rocket allowed it to be detected by a reconnaissance aircraft for the first time at Peenemunde toward the end of the Second World War. Similarly, the ill-fated USSR moon rocket was detected on its launch pad by a surveillance satellite using the shadow it cast.

Figure 9.11 illustrates how the sun synchronous orbit is achieved. If a satellite is in a perfectly circular LEO orbit over the poles of the earth, a carefully timed launch would put the orbit in such a position that the sun is directly behind the satellite on the sunward side of the first orbit. This is position A of Figure 9.11. However, a short while later, the earth will have moved in its orbit around the sun and the plane of the satellite's orbit (now in position B) will no longer be aligned with the direction of the sunlight. In order to make the satellite's orbital plane always keep pace with the apparent change in position of the sun, it must be launched into a retrograde orbit. A retrograde orbit has a velocity component in a westerly direction. In practice, a LEO satellite launched into an orbit with an inclination of close to 98° to the equator (measured counter clockwise from the equator looking east) will move the orbital plane in time to the earth's movement around the sun. Elliptical orbits with different retrograde inclinations will also yield sun synchronous orbits. The change (rotation) in the orbital plane is called precession. A key advantage of a sun synchronous orbit is that it will repeat its track every half day.

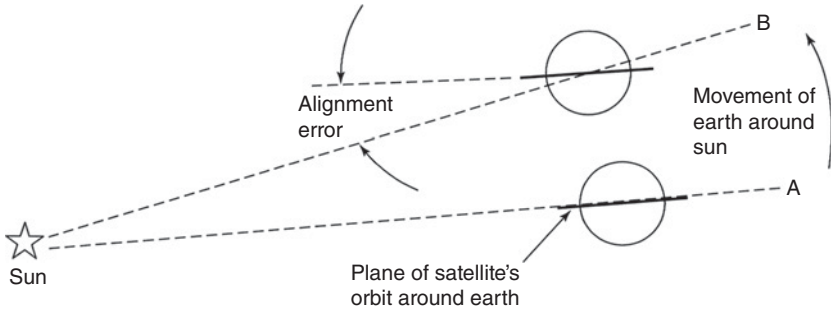


Figure 9.11 Illustration of the alignment changes of the orbital plane of a satellite due to the movement of the earth around the sun. In the figure above, a satellite has been launched into a LEO orbit (position A) in which the sun is directly overhead on the sunward side of the orbit. The view is from above the North Pole of the sun. When the earth has moved to position B, the plane of the satellite's orbit – which is fixed in inertial space – now has an alignment error with respect to the sun. If it is essential that the orbital plane of the satellite always be in line with the direction to the sun on a day-to-day basis over a long period, then the plane of the satellite's orbit will need to change at the same rate that the alignment error is increasing. That is, the plane of the orbit will need to precess to match the movement of the earth around the sun.

It can therefore be used to make measurements at given times of the day and night so that correlation exercises can be attempted.

One example of a spacecraft in a sun synchronous orbit was the Mars Explorer spacecraft, which was put into a sun synchronous orbit around Mars in 1998. The orbit was used to measure temperature at 2 a.m. and 2 p.m. local time equivalents over the same region so that local heating effects and cooling effects could be accurately tracked.

A sun synchronous orbit will pass over almost all of the earth at one time or another. Determining the instantaneous surface area of the planet seen by the satellite and over which information is required – or to which communications is to be established – is another issue. This portion of the earth's surface is called the coverage area or coverage region.

9.3 Coverage and Frequency Considerations

9.3.1 General Aspects

In some cases, the designer of a satellite system has few degrees of freedom in designing a payload to provide optimum coverage. This occurs in some missions where a shared spacecraft has to accommodate a number of payloads. Examples exist in the scientific community when (generally) low-cost, multiple missions are being developed for a single spacecraft. In other cases, more freedom exists in the design stage and the mission planners can vary a range of parameters iteratively to arrive at an optimum coverage.

The mission goals will directly determine the coverage that has to be achieved by a given satellite system. This in turn leads to the selection of orbit, payload technologies, and so on. For example, if a communications satellite system has to provide coverage of the European Union (EU), there is a minimum altitude at which a single satellite can operate and still cover all of the EU at once. If the coverage of the EU must be continuous, a GEO orbit can be selected or a constellation of NGSO satellites can be designed to provide the necessary coverage overlap between successive satellites. The determination of coverage area, while initially an exercise in simple geometry, is eventually heavily influenced by the available technology both on the ground and in space, and other aspects such as the radiation environment. We will consider first the geometrical aspects of determining an optimum coverage.

In Figure 9.12, a spacecraft orbits at distance r_s from the center of the earth, C. We will assume that the spacecraft is a communications satellite and that it needs to be in contact with an earth station located at E. The elevation angle to the satellite is θ. Using the sine rule we have

$$[r_s/\sin(90+\theta)] = [d/\sin(\gamma)] \tag{9.6}$$

which yields

$$\cos(\theta) = [r_s \sin(\gamma)]/d \tag{9.7}$$

All three parameters in Eq. (9.7) have key inputs to the architecture of the satellite system. The angle γ will yield the coverage area on the surface of the earth assuming the satellite has a symmetrical coverage about nadir. The distance d will determine the free

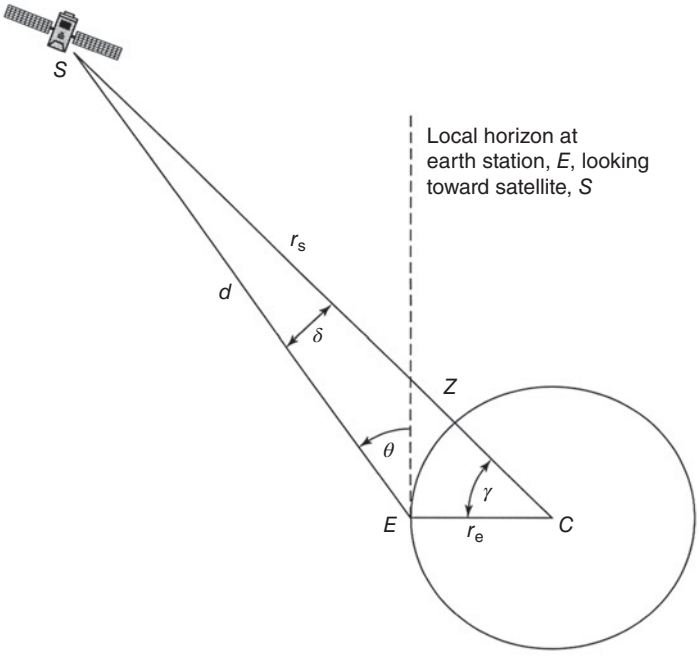


Figure 9.12 Geometry for calculating coverage area. The satellite, earth station, and the center of the earth are all in the same plane in this figure. *SCE* is the central angle, γ, and the elevation angle, θ, is the angle between the local horizon and the satellite at the earth station in the plane of the figure. The line *SC* joins the satellite and the center of the earth and cuts the surface of the earth at point *Z*. To an observer at point *Z*, the satellite is at zenith. The satellite is at a distance, *d*, from the earth station and a distance r_s from the center of the earth. The radius of the earth is given as r_e, a good average value for which is 6370 km.

space path loss along the propagation path, and will be a factor in the link budget design. The elevation angle θ will influence the G/T ratio of the antenna, the blockage probability from terrain and buildings near the antenna, and the likely propagation impairments that will be encountered along the path to the satellite. For systems that operate in frequency bands that suffer significant degradations in rain, the elevation angle can be the critical design element (see Chapter 7 for more details on propagation effects along earth-space paths).

9.3.2 Frequency Band

LEO satellite systems providing data and voice service to mobile users tend to use the lowest available RF frequency. The effective isotropically radiated power (EIRP) required by the satellite transponder to establish a given C/N ratio in the mobile receiver is proportional to the square of the RF frequency of the downlink, as the analysis in the next paragraph shows. The power that must be transmitted by a mobile transmitter is also proportional to RF frequency squared when the mobile uses an omnidirectional antenna. Since the cost of satellites increases as the EIRP of the transponders increases, a lower RF frequency yields a lower cost system. This is one reason why L-band is allocated for MSSs.

Consider a LEO satellite with a coverage zone on the earth's surface that has an area A m^2. A transponder on the satellite with output power P_t watts drives an antenna with a gain G_t to produce an EIRP from the satellite of P_tG_t watts. The average flux density, F across the coverage zone is therefore

$$F = P_tG_t/A \text{ watts/m}^2 \tag{9.8}$$

The value of the flux density is independent of frequency. The mobile receiver has an antenna that is omnidirectional, with a gain $G_r = 1$. The effective receiving area of this antenna is given by

$$A_e = \lambda^2/4\pi \tag{9.9}$$

The received power at the mobile earth station is given by $P_r = F \times A$, hence

$$P_r = \frac{P_tG_t\lambda^2}{4\pi A} \text{ watts} \tag{9.10}$$

Thus the received power at the mobile with an omnidirectional antenna increases as the square of the wavelength, or decreases as the square of the frequency. The lower the RF frequency, the greater the received power for any given coverage zone. By reciprocity, the same result will apply when the mobile terminal transmits with an omnidirectional antenna. It therefore makes sense for mobile systems, which are forced to use omnidirectional antennas so as to avoid having to steer a directional antenna, to use the lowest possible RF frequency. That is why Orbcomm's data relay LEO satellite system uses very high frequency (VHF) (30–300 MHz) and ultra high frequency (UHF) (300–3000 MHz) frequencies. Orbcomm satellites have a single transmit beam at the satellite that serves the entire coverage zone. For the same reasons, L-band (1–2 GHz) is allocated for MSS, but to achieve similar C/N ratios with L-band links as Orbcomm satellites achieve with VHF links, the L-band satellite must provide multiple beams from a high gain antenna.

One disadvantage of VHF and UHF frequency bands is a high noise power due to the natural environment. For this reason, the antenna noise temperature for a system operating at VHF or UHF will be much higher than the receiver noise temperature. Environmental noise temperature falls with increasing frequency; by L-band it is not a significant factor.

The worst possible choice of frequency for a mobile system is Ka-band (about 20–30 GHz), or above. A Ka-band mobile downlink operating at 20 GHz requires 22.5 dB more transmitted EIRP, or receiving antenna gain, than the same system operating at 1.5 GHz. Occasionally proposals are aired for Ka band mobile systems, but such a system can only succeed with a steered directional antenna on the mobile terminal. Conventional mechanically steered Ka band dishes are expensive – thousands of dollars more than a simple whip antenna, and self steering phased arrays are currently even more costly.

It is worth noting in passing that it is the omnidirectional antenna of a mobile terminal that drives up the cost of every transmitted bit in a mobile system. Suppose that a fixed terminal with an antenna of gain G_{rx} supports a bit rate of R_b bits per second. A mobile terminal with a low gain antenna, gain G_{rm}, and the same satellite EIRP and path loss can support a much lower data rate of $R_b \times (G_{rm}/G_{rx})$. For example, an 18 in. DBS-TV antenna operating at 12.5 GHz has a gain of 33 dB. A satellite link using the DBS-TV antenna can receive data at 2000 times the rate of a mobile terminal that has an omnidirectional antenna with a gain of 0 dB, with the same overall C/N value in the receiver.

Given equal costs for the space segment of the communications link, and a mobile system operating in Ku-band, a system operator must charge the mobile user 2000 times as much per delivered bit compared with the DBS-TV terminal owner. Looked at another way, for the same monthly fee, the DBS-TV customer can receive signals at 20 Mbps, equivalent to several compressed digital television signals, while the mobile terminal customer can receive only 10 kbps – a single voice channel.

Antenna gain is the system designer's friend. Mobile systems will become much more attractive economically when a self steering, self phasing, phased array is available for mobile terminals with even a moderate gain. A 10 dB increase in antenna gain translates directly to a 10-fold increase in bit rates.

9.3.3 Elevation Angle Considerations

As we have seen in Chapter 7, rain attenuation can cause significant attenuation on a slant path. At Ka-band (30/20 GHz), even light rain can cause appreciable signal loss. Light rain is usually stratified and so, the higher the elevation angle the lower the rain attenuation for a given rainfall rate. Figure 9.13 illustrates the geometry.

Most commercial satellite systems require that earth stations operate above certain minimum elevation angles. For example, INTELSAT requires that all earth stations using INTELSAT C-Band (6/4 GHz) satellites operate above 5°, otherwise the earth station does not meet INTELSAT's standard specification and must be qualified for operation on an individual basis. To qualify an earth station on an individual basis is an expensive undertaking. At Ku-band (14/11, 14/12 GHz) the standard antennas in the INTELSAT system are required to operate above a minimum elevation angle of 10°. In creating the original Teledesic system architecture (Teledesic 2000), the overriding design input for the coverages was that no earth station should operate at an elevation angle below 40°. Teledesic was the first satellite system proposed as an *internet in the sky*. The requirement for a minimum elevation angle of 40°, when coupled with an orbital height of around 800 km, led to an unrealistic initial constellation of 840 operational satellites to provide full global coverage. Most satellite systems now, whether for the MSS or the FSS at frequencies above 10 GHz, tend to limit the elevation

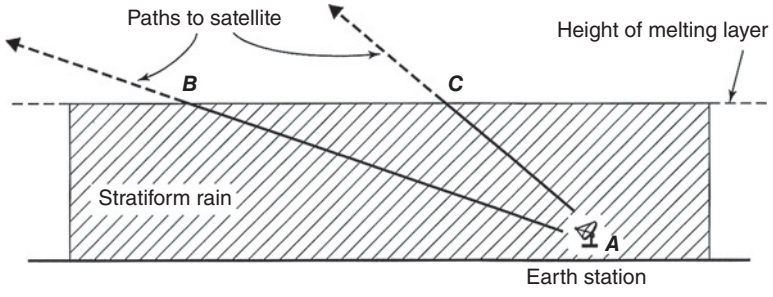


Figure 9.13 Illustration of the decrease in the path through rain as the elevation angle to the satellite increases. Light rain is generally formed at the melting layer height in stratiform clouds. The rain then falls fairly uniformly over a wide area. Freezing precipitation (hail, ice crystals, dry snow) does not cause any appreciable attenuation to radio waves. Since the rain is uniform, the attenuation per meter will be constant everywhere in the stratiform rain shower and the total path attenuation will be given by the length of the signal path in the rain. The higher elevation angle path (*AC*) will therefore suffer less attenuation than the lower elevation angle path *AB*.

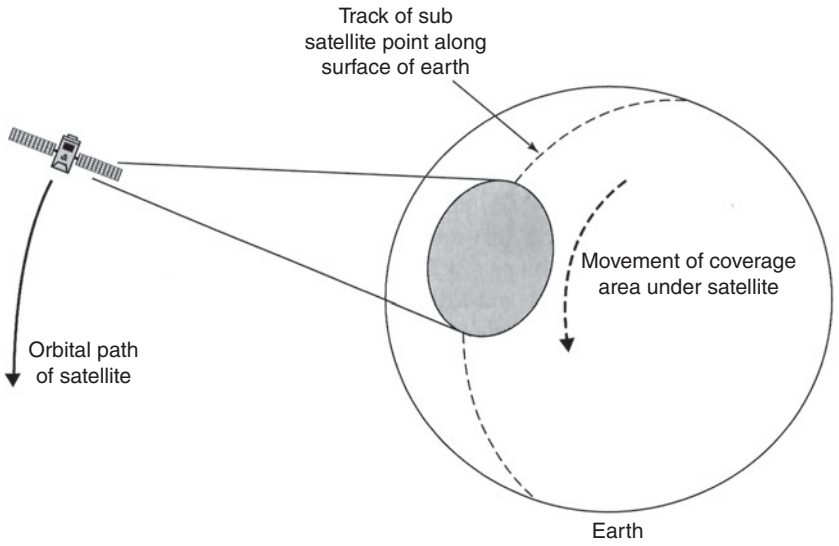


Figure 9.14 Illustration of coverage area under a satellite. In this example, an NGSO satellite moves along a path over the earth with a nadir pointing antenna, that is, the antenna has its electrical axis directed straight down toward the subsatellite point. The antenna will have a finite usable beamwidth, which will allow a given portion of the surface to be illuminated at the same time. This is shown above as the shaded portion in the figure. Increasing the altitude of the satellite's orbit will increase the coverage area. Alternatively, the altitude can remain fixed and the beamwidth increased in order to cover a larger area.

angle of the user to no less than 10° so that reliable service can be provided. Given a minimum elevation angle and an orbital height, the geometry set up in Figure 9.12 can be used to develop a coverage area, assuming that the satellite has a symmetrical beam aimed at nadir. A plot similar to that shown in Figure 9.14 results.

The shadowed area on the earth in Figure 9.14 is the maximum instantaneous coverage on the surface of the earth that can be achieved from that satellite. The calculation process for the instantaneous coverage has as input the minimum elevation angle a user can tolerate and the orbital altitude selected. (Instantaneous coverage means that, if a snapshot were taken and the motion of the satellite frozen at an instant in time, the region of the earth covered by the satellite's antenna at that particular time would be the instantaneous coverage from that satellite. The word instantaneous is used to separate out coverages that are developed by scanning or hopping satellite antenna beams. For the scanning/hopping beam concepts, full coverage is obtained by moving elemental beams around the coverage area to pick up traffic. A full, and instantaneous, coverage of the observed region is therefore not obtained with scanning/hopping beams). The instantaneous coverage from a satellite, however, is not always served by one beam from the satellite antenna due to the lack of available spectrum and a concomitant need for extensive frequency re-use. This is particularly true for MSS systems, which, like terrestrial microwave cellular systems (Cellular 2008), have to divide up their coverages into cells covered by separate beams in order to provide enough capacity into a given cellular structure. Each cell, here a separate beam from the satellite antenna, will have a portion of the spectrum allocated to it. The simplest spectrum re-use pattern is a

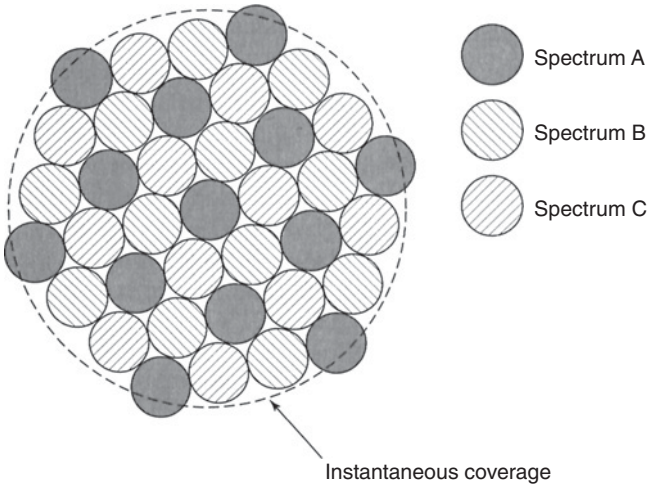


Figure 9.15 Illustration of a three-cell re-use pattern. The instantaneous coverage of the satellite antenna is shown as the circle with a broken line. Within this coverage, individual beams formed by the satellite antenna make a regular pattern that fills up the instantaneous coverage. The spectrum that has been allocated to this satellite has been divided up into three portions, called Spectrum A, Spectrum B, and Spectrum C. These different spectra are indicated with different shading. None of the three spectra are adjacent to the same spectral allocation. Note: In general, each of the individual beams will overlap their neighbors for two reasons. First, by overlapping the individual beams there are no holes in the instantaneous coverage. Second, physics will prevent the beams going from full power to zero power over a negligible distance. It is usual to develop coverages using the half-power (3 dB down) contour of the beams as the edge of coverage gain/power. There will therefore be energy spilling over into adjacent beam coverages. This is why it is necessary to employ a different spectral allocation in adjacent beams, unless a code division multiple access (CDMA) technique is used.

three-cell configuration. The spectrum is divided into three roughly equal portions and a three-cell pattern built up over the coverage area. Figure 9.15 illustrates the concept. There are many other cell re-use patterns that are possible (Cellular 2008).

9.3.4 Number of Beams per Coverage

The very small spectrum allocation available for MSS systems (<50 MHz), and the many competing systems that aim to provide MSSs, place a number of constraints on the system design. Table 9.3 shows the spectrum, antenna, and resulting capacity of the two major MSS systems: Iridium and Globalstar, known generically as big LEOs. They have now been joined two other Big MEO systems. One is O3B, which stands for the "Other 3 Billion" of dwellers on planet earth who currently do not have access to the internet. The other is OneWeb, which acquired the spectrum formerly owned by SkyBridge. It is instructive to compare the four systems below.

Figure 9.16a presents a snapshot of the multiple spot beams generated by an Iridium satellite within the instantaneous satellite coverage of one of the satellites. The ICO satellites are discussed in Section 9.5.3.

The requirement placed on the MSS satellite antenna to generate multiple beams within a given instantaneous coverage is a key driver in the payload technology. Traditional satellite antennas have evolved from simple, front-fed reflector antennas with

Table 9.3 Frequency, antenna, and capacity characteristics of the big LEOs and big MEOs

Parameter	Iridium	Globalstar	O3B	OneWeb
Frequency uplink/ downlink (GHz)	1.62135–1.6265	1.619–1.6215/ 2.4835–2.4985	27.5–30/ 17.7–20.2[a]	14.4–14.5/10.7–12.7[a] 27.5–29.1/17.8–18.6[a] 29.5–30/18.8–19.3[a]
Maximum bandwidth (MHz)	5.15	11.35	2500[a]	2500[a,b]
Spot beams per satellite	48	16	10 + 2	640
Nominal capacity per satellite	1110 voice ccts	2400 voice ccts	12 Gbps	≈6 Gbps
Orbital altitude (km)	780	1414	8062	1200
Orbit type	Polar	Inclined	Equatorial	Polar

[a]Spectrum is divided into three segments, some of which must be shared with GSO systems. Since GSO satellites were there first, O3B has to coordinate with the GSO systems, in the affected segments. Key sharing element is within ±5° of the equator. The individual bandwidth of each beam depends on the capacity. A 0.85 m earth terminal can deliver 100 Mbit/s on the uplink and receive 400 Mbit/s on the downlink.
[b]Some spectrum is only on a shared basis; first users receive protection.
Source: (Iridium 2018, Globalstar 2018). Data extracted from (Iridiumnext 2013, arstechnica 2015; skyrocket 2018b).

one feed horn, to offset-fed designs with more than 100 feeds (Terada 1999). Such multiple feed horn reflector antenna designs are necessarily large and heavy. The greater the number of individual beams to be generated, the heavier the reflector antenna and associated feed horns and beam forming network. Depending on the precise spacecraft mission, there is a threshold where the cost and complexity of a phased array antenna implementation will be less than that of the equivalent reflector antenna.

A phased array antenna usually has a non-mechanically steered array of radiators. The radiating elements can be passive devices (e.g., dipoles or feed horns) or active devices (e.g., patch elements, which include amplifiers). The steering of the beam is carried out by varying the phase (and amplitude for full sidelobe control) of the signal in each radiating element. For a passive device, the phase control is achieved in the feed matrix placed between the high power amplifier (HPA) and the radiating antenna elements while, for the active device, there is a phase shifter per element per beam. In many cases, it is possible to include the amplifier as part of the active phased array radiating element. This particular phased array concept is referred to as a direct radiating phased array. Figure 9.17 illustrates the two phased array approaches. In either approach, the scan angle is often a critical design limitation.

9.3.5 Off-Axis Scanning

The design of a point-to-point wireless communications system requires that the antennas at either end be directed toward each other for maximum gain advantage. This was the approach adopted for the fixed service (FS), the terrestrial microwave communications service. If the transmitting antenna has to communicate with more

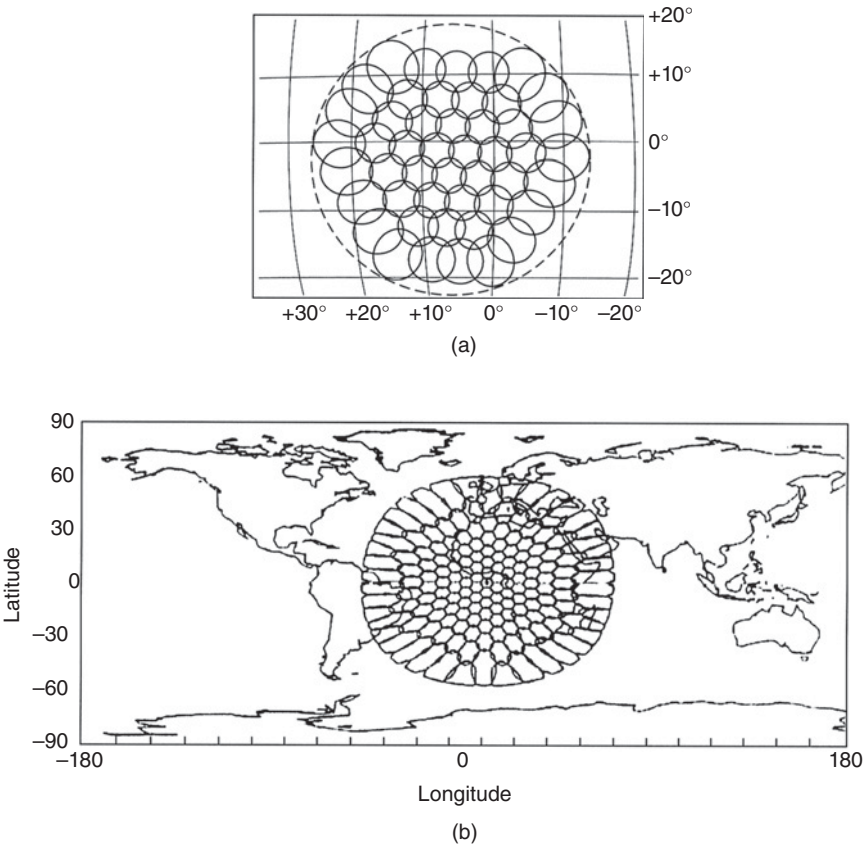


Figure 9.16 (a) User spot beams developed by an Iridium satellite. Source: Figure 11 of Evans (1997), reproduced with permission. The satellite covers about 40° of the earth's surface from its orbital height of 800 km, which translates into about 4000 km diameter main coverage. This coverage is divided up into 48 spot beams. Each of the spot beams has the same beamwidth, but the curvature of the earth has caused the outer spot beams to appear elliptical. The boundary of each spot beam denotes the -3 dB contour of that spot beam. (b) User spot beams developed by an ICO-Global satellite. Source: Figure 20 Evans 1997 reproduced with permission. The satellite covers about 110° of the earth's surface from an orbital height of 10, 355 km, which translates into about a 12 000 km diameter coverage. This coverage is divided up into 163 spot beams. Each of the spot beams has the same beamwidth, but the curvature of the earth has caused the outer spot beams to appear elliptical. The boundary of each spot beam denotes the -3 dB contour of that spot beam. Note that the spot beam size of the ICO-Global satellite is similar to that of the Iridium satellites.

than one receiving antenna, and these antennas are located in different positions, a compromise must be reached between the gain of the transmitting antenna toward the various receiving antennas. In this case, most, if not all, of the receiving antennas will not be on the boresight (main beam axis) of the transmitting antenna. Figure 9.18 illustrates the problem.

Exactly the same design compromise illustrated in Figure 9.18b faces satellite system designers who have to provide coverage over a large instantaneous area from a single satellite. A satellite is a prime example of a point-to-multipoint system. There are two basic input geometrical parameters that are used in the initial design phase of a satellite

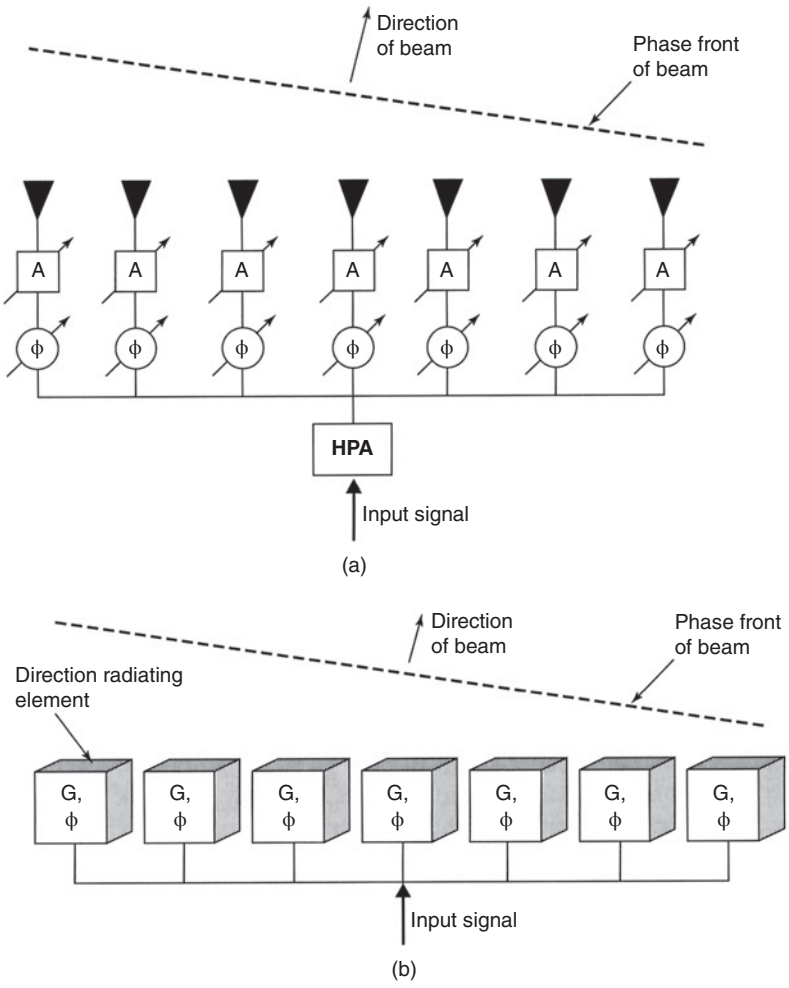


Figure 9.17 Illustration of scan angle control mechanisms for phased array antennas. (a) Passive phased array (b) Direct radiating array. In (a), the high power amplifier (HPA) has had the power divided up among a number of different feed lines. Each feed line is acted on by a variable phase-change (φ) and a variable attenuator (A) device. The resultant output signal is then fed to a passive feed horn. The sum of the many phases and amplitudes generated by the feed horn cluster will develop the antenna coverage. In (b), the phase and amplitude are controlled by the direct radiating device at the end of the feed line. The amplitude is controlled by the gain of the radiating amplifier, G, and the phase, φ, can either be controlled within the amplifier unit itself or by a phase element associated with the radiating device. To develop a large number of beams, many signal lines will feed each element and a complex phase front for that signal; each beam shape will be given by the number of individual elements contributing to the development of the phase front.

antenna: the orbital height and the instantaneous coverage requirements for a single satellite. Figure 9.19 presents the three main design options for orbital altitude: LEO, MEO, or GEO, and Table 9.4 lists some scan angle requirements for various satellite altitudes, with atmospheric refraction ignored.

A fixed antenna with a parabolic reflector is able to scan its main beam away from the electrical boresight axis by repositioning the feed transversely from the prime focus.

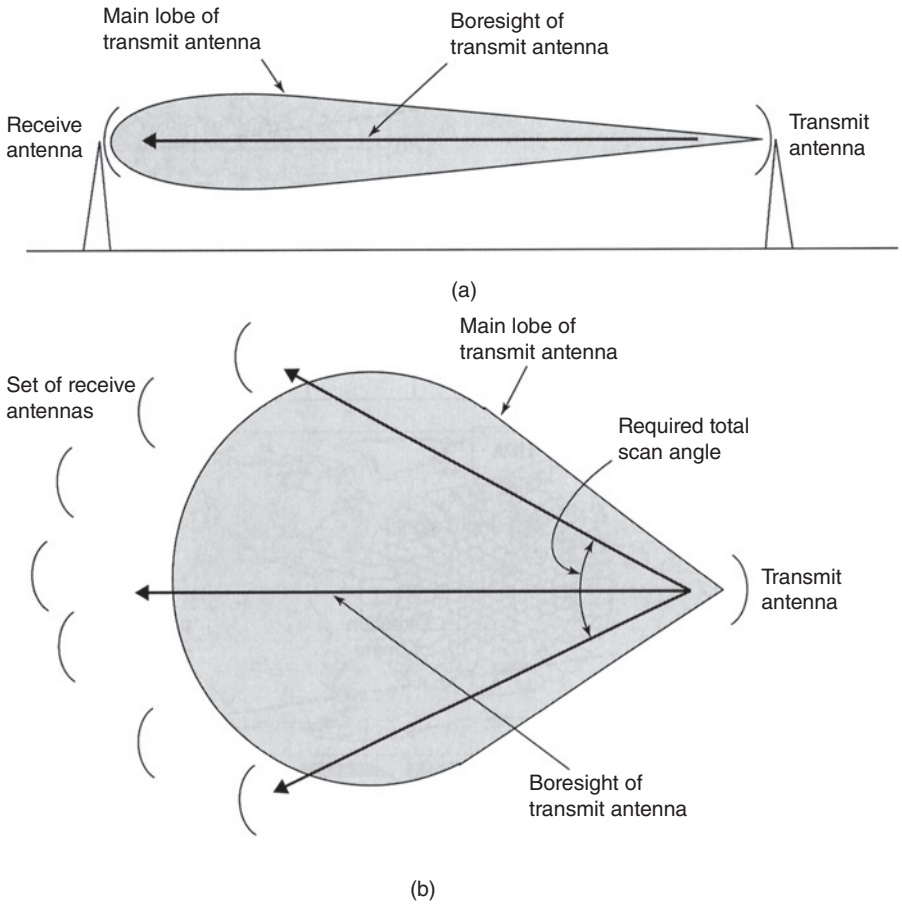


Figure 9.18 (a) Point-to-point line-of-sight terrestrial communications link. The transmit antenna illuminates the receive antenna along its electrical boresight, providing the maximum gain for the link. The receive antenna has its electrical boresight directed toward the transmit antenna (not shown explicitly here). The transmit and receive gains are therefore maximized. (b) Point-to-multipoint line of sight terrestrial communications link. In this plan view of a point-to-multipoint system, one example of which is called LMDS – local multipoint delivery system – the transmit antenna has to cover a number of receive antennas spread over a large scan angle. There are two main options available to the link designer: use a single wide-angle beam to cover the receive antennas (as has been illustrated here); or set up several different transmit antennas, each directed toward a given receive antenna. In this latter concept, the function of the transmit antenna can be divided up amongst a small group of antennas that provide 360° coverage, as in the sectored antenna approach of cellular systems.

However, the plane wave that is present in the aperture of a focused parabolic reflector antenna becomes distorted when the feed horn is moved away from the focus, resulting in an effect known as coma. Coma causes a reduction in antenna gain, an increase in side-lobe levels, and an increase in cross polarization. The reduction in gain and polarization purity can be held to relatively small values if the focal length, f, of the antenna is long with respect to the antenna diameter, D, and the off-axis scan angle is small. A value of $f/D \geq 1$ is generally taken as the required design goal, and is usually implemented by employing a double reflector configuration such as the Cassegrain

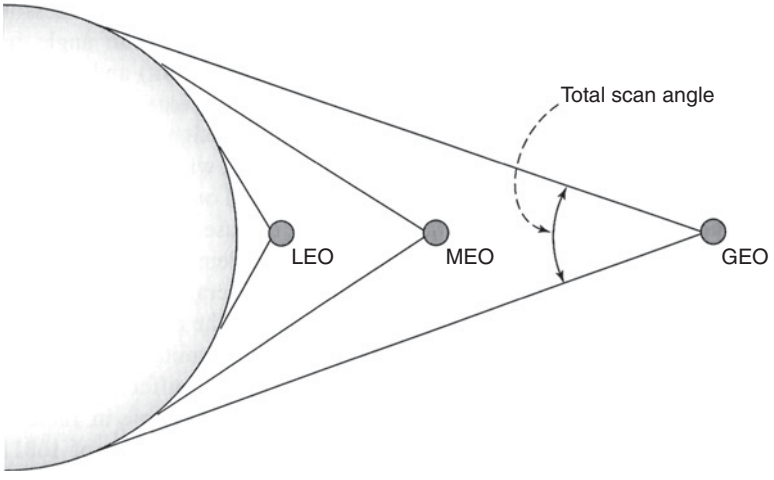


Figure 9.19 Schematic of the total scan angles for LEO, MEO, and GEO satellites. The further away the satellite is from the earth, the smaller the satellite scan angle needed to provide an instantaneous coverage out to a given user elevation angle minimum. In the above figure, the satellites are all in an equatorial orbit. The view is from above the earth with one side of the earth in sunlight and the other in darkness. The terminator is directly under all three satellites.

antenna (see Appendix B). Cassegrain antennas have a large equivalent f/D ratio while being mechanically compact. GEO satellite antenna designs that scan over the full earth coverage ($\pm 8.7°$, which corresponds to 0° elevation angle at the extreme earth station locations) have been successfully implemented using reflector antenna technology. However, as the scan angle requirement increases (due to the satellite moving from GEO to a lower orbit) and the number of individual beams that are needed within the instantaneous coverage grows large, phased array antennas prove to be the best match to the system design (Evans 1997). Phased array antennas are used on the Iridium and Globalstar satellites, while Cassegrain or Gregorian (another double reflector configuration) antennas are used on many large GEO satellites.

A number of factors influence the coverage of a phased array antenna from a given satellite. While the phase of the signal in the radiating elements determines the steering of the beam, it is usual to have the main beam axis (generally normal to the surface of

Table 9.4 Scan angle and latitude/longitude ranges for different satellite altitudes

Orbit	LEO		MEO		GEO
Orbital height (km)	750	1800	10 000	14 000	35 786
Scan angle	±57.2°	±47.1°	±21.5°	±17.1°	±8.25°
Latitude/longitude range	±12.8°	±22.9°	±48.5°	±52.9°	±61.8°

The minimum elevation angle to the user terminal for the data in this table is 20°. The coverage is assumed to be a cone of revolution around a nadir pointing direction. Note that, even though the scan angle is smaller for satellites at higher altitudes, the latitude (and longitude) coverage increases with altitude. Thus, for a given scan angle, instantaneous coverage increases with satellite altitude.

Source: Some of the data have been extracted from (Globalstar 2018).

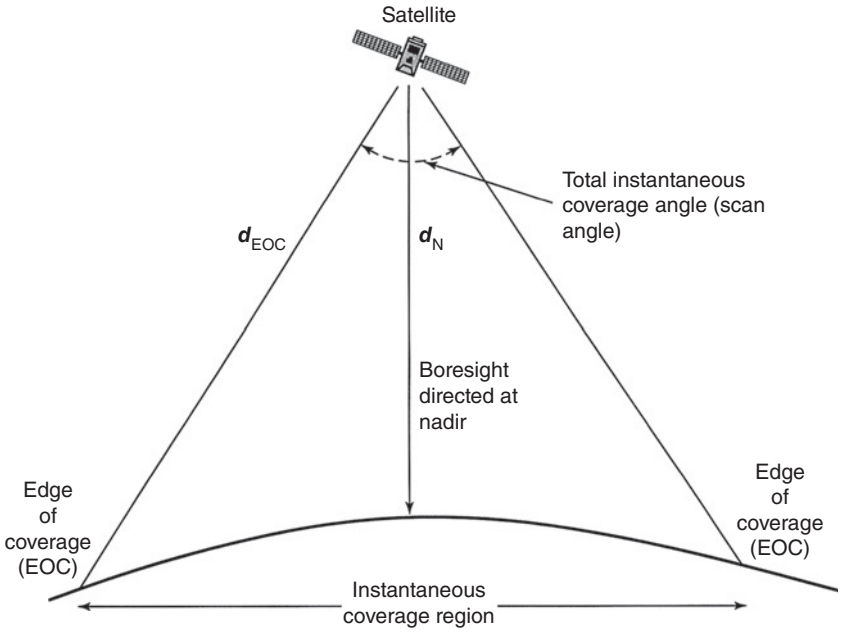


Figure 9.20 Illustration of path loss and scan angle loss evaluation for a phased array. The phased array has as its prime axis pointed at nadir. The energy received at nadir from the satellite will be greater than that received at edge of coverage (EOC) for two reasons. First, the path loss will be less since the nadir distance, d_N, is less than the EOC distance, d_{EOC}. Second, there will be a scan loss associated with the satellite antenna reaching out to cover the EOC region.

the antenna array panel) directed at nadir. This will lead to the edge of coverage (EOC) users suffering two loss components that are larger than for a signal transmitted in the nadir direction. First, they will be further away from the satellite, and so will suffer a free space path loss that is increased with respect to a nadir user. For the LEO example in Table 9.4 that is at an altitude of 750 km, the 57.2° scan angle leads to a slant range of 1681 km at the 20° EOC. The difference in path loss between the range at nadir (750 km) and EOC (1681 km) is 7.0 dB. The 1800 km orbit LEO satellite has a 5.5 dB path loss difference between nadir and EOC. For mobile satellite systems that must operate with this rapid variation in path loss as the satellite passes by the user, power control is employed to offset changes in perceived power level at both the satellite and the earth terminal (the handset). The second higher loss component experienced by EOC users is that the satellite antenna will incur a scan loss as it attempts to direct energy away from the main beam axis (the boresight directed at nadir) out to the edge-of-coverage user. Figure 9.20 illustrates the change in path loss with scan angle.

Scan loss for a phased array antenna normally follows the relationship (Schuss et al. 1999)

$$\text{scan loss} = -(\text{cosine}\,\phi)^k \tag{9.11}$$

where ϕ is the scan angle off boresight and k is an empirical number between 1.2 and 1.5. Note that the negative sign in Eq. (9.11), which was not incorporated in the referenced

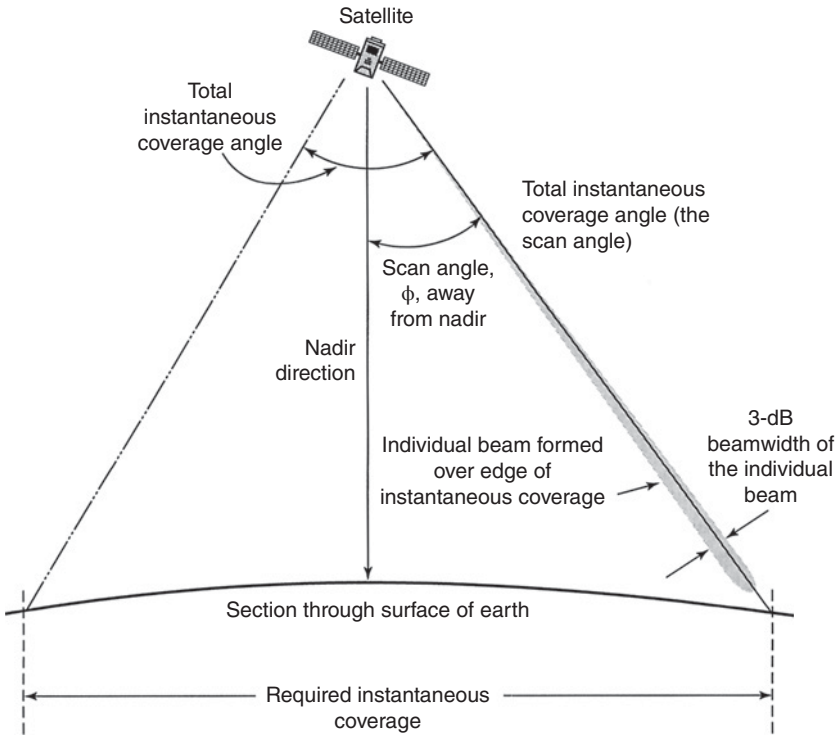


Figure 9.21 Illustration of the scan angle of an individual beam within an instantaneous coverage. The instantaneous coverage is developed through many smaller beams spread over the region to provide sufficient frequency reuse for the users in that area. Only one of the small, individual beams is shown above as a shade area on the right. This beam is scanned to the edge of the instantaneous coverage. Note that, a user *within* the small, individual beam will have to factor two components into the link budget: (i) the gain loss due to not being at the center of the individual beam; and (ii) the scan loss due to not being at the nadir point of the instantaneous coverage region (here assumed to be the boresight of the phased array in the satellite).

article, allows the sign on both sides of the equation to agree (see the example in Eq. (9.12) below). Figure 9.21 illustrates the geometry for Eq. (9.11).

A typical value to use for the parameter k is 1.3 (Schuss et al. 1999). For example, a LEO system that needs to scan 57.2° away from boresight will have a scan loss

$$\text{Scan loss} = -(\text{cosine } 57.2)^{1.3} = 0.4507 \Rightarrow 3.5 \text{ dB} \tag{9.12}$$

Thus the scan loss is 3.5 dB for a beam transmitted to the edge of an instantaneous coverage of ±57.2°. The EOC path will also suffer an additional path loss compared with the nadir path of 7 dB for a LEO satellite orbiting at a height of 750 km. The EOC signal is therefore 10.5 dB below the nadir signal in this example. To counteract these two loss components – scan loss and enhanced path loss – the phased array boresight can be redirected toward the EOC. The problem with this approach is that at least three phased array panels are required on the spacecraft to illuminate the full instantaneous coverage. With three antenna array panels, the scan loss is reduced by at least 1.5 dB (Schuss et al. 1999). This solution was adopted by Iridium. The three phased array antenna panels can be seen clearly in Figure 9.22 (Iridium) and (Schuss et al. 1999). The second generation

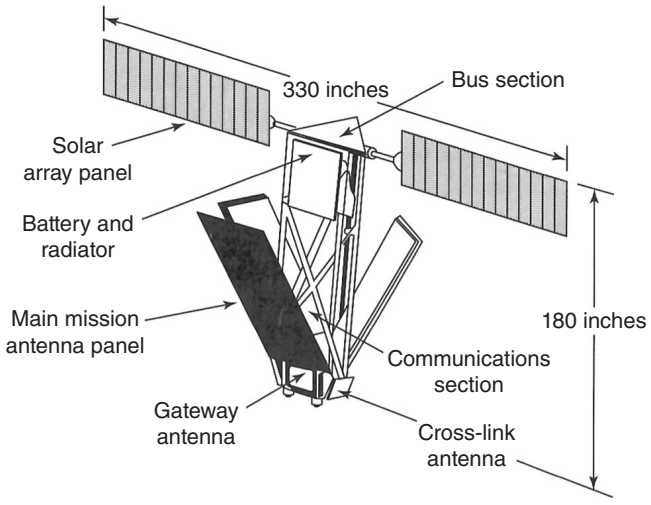


Figure 9.22 A sketch of an Iridium satellite. Source: Figure 10 of Evans 1997, reproduced with permission of Motorola. One of the three phased array antennas is shown as the main mission antenna panel in the above figure. The instantaneous coverage is developed using these three phased array antennas, much like a three sector microwave cellular coverage within a cell. Iridium uses an FDMA/TDMA multiple access technique. Communications can be established from two of the antennas into an area that is on the joint between the two sector coverages, thus the signals must be accurately controlled in time so that they arrive at all three antenna array panels at the same instant and the TDMA bursts do not overlap in time. Not shown clearly in this figure are the four ISL antennas that communicate with the other satellites in the constellation. There are two antennas in the plane of the orbit, one directed northward and one southward; there are also two antennas used to cross-link to the side, one to the east and one to the west. Iridium ISL links are only possible with satellites that are moving in the same general direction. That is north-going satellites cannot establish an ISL link with a south-going satellite.

Iridium satellites incorporate an advanced phased array that, while it still provides the 48 individual beams of the first series of spacecraft, achieves this with a single, flat array panel (skyrocket 2018b).

In addition to the required antenna scan angle, the height of the orbit is the other key geometrical parameter that influences the design of a LEO constellation.

9.3.6 Determination of Optimum Orbital Altitude

The earth station locations at the EOC within the instantaneous coverage region normally present the greatest problems in the design of a satellite service. It is at the EOC that the power flux density into the user terminal is at its lowest. Even if the user is at the center of the individual beam that serves the EOC (see Figure 9.21) there are still the two additional factors that determine whether the link can provide adequate service: the scan loss and the added free space path loss at EOC when compared with nadir. Minimizing the total additional loss in the transmission path to EOC is a design goal. If the orbital altitude is increased, the free space path loss will increase, but the scan loss will decrease. For a LEO constellation of satellites with a given large instantaneous coverage requirement, as the orbital altitude increases from a minimum of 500 km, the scan loss will decrease faster than the path loss increases. There will therefore be an optimum altitude for a LEO constellation, based upon the number of satellites per plane and whether

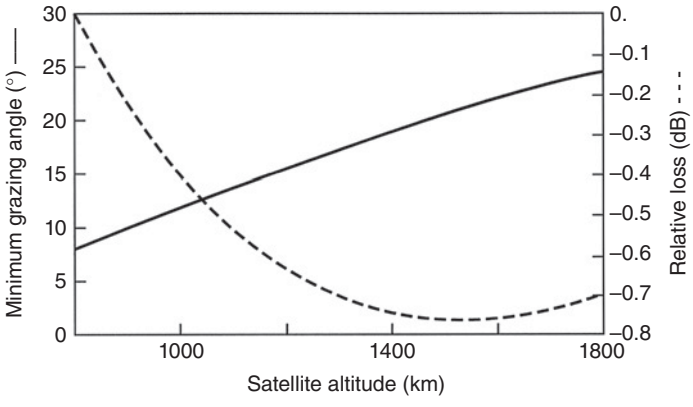


Figure 9.23 Relative transmission loss and minimum grazing angle vs. satellite altitude for a constellation of LEO satellites with 10 satellites per plane and double coverage. Source: Figure 1 of Chiavacci 1999, reproduced with permission of Microwave Journal. In the calculations for this figure, it was assumed that there would always be two satellites in view for any user. With 10 satellites per plane, a minimum scan angle at the satellite is calculated and, from this, the elevation angle for the edge-of-coverage user is found. This angle is referred to as the *minimum grazing angle* in the figure. The scan loss + free space path loss for edge of coverage are normalized to an orbital height of 800 km. As the orbital height increases, the scan loss + free space path loss reaches a shallow minimum between about 1350 and 1800 km altitude above the earth.

more than one satellite must be in view to any given user at all times. Figure 9.23 shows this trade off for a constellation with 10 satellites per plane and double coverage (i.e., two satellites always in view from all possible user sites). Figure 9.24 shows a similar trade off for a constellation with 15 satellites per plane with the same double coverage requirement (skyrocket 2018b).

Figures 9.23 and 9.24 indicate that there are a number of iterative analyses that can be performed to balance scan loss, free space path loss, number of satellites in view at any user site, and orbital altitude. Once these geometrical analyses have been performed, it is necessary to look at other factors. From the satellite hardware design aspect, a critical factor is the radiation environment: the higher the LEO orbit altitude, the worse the radiation environment becomes as it approaches the first main Van Allen radiation belt at around 1500 km. Perhaps the most critical factor is the RF transmit power available from the user's handset. Battery and adaptive handset antenna technology may be able increase the available EIRP from the user's phone, but the biological radiation limits imposed for safe usage will place an upper bound on handset EIRP.

9.3.7 Radiation Safety and Satellite Telephones

In the United States, the Federal Communications Committee (FCC) mandates strict limits on radiated power levels throughout the spectrum. Their rules are usually issued in dockets (see, for example, (FCC 2018a), which provides the main FCC web site and the dockets for evaluating the environmental effects of radio frequency radiation). The Office of Engineering of the FCC has also posted an RF safety program on its web site (FCC 2018b), which provides guidance on the specific absorption rate (SAR) for wireless phones and devices. Many of these guidelines have been developed through IEEE

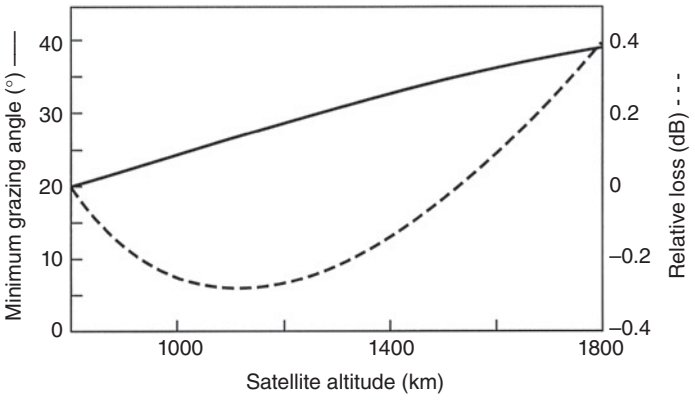


Figure 9.24 Relative transmission loss and minimum grazing angle vs. satellite altitude for a constellation of LEO satellites with 15 satellites per plane and double coverage. Source: Figure 2 of Chiavacci 1999, reproduced with permission of Microwave Journal. The calculations for this figure are similar to those carried out in Figure 9.23, except for this constellation there are 15 satellites per plane. With 15 satellites per plane, a minimum scan angle at the satellite is calculated and, from this, the elevation angle at edge-of-coverage user is found. This angle is referred to as *minimum grazing angle* in the figure. The scan loss + free space path loss for edge of coverage are normalized to an orbital height of 800 km. As the orbital height increases the scan loss + free space path loss reaches a more pronounced minimum than in Figure 9.23. This time the minimum is between 950 and 1300 km above the earth rather than 1350 and 1800 km found for the constellation with one third the number of satellites per plane.

Committees (e.g., see (ANSI/IEEE, 1992)), which have made many proposals to the American National Standards Institute (ANSI) on this issue. Safe exposure levels are given as 0.08 W/kg as averaged over the whole body for the general population and 0.4 W/kg for occupational or controlled exposure for professionals working in this area. These values do not provide enough insight into handheld units held close to the head and many studies are still underway, some of which are reported in (Foster and Moulder 2000). It is clear that handset power levels are well below those that cause ionization damage to tissue. However, while the short-term effects of the power levels used in handsets have been proven to be negligible, there have been insufficient studies at present to provide the long-term effects of such exposure, that is, over more than 10 years of handset use. A number of international groups, in particular European Telecommunications Standardization Institute (ETSI) in Europe, are collaborating on such studies.

9.3.8 Projected NGSO System Customer Service Base

A single satellite in an NGSO system will not provide continuous 24-hour coverage over a given area. If a national or regional coverage is desired, a constellation of NGSO satellites is required with orbits tailored to match the coverage. This was the approach adopted for the Molniya system where a minimum of two satellites in two Molniya orbits could provide continuous 24-hour service. Most of the new NGSO systems have been aimed at mobile users. For mobile users, the problem is to generate sufficient transmit power in a handheld terminal without exceeding the limits for electromagnetic radiation from the antenna into the head and body of the user. Low power transmissions from the hand held unit requires either a satellite in LEO or a very large antenna on a MEO

or GEO satellite. All three alternatives have been developed (Evans 1997). The driving forces behind the decisions made in choosing a system architecture will be discussed in Section 9.4 and some typical systems will be reviewed in Section 9.4. We will now look at two closely associated elements of an NGSO system – or any telecommunications system for that matter – that can have significant implications on customer acceptance: delay and throughput.

9.3.9 Delay and Throughput Considerations

Delay in a communications link is not normally a problem unless the interactions between the users are very rapid – a few ms apart in response time. Long delays, such as those associated with manned missions to the moon, required the development of agreed procedures, much like tactical military or police communications requires specific handoff code words such as *over* to signal the end of one user's input. For most commercial satellite links that are over long distances, particularly those with satellites in geostationary orbit, the main problem was not delay, but echo. A mismatched transmission line will always have a reflected signal. If the mismatch is large, a strong echo will return. Over a GEO satellite link, the echo arrives back in the telephone headset about half a second after the speaker has spoken, and usually while the speaker is still speaking. This will interrupt the speaker and the conversation becomes fragmented. The development of echo suppressors and, even better, echo cancellers, solved the problem. Figure 9.25 illustrates the one-way propagation time for a typical LEO, MEO, and GEO system.

Based on the calculations shown in Figure 9.25, the time delay for a signal passing between LEO user 1 and LEO user 2 in the same instantaneous coverage is 5.4 ms (2.7 ms up and 2.7 ms down) and the go and return (round trip) delay between the two users is twice this at 10.8 ms. It is rare, however, for a user to be immediately underneath a LEO satellite and, for LEO satellites in higher orbits, the round trip delays due to propagation time can be more than double this. Globalstar, which has a maximum pathlength from the satellite to the user of 2500 km, will have a maximum round trip delay time of 33 ms. For GEO users, the up and down (forward) link delay is typically 240 ms with the round trip delay 480 ms. However, Figure 9.25 does not tell the whole story. Most MSS systems use voice compression to reduce the bandwidth required for a single voice channel. The coded bit rates for a single voice channel range from 2.4 kbps for Globalstar to 6.25 kbps for Iridium (Evans 1997). The vocoders sample the incoming analog voice signal and produce excellent, low data rate digital reproductions – but at a price in delay. The access scheme can also add additional delay. If the channel is operated in a simplex fashion, that is, you cannot send at the same time as you are receiving, there can be a delay in response. The Iridium time division multiple access (TDMA) access mode uses a time division duplexing (TDD) scheme. A TDD scheme allows transmissions to occur for a certain period (while receive functions are off) and then transmissions cease while receive operations are in use. In the present Iridium TDMA access scheme, eight users share a frequency assignment and, within this frequency channel, share a 45 ms transmit frame and a 45 ms receive frame. There can therefore be up to 90 ms between transmission and reception of specific parts of a message. On the Iridium satellite, the onboard processing system translates the received signal to baseband, the header address information is read, and the appropriate route selected for onward transmission. The baseband signal is then reformatted, up-converted to the RF band, and transmitted. All of this

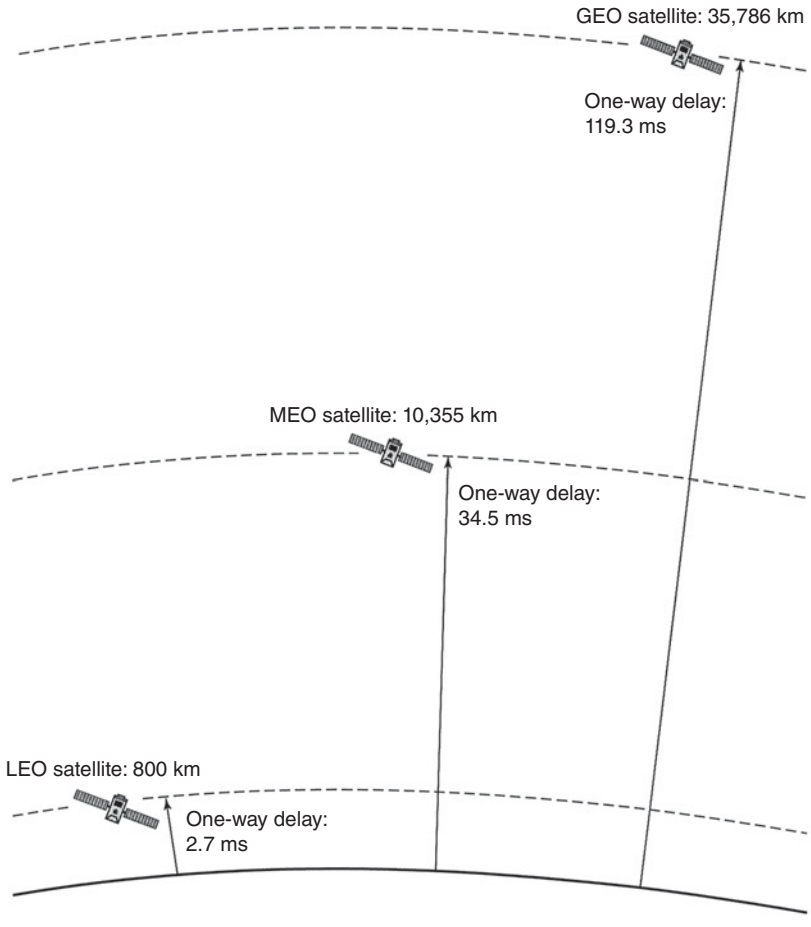


Figure 9.25 One-way propagation delay for the three orbits shown: LEO, MEO, and GEO. The one-way delay figures shown above have been calculated assuming the radio signal propagates at the speed of light in a vacuum, that is, 3×10^8 m/s. That is, no account has been taken of any delay due to the refractive index of the atmosphere not being unity. Also, no account has been taken of any processing delay imposed on the signal from any source coding, channel coding, modulation, or access scheme.

takes time. The forward delay (ground-to-satellite plus satellite-to-ground) within the same instantaneous coverage averaged 153 ms in the initial operational tests of Iridium. A transoceanic link delay using ISLs averaged 253 ms – almost the same as for a GEO satellite link. Delay can also have an adverse effect on the throughput of the signal, as noted in Chapter 8. If the protocol used in the link is not adapted for the particular delay environment, appreciable reduction in throughput will occur. Customer acceptance of a service has been found to be driven by three prime factors: *access ability* (i.e., can the required connection be obtained immediately on request?), *availability* (i.e., once connected, will the call be dropped?), and *performance* (i.e., is the error rate low and the throughput high?). Pricing will attract customers but it will not keep them for long if all three prime factors are not met.

As a final element in the discussion on delay, it is worth noting the challenges that face system designers when ISLs are employed to relay signals around a LEO constellation. It

is a fairly straightforward matter to design an ISL to connect two GEO satellites or a LEO satellite to a GEO satellite: the relative motions are not that large. Consider now a LEO system attempting to establish connections across the constellation. The connections will have to be both in plane (i.e., around the same orbit plane of that particular ring of satellites) and across planes. When the satellites are close to the equator, the orbital planes are at their furthest separation and the rate of change between two LEO satellites traveling in the same direction is at a minimum. As satellites move closer to the poles, the more rapidly they have to steer their ISL antennas to maintain contact. In some operational modes, Iridium switches off the across-plane ISL links when the spacecraft are above latitudes of about 60° (Evans 1997). In no case, however, can Iridium maintain an ISL link between planes where the satellites are moving in opposite directions. There will therefore be a *seam* in the constellation across which no ISL links can operate. This is illustrated in Figures 9.26 and 9.27 depicts a polar view of the constellation with the ISL directions. In the original Teledesic system (Teledesic 2000) 840 satellites were proposed for the full constellation, then 288 satellites, and eventually fewer than 200 satellites. It was reportedly designed to operate across the LEO seam and so it is

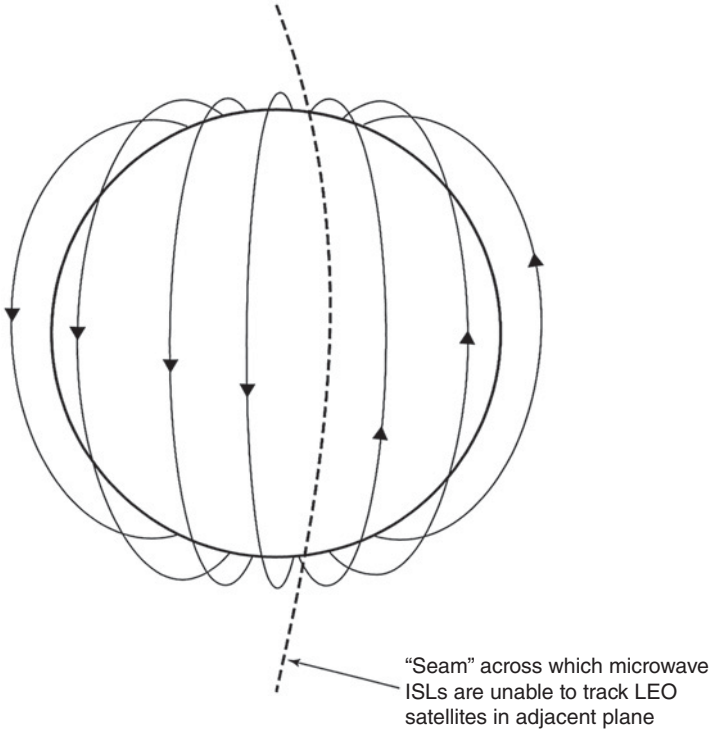


Figure 9.26 Schematic of the ISL seam in the Iridium constellation. The Iridium satellites are in an orbit that is close to polar (86.5° inclination). There are four ISL antennas on each satellite that are used to communicate with adjacent satellites. The ISLs operate at 23 GHz and use solid reflector, tracking antennas. The inertial mass of the antennas combined with the need to have a stable satellite platform for the normal communications mode to the earth limits the rate of change of the tracking mechanism. Satellites across the seam are traveling at a closing speed of about 36 000 mph (~58 000 km/h) and it is likely that only lightweight optical ISLs will be able to track at the angular rates required across an LEO seam (see Figure 9.27).

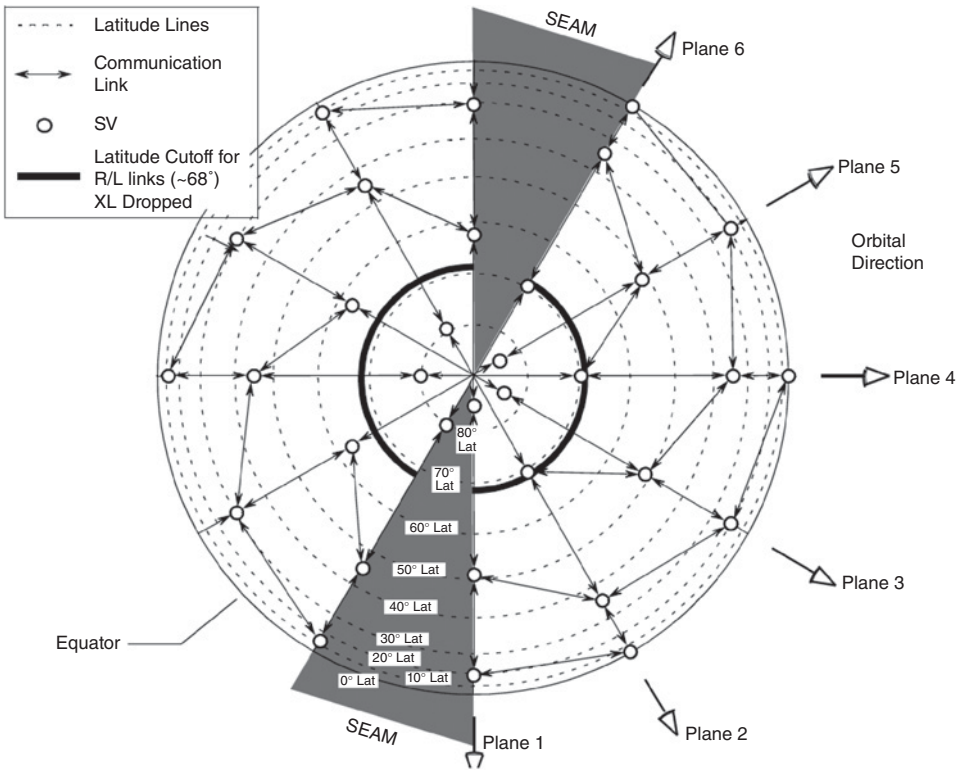


Figure 9.27 Polar view of the Iridium next constellation. (Source: From Figure C, Iridium 2013.) Note the shaded area that denotes the *seam* of the Iridium constellation, across which the ISLs will not work satisfactorily. To communicate with satellites on the other side of the seam, Iridium satellites communicate northwards or southwards with satellites in their plane until they can find another Iridium satellite that is operating, or about to operate, on the other side of the seam.

likely that the ISLs were to be optical and not microwave. Optical ISL antennas are much smaller and lighter than microwave ISL antennas and so impose fewer tracking restrictions due to inertial forces when under acceleration. Iridium and Iridium Next use microwave antennas around 2 m in diameter, which have considerable inertial mass when being repointed. Whether or not to use ISLs; whether to design to operate across the seam if ISLs are used; selecting an orbital height, number of satellites visible at any instant, coverage region, and so on; all interact in the overall system design.

Whether the satellite system is to communicate directly with the end user, to a subscriber through a wireless local loop (WLL), or via a portal in the public switched telephony network (PSTN), the characteristics of the earth segment are of critical importance in the overall system design.

9.3.10 Earth Segment Aspects

A communications satellite can be considered as a point-to-multipoint distribution node. It may have a single uplink (as in the direct broadcasting satellite service – DBSS) or a multitude of uplinks from, for instance, a large number of small earth terminals. The

earth terminals can be mobile or fixed, the latter meaning the terminal does not move, but maintains a constant location. The large antennas used in the INTELSAT system started as 30 m diameter Standard A paraboloids, before reducing in diameter as satellite EIRP increased. Nevertheless, there was considerable demand for smaller antennas and, over time, the fixed earth terminals gradually reduced in size from a diameter of about 5 m to a diameter of 1.8 m. The smaller antennas are generically called VSATs – very small aperture terminals. The antenna size used for DTH services is usually around 60 cm in diameter, although the United States does not require a user to seek coordination for a dish size up to 1 m. All of these antennas are specifically designed to operate with a geostationary satellite, and so no tracking is required. Before investigating how an earth terminal can operate with the new constellations of NGSO satellites without mechanical tracking, we will investigate the bandwidth required for a communications system. We will look first at frequency division multiple access (FDMA).

FDMA generally offers the lowest costs for entry-level VSAT systems from the user's perspective since the receiver bandwidth and terminal transmit power required are the lowest. These systems were initially designed to carry thin route traffic, typically the equivalent of one digital voice channel at 64 kbps. The occupied bandwidth of an RF channel carrying a digital signal with a symbol rate R_s and using error control coding with a code rate R_c is given by

$$B = R_s(1 + \alpha)/R_c \text{ Hz} \tag{9.13}$$

where α is the roll off factor of the root raised cosine (RRC) filters in the link. For example, in a link using quadrature phase shift keying (QPSK) modulation where two bits of information are carried by each transmitted symbol, a message information rate of 64 kbps results in a transmission symbol rate of $R_s = 32$ kbps. If the message data bits are encoded with one half rate forward error correction (FEC), code rate $R_c = ½$, the occupied bandwidth required for a 64 kbps signal is B_{occ} where, from Eq. (9.13)

$$B_{occ} = 32000 \times (1 + \alpha)/(1/2) \text{ Hz} = 64 \times (1 + \alpha) \text{ kHz} \tag{9.14}$$

Typical values of α for satellite links lie between 0.25 and 0.5, with the higher value being easier, and thus cheaper, to realize when conventional analog filters are used in the transmitter and receiver. If an $\alpha = 0.5$ SRRC filter is used, the occupied RF bandwidth of this signal is

$$B_{occ} = 64 \times (1 + 0.5) = 96 \text{ kHz} \tag{9.15}$$

A VSAT that is required to transmit a 64 kbps stream of data using QPSK modulation, with half rate FEC and a square root raised cosine (SRRC) filter having a roll off factor of 0.5, therefore needs an RF channel bandwidth of 96 kHz and has a receiver noise bandwidth of 64 kHz. Note that the roll off of the SRRC filter, while adding additional spectrum requirements, does not alter the noise bandwidth; all RF and IF SRRC filters used in digital radio links have a noise bandwidth in hertz equal to the symbol rate in symbols per second. In practice, a guard band will have to be added between FDMA channels so that adjacent signals do not overlap in frequency at the satellite, and to allow the filters in the receiver that extract individual channels to roll off between channels. Most VSATs will operate unattended for most of the time and will be exposed to all weathers. The frequency of the fundamental oscillator may therefore drift. For this reason, fairly large guard bands need to be designed in, which generally leads to a spectrum allocation at the satellite of around 120 kHz for each 64 kbps (voice) channel of the

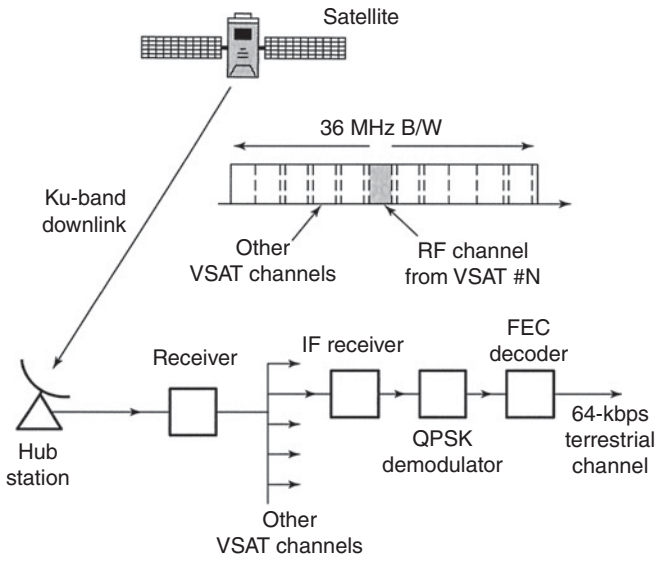


Figure 9.28 Schematic of a 64 kbps equivalent voice channel accessing a satellite using FDMA. The 64 kbps information rate is contained in a bandwidth of 96 kHz when transmitting to the satellite. The bandwidth of the satellite transponder (from frequency f1 to f2) is divided up, or channelized, into increments of 96 kHz so that a large number of VSATs can access the transponder at the same time. Each of the 96 kHz channels requires a certain amount of spectrum on either side to guard against drift in frequency, poor VSAT filtering, and so on. The 96 kHz channels plus the guard bands on either side add up to a channel allocation of about 120 kHz per VSAT. From a spectrum allocation viewpoint, therefore, a typical 36 MHz satellite transponder would permit the simultaneous access of 300 VSATs, each of which is transmitting the equivalent of a 64 kbps voice channel. Because each VSAT uses a single channel continuously on the uplink, it is often referred to as SCPC (single channel per carrier) FDMA.

type described above. This situation, where the VSAT is acting as a hub for many users, is illustrated in Figure 9.28.

The VSAT described above could also be the hub for a WLL. The WLL would essentially be a digital mobile cell service that offers communications to remote locations that are not served by any other systems. The VSAT/WLL fits into a segment of the market depicted in Figure 9.29.

With a high user density, it is economic to introduce high-speed fiber optic service. At the other end of the scale, with fewer than 10 users per km^2, heavy subsidies are required to implement and run a communications service. VSATs themselves, however, appear to be at an inflection point (Satellite Today, 2017a,b). It may be that VSAT services need to adapt to network architectures traditionally associated with mobile cell service, such as using a packet core as framework (Satellite Today, 2017a,b). With each service being essentially data-centric, the end user will demand the ability to move from audio, to video, or to a combination of both without changing supplier. The long-term-evolution (LTE) of terrestrial mobile service and the many proposals for new LEO and MEO communications satellite systems will accelerate the LTE/LEO/MEO/GEO interoperability, which, in turn, will drive the need for a standards-based approach (Satellite Today, 2017a,b). One of the key nodes in all of this is the VSAT itself. Surveys of users (Satellite Today, 2017a,b) have shown that the capability of the VSAT is a priority, with

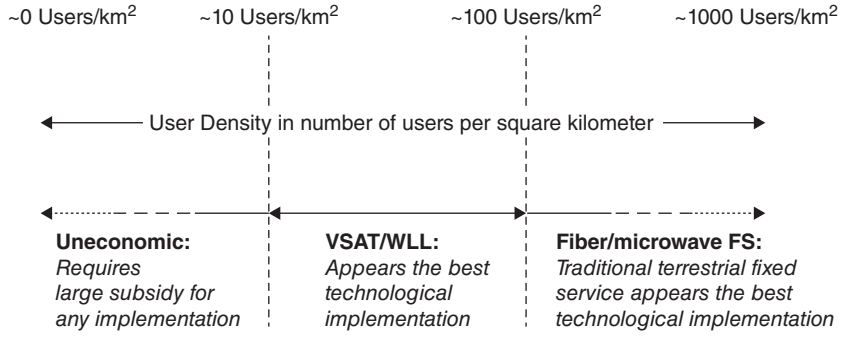


Figure 9.29 Approximate economic break points in the implementation choices for serving new regions with different population densities. Physical distance, major transportation routes, and geographic barriers, as well as the individual country's demographics and political influences, can alter the break points.

cost only secondary. The size of the VSAT antenna in this survey was considered to be insignificant. It is easy to understand why this is, as the O3B user earth terminals can be as small as 0.65 m and still provide a minimum of 100 Mbit/s.

9.4 System Considerations

There are four important factors that influence the design of any satellite communications system: incremental growth, interim operations, (satellite) replenishment options, and end-to-end system implementation.

9.4.1 Incremental Growth

The 1964 decision by the Interim Communications Satellite Committee (soon to become the International Telecommunications Satellite Organization INTELSAT a few years later) to select a GEO satellite system rather than a 12-satellite MEO system that was supported by major entities on both sides of the north Atlantic at that time was driven by incremental growth plans as well as by launcher technology. The primary international traffic route was across the Atlantic Ocean, followed (a long way second) by the Indian Ocean region, and (an even longer way third, at that time) by the Pacific Ocean region. The system could be grown incrementally with a GEO architecture. The first GEO satellite – early bird – was placed over the Atlantic Ocean region in 1965. For the first decade of operations, new satellites were launched into the Atlantic Ocean region to replace satellites that had been operating there. The satellites being replaced were moved to the Indian Ocean region and the satellites replaced in the Indian Ocean region were moved to the Pacific Ocean region. It was not until INTELSAT VII that Intelsat specifically designed a satellite for the Pacific Ocean region from scratch. This approach to incremental growth served Intelsat well. By comparison, the new LEO and MEO mobile service systems now in operation require all of the satellites to be in operation before full operations can begin. However, most of the LEO and MEO system operators developed interim operations plans where a reduced number of satellites could

provide useful service. We will now look at other system considerations that can affect the design of the satellite network in other respects.

9.4.2 Interim Operations

Interim operations for LEO and MEO systems serve two functions: they can bring a service on line gradually, introducing the technology to the market while teething problems are sorted out; and they can act as fall back plans should multiple satellite failures occur over a short period. Nearly all of the LEO and MEO systems undertook such interim operations. Orbcomm began commercial operations with less than half of its 36-satellite constellation in place, thus becoming the first commercial LEO system to establish a revenue stream. Globalstar began with 32 out of the planned 48-satellite constellation and O3B plans to start operations with eight satellites spaced 45° apart in an equatorial orbit. Iridium, since it uses ISLs to complete the network, required all 66 satellites to be available before it began beta testing in November 1998. SpaceX launched MicroSat-2A and 2B to test the ISLs and phased array antennas to be used in their planned NGSO constellation of up to 7500 spacecraft (skyrocket 2018c). SpaceX is also planning to launch a new constellation of satellites that operate in V-band (40–75 GHz) (FCC 2018b). Some details are given in Table 9.5.

The technical planning for interim operations includes relaxing the number of satellites visible to any user at any particular time, which lowers the number of satellites required to complete the constellation. The elevation angle minimum for users is also usually lowered, the gaps between operational satellites in the same plane are made symmetrical, and the orbits adjusted if possible to maximize coverage over those parts of the day when user service requests are highest. Most LEO constellations have at least four satellites per plane and multiple spacecraft launches are used in the constellation

Table 9.5 SpaceX V-band proposal

LEO constellation					
	Initial deployment		**Final deployment**		
Parameter					
Orbital planes	32	32	8	5	6
Satellites/Orb. plane	50	50	50	75	75
Altitude (km)	1150	1100	1130	1275	1325
Inclination	53°	53.8°	74°	81°	70°

VLEO constellation			
Initial deployment	**Final deployment**		
Parameter			
Satellites per altitude	2547	2478	2493
Altitude (km)	345.6	340.8	335.9
Inclination	53°	48°	42°

Source: Data extracted from (FCC 2018b).

Table 9.6 Primary and replenishment launchers for the NGSO systems

LEO/MEO system	Primary launchers (number per launch)	Secondary/replenishment launchers (number per launch)
Iridium NEXT	SpaceX Falcon 9 (10)	Dnepr (2)[a]
Globalstar-2	Soyuz 2-1a[b] (6)	[d]
O3B	Soyuz 2-1a[b] (6)	[d]
OneWeb	Soyuz 2-1z[c] (32–36)	Virgin Galactic (1)[e]
SpaceX LEO	SpaceX Falcon 9 (12–24)	[d]
SpaceX VLEO	SpaceX Falcon 9 (24–36)	[d]
Orbcomm OG2	SpaceX Falcon 9 (11)	[d]

[a]Dnepr launch site is in Russia
[b]Launch site is Kourou
[c]Launch sites are Kourou and Kazakhstan
[d]No replenishment launcher announced
[e]39 replenishment launches have been announced, using the air-launched Launcher 1

build-up. A SpaceX Falcon 9 rocket carried 11 Orbcomm satellites into orbit (Orbcomm 2018), A Soyuz II carried six Globalstar satellites, and a SpaceX Falcon 9 rocket carried 10 Iridium Next into orbit. Most NGSO systems populate their constellation with more spacecraft than are needed at any one time and so, when a satellite fails in service, there is usually an in-orbit spare to take its place. If more than one satellite fails in a plane, additional satellites must be launched to replenish the system.

9.4.3 Replenishment Options

Launching five or more satellites to replace one failed satellite makes little economic sense. As a result, the LEO service providers may use smaller rockets to replenish their system. With the new very low earth orbit (VLEO), LEO, MEO, and other NGSO constellations that have up to 4000+ satellites, however, it is more likely that no single satellite replenishment will be contemplated in most cases. The full system complement of satellites will contain several "surplus" spacecraft in each of the orbital planes, which will make it relatively simple to replace a failed satellite with a spare satellite in orbit. Table 9.6 lists the primary and replenishment launchers proposed to be used by the NGSO systems.

9.4.4 End-to-End System Implementation

A communications system can be part of a larger network (e.g., just providing the long distance portion of the connection) or it can provide the full end-to-end system implementation, from user to user. AT&T and Intelsat, when they were first set up, did not provide end-to-end service: AT&T provided long-distance capacity for local telephone companies and Intelsat provided satellite capacity for entities such as AT&T to carry their international traffic. Neither company interacted directly with the end-user. Indeed, at that time, there were specific laws or protocols that prevented this from happening. These laws no longer exist, but at that time, they had a significant impact on the early satellite systems, influencing the systems design in a number of ways.

The systems architecture of an NGSO constellation will be heavily influenced by the decision on whether or not to provide service directly to the end-user. It will also be impacted by the decision on whether or not to include established telephone companies in the delivery of the service. By their very nature, mobile satellite systems have committed to serve the end-user directly. However, different approaches have been taken with regard to including established telephone companies. Two examples of organizations that took opposite decisions are Globalstar and Iridium. Globalstar elected not to bypass existing telephone companies while Iridium did. These decisions led to a very different architecture for the two systems, which will be discussed in the next section.

9.5 Operational and Proposed NGSO Constellation Designs

Ten NGSO satellite constellation designs are reviewed briefly in the following discussion, four MSS offerings with multiple beams, one with single beam coverage providing both two-way services and one-way store-and-forward services, and five internet-multimedia satellite systems. Four of the systems never had any of their proposed satellites launched (Ellipso, New ICO, Skybridge, and Teledesic) but their various approaches to establishing a satellite architecture are very instructive. Not only do they show the level of confidence – or rather lack of confidence – in the technology existing at the time they were proposed, but the degree to which they each sought to capture a specific market.

9.5.1 Ellipso

The Ellipso constellation drew from studies of the world's population distribution and the potential market for MSS users. Figure 9.30 (abstracted from data in (Ellipso, 1998)) shows that more than 85% of the world's population lives north of the equator. Additional studies (Ellipso, A) concluded that an equatorial constellation of MEO satellites could serve the bulk of the world's population. Ellipso therefore adopted an incremental approach to their service offering. The first set of satellites would be in a circular equatorial orbit. The second set would be in elliptical equatorial orbit, with the ellipticity of the orbit designed to provide higher dwell times over the regions of greater demand. The third set of satellites would be in sun synchronous three-hour orbits inclined at 116.6° to provide coverage over the highly industrialized Northern Hemisphere regions. The equatorial orbit groups of the Ellipso system were called Concordia™ and the sun synchronous group was called Borealis™. Details can be found in Table 9.7. The Ellipso spacecraft was based on the Boeing GPS satellite bus and up to five satellites could be launched by a single rocket. No onboard processing was considered; the signals received at the satellite are simply transponded down to GESs for onward routing via the terrestrial PSTN or satellite network. No ISLs were proposed between spacecraft.

9.5.2 Globalstar

In a similar manner as Ellipso, Globalstar elected to develop a constellation that was aimed at the populous regions of the earth. The Globalstar orbital planes are therefore inclined at 52° to the equator, thus ignoring the sparsely populated high latitude regions. To minimize the power requirements of the user handset, the constellation altitude was

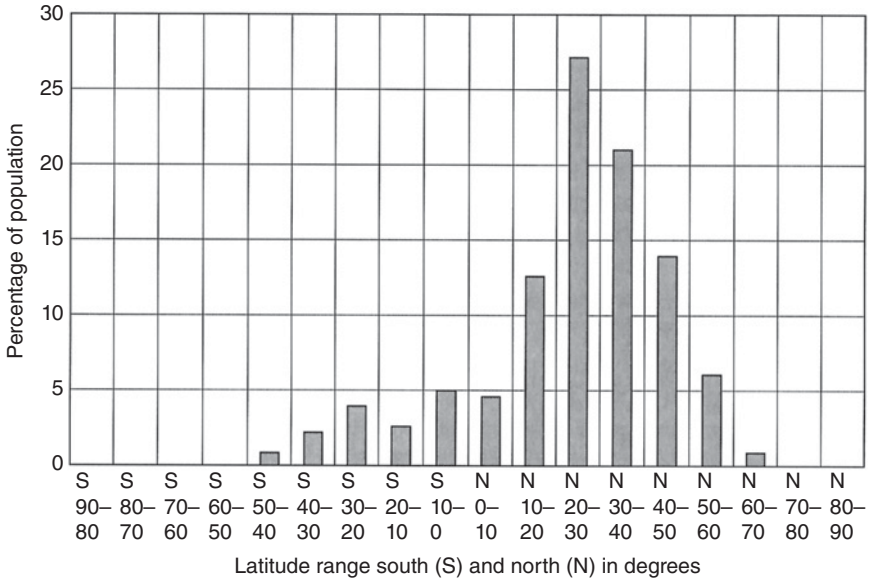


Figure 9.30 Percentage of the world's population living in the given latitude ranges. Source: Data originally extracted from Ellipso A, 1998; Draim 1998, reproduced with permission of Springer Nature. The data in the figure show that more than 85% of the world's population lives in the northern hemisphere. Designing a satellite system that spends most of its time in the northern hemisphere would therefore cover the world's population more efficiently than one that equally divides its time over the whole globe.

Table 9.7 System parameters of five NGSO constellations aimed at data and voice communications

System parameter	Ellipso	Globalstar	New ICO	Iridium	Orbcomm
Number of planes	$1 \rightarrow 3 \rightarrow 5$	6	2	6	$4 \rightarrow 5$
Satellites per plane	1×7 then 1×7 and 2×3 then 1×7, 2×3, 2×5	8	5	11	4×8 then 4×8 and 1×4
Total complement	23	48	10	66	36
Orbital inclination	3 at 0°, 2 at 116.6°	52°	45°	86.5°	4 at 45°, 1 at 72°
Orbit type	1 circular (0°) 2 elliptical (0°) 2 sun synchronous	Circular	Circular	Circular	Circular (45° and 72°)
Orbital height (km)	1 circular 8050 2 elliptical 6149–8050 2 sun synchronous 633–7605	1414	10 255	780	775
Spot beams per satellite	61	16	163	48	1
Satellite lifetime	5–7 yr	~7.5 yr	~12 yr	5–7 yr	5–7 yr

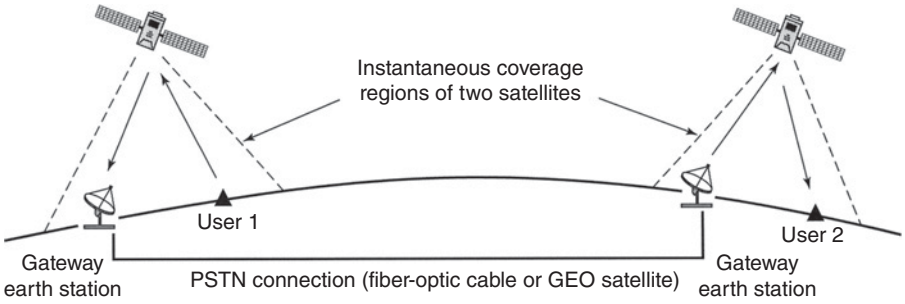


Figure 9.31 Schematic of end-to-end connection of satellites that have no onboard processing or ISLs. User 1 is in a different instantaneous coverage region than that occupied by user 2. The signal from user 1 is picked up by the gateway earth station and relayed to the gateway earth station of user 2. The signal is then sent up to the satellite from the second gateway earth station and then down to user 2. If user 1 or user 2 is using a fixed telephone (or computer) the signal would simply pass over the regular PSTN circuits and not via the space segment. Because the users must be in line-of-sight contact with the gateway earth station, no maritime traffic can be picked up unless the ship is close to land (and a gateway earth station).

lowered to just below the first Van Allen radiation belt. This increased the total number of satellites needed to 48. No onboard processing or ISLs are used; the signals received at the satellite are simply transponded down and the GESs process the signals for onward routing (see Figure 9.31). Like Ellipso, service over water is restricted to coastal regions where the satellite is within radio range of a GES.

9.5.3 New ICO

ICO Global was the company that was spun off from the International Maritime Satellite Organization (INMARSAT); New ICO is the company that emerged from bankruptcy protection in 2000. INMARSAT was initially set up solely for the purpose of providing reliable communications to maritime traffic. Later, INMARSAT also provided aeronautical services, in addition to priority links for safety communications, whether on land or sea. New ICO, although primarily aimed at the LMSs market (land mobile services), also needed to provide capacity for maritime links. New ICO elected not to include ISLs in their system architecture nor any significant onboard processing. Since a LEO constellation would not provide maritime coverage without ISLs, a higher orbit was necessary. If little onboard processing is used, traffic routing from mobile to mobile would have to be carried out at the GESs (as it was for Ellipso and is for Globalstar) necessitating a double-hop link. A double-hop link involves two uplinks and two downlinks. (A double hop is shown in Figure 9.31; two different earth-space links are used to complete the connection.) A double hop configuration is not feasible for a GEO constellation since the overall delay would be completely unacceptable at about one second. New ICO therefore adopted a MEO constellation. An inclination of 45° was used, and since the orbit altitude was so high, full global coverage was possible.

9.5.4 Iridium

As we learned in earlier sections of this chapter, the genesis of Iridium was formed round the need to communicate from anywhere to anywhere on the surface of the world, even

where no telecommunications infrastructure existed. The system therefore had to be standalone. From this – and the need for a low power handset – came the concept of first 77, and then 66, almost-polar orbiting LEO satellites linked via ISLs. Each of the satellites in the constellation acts as a switching node. Uplink signals are received and demodulated at the satellite using onboard processing to recover individual data packets at baseband so that the header information can be read. Using this information and links to the network control stations, the next node for each packet is determined and the packet is reformatted with the next address. The baseband data packet is then processed and up-converted for transmission either to the ground at L-band directly to another Iridium user, or at 20 GHz to a GES, or over one of the four ISL links available (at 23 GHz) to the next satellite in the chain. Onboard processing is needed to carry out the entire message routing and formatting functions.

9.5.5 Orbcomm

Many research organizations and businesses need to obtain data from locations that are either inaccessible on a regular basis or are moving within areas without good cellular telephone coverage. Examples are buoys measuring water characteristics in rivers and at sea, and delivery trucks. Tracking of high value cargo on trucks is another application that needs to send a short message to a central station at regular intervals. A GPS receiver on the cargo determines its location and this information is sent with an ID number via an Orbcomm satellite. If the truck carrying the cargo is hijacked, its route can be followed and the truck intercepted. Much of this information is usually not required in real time nor does it need a high capacity link. Orbcomm developed their system around this requirement and have orbited a constellation of satellites with both two-way data communications and store-and-forward capabilities (see Table 9.7). The satellites are lightweight (40 kg) and simple in design and execution (Orbcomm 2018). The small circular payload unfolds a long, UHF antenna underneath it. The long antenna boom provides gravity gradient stabilization, and so no attitude thrusters are needed to maintain pointing toward the earth. A single beam is used to develop the instantaneous coverage and no onboard processing is used.

A terminal that is within the coverage area of a satellite and a gateway station (which includes almost all of the United States) can send short messages to the gateway station in real time. The message length is limited to a few hundred bytes. A terminal that has data waiting to be uploaded for store and forward listens for the passage of a satellite and then uploads its data when the satellite is in view. The data, in the form of a packet with the address of the intended recipient, are stored and transmitted to a gateway station for onward transmission to the recipient when the satellite is within range of the gateway station. Orbcomm satellites carry short messages, with a relatively high cost per transmitted bit. The system is therefore most attractive to users who want to send a small number of high value bits, such as requests for help in emergency situations or tracking information for high value cargo. Orbcomm satellites can be classified as smallsats (see Chapter 8).

None of the five NGSO constellations above were initially designed to carry traffic at rates higher than 10 kbps. This is not adequate for internet access, which has emerged as a vitally important requirement in mobile systems. Two NGSO constellations were proposed to address this market, Skybridge and Teledesic. Neither succeeded in, literally, getting off the ground. Three NGSO systems drew on some of their ideas, O3B,

OneWeb (formerly WorldVu), and SpaceX LEO and VLEO. All five NGSO systems are discussed below.

9.5.6 Skybridge

Skybridge evolved a similar approach to coverage as Globalstar, by selecting an inclined orbit that covers the major population densities. Like Globalstar, Skybridge satellites carried a non-processing payload and did not have ISLs; so all traffic was to be transponded down to the GESs for processing and onward routing. However, Skybridge satellites were intended to carry wide-band traffic and therefore were designed to operate at frequencies above 10 GHz. They chose to employ the same Ku-band frequencies that the FSS service in GEO uses: 12.75–14.5 GHz on the uplink and 10.7–12.75 GHz on the downlink. To allow successful coordination with existing FSS GEO systems, they elected to prevent any operations (up or down) whenever a satellite look angle was within 10° of the GEO orbital plane. This requirement led to a relatively large number of satellites (80 vs. 48) for the constellation. The decision not to use ISLs also required a very large number of GESs (on the order of 200). Skybridge also proposed to use the concept of a fixed earth cell (see Figure 9.32). More details of Skybridge can be found in Table 9.8. When

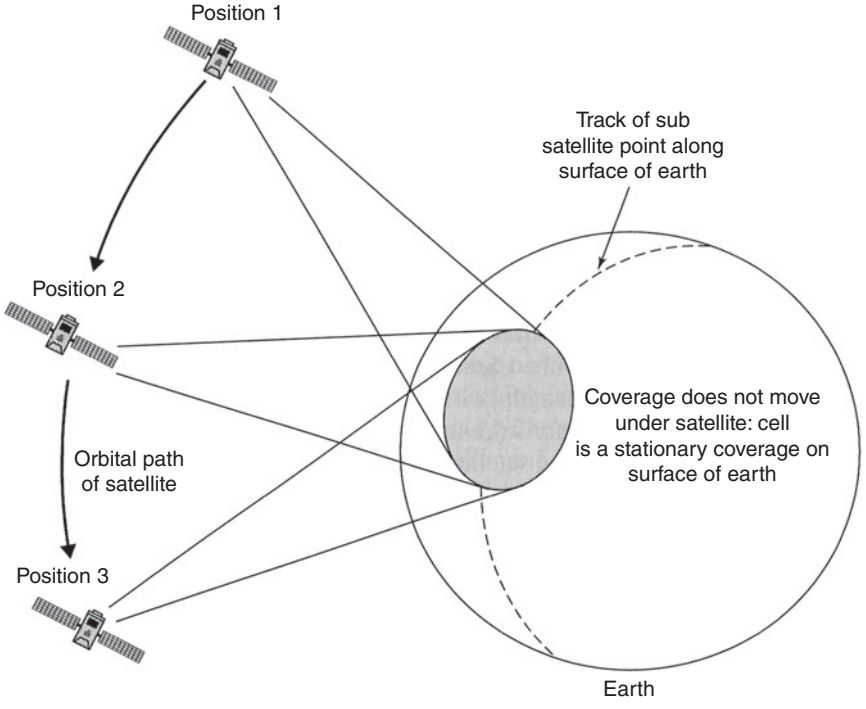


Figure 9.32 Concept of a stationary cell. Unlike the coverage of the NGSO satellite shown in Figure 9.14, a stationary coverage (or fixed earth cell) of an NGSO satellite does not move with the satellite. The phased array antenna on the satellite steers the beam, while the satellite transits, to keep the coverage on the surface of the earth constant. As the satellite moves between positions 1, 2, and 3, the stationary coverage is maintained on the surface of the earth. Separate antennas are used for communications coverage and gateway links. In this way, a gateway need not be within a given stationary coverage.

Table 9.8 System parameters of two NGSO constellations aimed at internet multimedia communications

System parameter	Skybridge	Teledesic
Number of planes	20	12
Satellites per plane	4	24
Total complement	80	288
Orbital inclination	53°	~90°
Orbit type	circular	circular
Orbital height (km)	1469	~1400
Spot-beams per satellite	18	—
Satellite lifetime	~7 yr	~7 yr

Skybridge ceased work, the spectrum it was proposing to use was acquired by OneWeb, formerly WorldVu (see Section 9.5.8).

9.5.7 Teledesic

Teledesic started from the same precept as Iridium, but was designed for internet-like data traffic rather than voice communication. Any user could access any other user or ISP (internet service provider) independent of location and the existing telecommunications infrastructure. The concept of Teledesic was to provide a complete worldwide data communications system above the surface of the earth using satellites, instead of on the earth's surface using fiber optic cables. This requirement dictated the use of wideband data links, onboard processing, and ISL links. To avoid the necessity of coordinating with existing systems, Teledesic chose to move their operations completely into Ka-band (there was no Ka-band satellite system in orbit then).

As noted earlier, to reduce the impact of rain, Teledesic also limited the elevation angle at which users could access the satellites (the mask angle) to 40°. The initial Teledesic constellation had a complement of 840 satellites (22 planes with 40 operational satellites per plane) plus 40 spare satellites in orbit. The orbital altitude was later moved up from 700 to about 1400 km, which reduced the number of planes to 12, with 24 operational satellites in each plane (see Table 9.8). The early estimates of Teledesic's system cost were between US$9B and $12B, using 840 satellites. Reduction in the number of satellites to 288 lowered the cost significantly, and further reductions in the number of satellites would probably have made the cost of creating the system more acceptable (Teledesic 2017). Internet traffic has driven many of the new satellite constellations, and it is worth looking at a snapshot of this traffic at the end of the twentieth century in Table 9.9. Whether the internet traffic originates from a rural household or a major Wall Street company, it all has to go via a server so that the traffic is routed successfully. In 2016, 70% of the entire world's internet traffic made its way through a huge server in Loudoun County, Virginia, United States (Washingtonian 2016).

In the first decade of the twenty-first century, the GEO system approach to communications seemed to be king. The GEO appeared to have the unique characteristic of providing data transfer by satellite at the lowest cost per bit. None of the scheduled or

Table 9.9 Internet traffic centers

1. London	(18 terabits per second)	9. Washington, DC	(4.0 terabits per second)
2. New York	(13.2 terabits per second)	10. San Francisco	(3.9 terabits per second)
3. Amsterdam	(10.9 terabits per second)	11. Toronto	(3.5 terabits per second)
4. Frankfurt	(10.5 terabits per second)	12. Chicago	(2.7 terabits per second)
5. Paris	(9.7 terabits per second)	13. Seattle	(2.6 terabits per second)
6. Brussels	(6.2 terabits per second)	14. Vancouver	(2.5 terabits per second)
7. Geneva	(5.9 terabits per second)	15. Tokyo	(2.4 terabits per second)
8. Stockholm	(4.4 terabits per second)		

Source: Data provided by Dr. Feng of Virginia of Tech in 2000, private communication.

operating LEO and MEO satellite constellations seemed to be able to demonstrate a significant added value from the use of their particular service when a commercial return on investment was required. There was a clear military requirement for many of the new constellations – from anywhere to anywhere – without any intervening infrastructure, but the growth in terrestrial cellular systems and optical fiber links appeared to have removed much of the potential commercial demand for these new services. At the turn of the twenty-first century, more than 90% of all internet traffic flowed through about 30 metropolitan areas. If these conurbations are connected via optical fibers or through high powered spot beam antennas from GEO, the remaining traffic is what a LEO or MEO system would pick up. The same appeared to be true for cellular telephony: what the major cities could not provide seemed to leave very little traffic for a high priced LEO or MEO alternative. But, as always, things never stay the same for long, especially with technology.

Three remarkable game changers occurred in the second decade of the twenty-first century. First: the demand for geostationary satellites trended downwards. From about 25 geostationary satellite procurements a year, only about 15 a year were being ordered. Second: microminiaturization of all aspects of satellite payloads also trended downward – from just small to minute! By the end of 2017, more than 10 times more smallsats (see Chapter 8) were being launched per year than geostationary communications satellites. And finally, third: with both the reduction in satellite size and concomitant launch costs, there now appeared to be a financial case for attempting to serve those regions of the world where no internet service had been feasible before. The three ventures that are attempting to do this – OneWeb, O3B, and SpaceX LEO/SpaceX VLEO – are discussed below.

9.5.8 OneWeb

Formerly known as WorldVu, OneWeb acquired the spectrum of SkyBridge. Several system designs were proposed by OneWeb. The first was a constellation of 648 satellites in 18 near-polar orbits that would provide global internet service. The number of satellites changed to 720 and then 640, all in a 1200 km orbit. The proposed spacecraft are quite small, about 125 kg, designed to provide service to small, low cost terminals that, through a WLL, would provide LTE and 3G mobile as well as Wi-Fi. Using Ku-band frequencies, apparently most of the capacity has already been sold, so OneWeb is considering quadrupling the constellation to 1972 satellites. In June 2015, Ariane was

selected to provide 21 multi-satellite launches using Soyuz rockets. Virgin Galactic was also put under contract to provide 39 single-satellite launches using its LauncherOne smallsat launch vehicle. Later in 2017, OneWeb said it would build an additional 2000 V-Band satellites: 720 in LEO (1200 km altitude) and 1280 in MEO. The user terminals will be phased array antennas approximately 36 by 16 cm that reportedly will provide internet access at 50 Mbit/s. The first six production satellites were launched in February 2019. If these work as required, OneWeb will begin implementing the fleet of NGSO satellites later in 2019.

9.5.9 O3B

The mission of O3B is similar to that of OneWeb – the provisioning of internet services worldwide, although the reach of the satellites is limited to ±45° of the equator as the circular orbits of O3B satellites are around the equator. There are apparently plans to have some O3B satellites in elliptical inclined orbit to offer services to higher latitude. The O3B satellites are larger than the OneWeb satellites at 700 kg (dry mass 450 kg). The beginning-of-life power is 2400 W, with the end-of-life power being 1700 W.

The spacecraft are three-axis stabilized with 12 antennas on the earth-facing panel. All 12 antennas are fully steerable: 10 cover the same urban centers on each orbit, similar to the Teledesic concept, and as shown in Figure 9.32. The remaining two antennas are for GESs. The total satellite throughput is 12 Gbps, with each downlink and uplink capable of 600 Mbit/s. The VSAT earth stations have diameters from 0.85 to 2.4 m. The 0.85 m dish can handle 100 Mbps on the uplink and 400 Mbit/s on the downlink. The downlink Ka-band frequencies are in the range 17.7–20.2 GHz, with the Ka-band uplinks from 27.5 to 30 GHz. The O3B constellation is required to coordinate with existing Ka-band GEO satellites, but the major problems are only within 5° of the equator, where individual system coordination must take place with satellites that are already in orbit, or have been authorized to be launched.

9.5.10 SpaceX LEO/SpaceX VLEO

The SpaceX LEO and VLEO constellations are probably the most ambitious of all the NGSO systems proposed. A March 2017 filing with the FCC indicated plans to launch 7500 V-Band (40–75 GHz) satellites into an NGSO constellation. By 2024, SpaceX proposed to have a total of 7518 satellites in their system. This constellation might be joined by one proposed by Samsung (Samsung 2018) that would orbit 4600 satellites in 1400 km orbits. Samsung claims that they could provide 200 Gbps per month to 5 000 000 000 inhabitants of the earth. There are no clear details as yet of the Samsung constellation, while SpaceX is quite specific of its two constellations (see Tables 9.10a and 9.10b below).

Table 9.10a SpaceX V-band LEO system

Parameter	Initial deployment		Final deployment		
Orbital planes	32	32	8	5	6
Sats. per orbit/plane	50	50	50	75	75
Altitude (km)	1150	1100	1130	1275	1325
Inclination	53°	53.8°	74°	81°	70°

Table 9.10b SpaceX V-band VLEO system

Satellites per altitude	2547	2478	2493
Altitude (km)	345.6	340.8	335.9
Inclination	53°	48°	42°

Each satellite earth-pointing antenna has a 1.5° beam that provides a 52 km^2 beam from an average altitude of 340 km. The minimum operational elevation angle of the steerable satellite beams as viewed from the earth is 35°. The frequency assignments are shown in Table 9.10c below.

9.6 System Design Example

A company wishes to develop a LEO constellation that provides continuous global coverage. They are restricted to an orbital height of 750 km due to user terminal power, operating time between battery charges, and satellite launcher capabilities. The following design data are required:

- The length of the coverage arc on the surface of the earth within the instantaneous coverage;
- The gain of the satellite antenna if one beam is to illuminate this coverage;
- The number of satellites needed to complete one plane with a suitable overlap; and
- The number of satellites needed to complete a global system.

9.6.1 Length of Coverage Arc

Figure 9.12 illustrates the geometry of a satellite and user terminal. If the minimum elevation angle is set at 10°, we know $r_s = r_e + 750$ km (the orbital height), $\theta = 10°$, and $r_e = 6378$ km (average radius of the earth). We need to find the central angle, γ, (angle ECZ in Figure 9.12), which will allow us to find the length of half of the arc under the coverage-arc EZ. Using the sine rule, we have

$$\sin(\delta)/r_e = \sin(\text{angleSEC})/r_s \tag{9.16}$$

The angle SEC $= \theta + 90° = 100°$ and this yields $\delta = 61.7859 = 61.79°$
If $\delta = 61.79°$, then $\gamma = 180–100 - 61.79 = 18.21° \rightarrow 0.3178$ rad

Table 9.10c SpaceX V-band VLEO frequency assignments

	Frequency ranges (GHz)
Downlink channels	
Satellite user terminal or	
Satellite gateway	37.5–42.5
TT&C downlink	37.5–37.7
Uplink channels	
User terminal to satellite or	47.2–50.2
Gateway to satellite	50.4–52.4
TT&C uplink	47.2–47.7

Arc EZ is therefore given by $r_e \times \gamma$ (with γ in radians) $= 6378 \times 0.3178 = 2027.1$ km. The diameter of the instantaneous coverage region is therefore $2 \times 2027 = 4054$ km and the coverage angle measured at the center of the earth is $18.21° \times 2 = 36.42°$.

(Note that this assumes the coverage is symmetrical about the nadir pointing direction SC in Figure 9.12). Alternatively, we could have derived the same result by noting that the circumference of the earth is $\pi \times$ diameter of the earth $= \pi \times 6378 \times 2 = 40\,074$ km. The fraction of this illuminated by the satellite is $(2\gamma)/(2\pi)$ (with γ in radians) $= 0.1012$.

Thus, the total coverage diameter arc $= 40\,074 \times 0.1012 = 4055$ km.

9.6.2 Gain of Satellite Antenna

The angle δ in Figure 9.12 is half of the antenna beamwidth. The full angle of the antenna beamwidth at the satellite is therefore $2 \times \delta = 61.79° \times 2 = 123.6°$. The gain of an antenna can be related to the 3 dB beamwidth using the approximate relationship

$$\text{Gain ratio} = 33\,000/(\text{3dB beamwidth in degrees})^2 = G \tag{9.17}$$

which gives $G = 33\,000/(123.6°)^2 = 33\,000/15\,276.96 = 2.16 \Rightarrow 3.3$ dB.

9.6.3 Number of Satellites per Plane

We now have the situation set up in Figure 9.33. Since each of the satellites will cover $36.42/360 \stackrel{\text{def}}{=} 0.1$ of the earth's circumference, we will need a minimum of 10 satellites in one plane.

9.6.4 Total Number of Satellites for a Global System

By the same logic used above, if 10 satellites are required to complete coverage around (say) the equator, 5 complete planes of satellites will be needed to complete the full global coverage. (Remember that one plane of satellites, if in a polar orbit, will have satellites on both hemispheres of the earth, some going northward and some southward. There will therefore be 10 slices around the earth made up of five planes of satellites). The total minimum number of satellites needed is therefore 50. It should be noted that this is an absolute minimum number. In addition to coverage gaps potentially existing, there will be a need to have spare satellites in orbit to take care of satellite failures.

Other architecture requirements can now be imposed. One could be the need for simultaneous coverage of any user by two satellites. The coverage of each satellite is unchanged, but the satellites must be half the distance apart so that two satellites always are in view throughout the constellation. A second architecture rule might be that no user is required to operate below 20° (rather than 10°). There are many other possible variations – for example, covering only the latitudes between $\pm 65°$ of the equator, inserting elliptical orbits to increase dwell time over a particular region. It is thus easy to see why some constellations need many dozens of satellites to complete the full architectural requirements.

9.7 Summary

Launching satellites into a geostationary orbit is a complex task that was not achieved successfully until six years after the first satellite was orbited in 1957. The quest for the

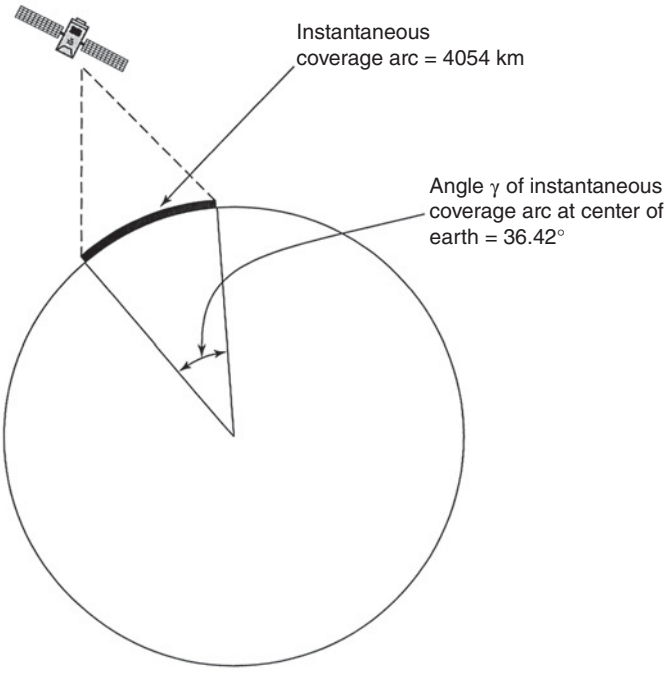


Figure 9.33 Coverage results from system design example in Section 9.6. The coverage of one satellite orbiting at an altitude of 750 km is shown. The coverage arc has been calculated to be 4054 km when a minimum elevation angle of 10° is assumed. The angle at the center of the earth that subtends the instantaneous coverage arc, γ, has been found to be 36.42°. To complete the coverage around the earth using this satellite configuration would require (360)/(36.42) satellites. Rounding to whole numbers yields a minimum number of 10 satellites per plane.

geostationary orbit grew from a paper by Arthur C. Clarke that established this orbit, at that time, as the prime location for communications satellites. If the satellite is geostationary, he argued, the earth station antenna need not be steerable, greatly reducing the cost of the system. Non-geostationary satellites, however, continued to be launched for a huge range of missions. TIROS satellites were the first to photograph the weather over the earth in sun synchronous orbits, orbits that precessed to match the rotation of the earth around the sun. TRANSIT satellites began the experiments into navigational aids that were space borne. MIDAS, the first early warning satellite; IRAS, the first infrared astronomy satellite; and Explorer satellites that probed the inner reaches of space above the earth continued the ever expanding list of experimental spacecraft launched into non-geostationary orbits for research purposes. *Molniya*, which means flash of lightning in Russian, was a satellite series that established the first regional satellite system in 1965 over what was the USSR, a few months after Canadian satellite *Anik* became the first domestic satellite system. The Molniya orbit has found many uses for a range of other satellite systems in non-geostationary orbit.

In the 1990s, a whole series of proposals arose for constellations of non-geostationary satellite communications systems, some in LEO and others in MEO. The thrust for moving the satellites down from geostationary orbit can be summed up in one word: power (or, more strictly speaking, four words: EIRP). The non-geostationary satellite

systems were aimed at the mobile user and so required that the user terminal antennas be essentially non-directional. Limited by radiation dosage limits, the telephone handsets for the mobile satellite systems could radiate only feeble amounts of power. To compensate for the low power and minimal G/T in the handsets, either technologically adventurous antenna systems were required in geostationary orbit or equally adventurous (as it turned out, but this time in an economic sense) constellations of non-geostationary satellites were required for a global mobile telephony system. Satellite constellation design is a complex mix of coverage requirements, capacity, and connectivity. The iterations in a design will also have to bear in mind the radiation environment in space caused by the Van Allen radiation belts and the need to have both an incremental design philosophy and replenishment options for the satellites that fail.

From the design examples shown, it is clear that geostationary satellites will always provide lower costs on a per bit basis than a satellite system in any other orbit until technology breakthroughs in smart antenna design for the user handsets. Even then, care must be taken that terrestrial systems have not taken away the customer base. This is a very interesting period in the evolution of telecommunications. The internet, and by implication digital data communications, has become the greatest growth area in the transfer of information globally: indeed, like a transportation infrastructure, the internet is seen as the key to a country's growth. Given an information system that provides a communications lifeline for remote and/or under-served regions of the globe, advances in education and telemedicine, so necessary to advance the living standards of those peoples – the excellently styled *Other 3 Billion* – will significantly increase their living standards. All inhabitants of the earth need to be given opportunities to advance. The extent to which satellites in non-geostationary orbit can fit into the global information infrastructure successfully will determine whether they have any commercial future in this field. That they have a future in navigation (GPS) and geographic information systems is assured; let us hope the next decades of the twenty-first century also bring success to the new NGSO systems designed to provide communications – especially the internet – to the whole world.

Exercises

9.1 Where is the optimum launch site located to minimize launcher energy requirements when a NGSO satellite is launched into
- **a.** A circular equatorial orbit
- **b.** A circular polar orbit
- **c.** An inclined orbit with an inclination angle of 27°.

9.2 A constellation of NGSO satellites provides internet access from any point on the earth's surface within latitudes 70° North and South. The round trip delay time from a user terminal to an earth station that connects to an ISP must not exceed 50 ms to ensure that the link can employ standard transmission control protocol/internet protocol (TCP/IP) that time out after 60 ms.
- **a.** What is the maximum slant path length to a satellite that meets this requirement?
- **b.** If user terminals and earth stations are restricted to operate above an elevation angle of 10°, what is the maximum orbital altitude of the satellites?

9.3 The satellite digital radio company Sirius began broadcasts over the United States in 2002 using three satellites in NGSO. The orbits, called *tundra orbits,* were highly elliptical and had an orbital period of one sidereal day. The three orbits were spaced by 120° in longitude and had an apogee over Hudson Bay, in Canada. XM Satellite Radio began broadcasting to the United States from two GEO satellites with longitudes 85°W and 115°W. XM radio also installed hundreds of terrestrial transmitters in cities throughout the United States.

a. Explain why Sirius selected the tundra orbit for its satellites, and why the apogee was set over the coordinates of Hudson Bay.

b. Why did XM Radio select GEO satellites with longitudes 85°W and 115°W for its SDAR service? Why were so many terrestrial radio transmitters required?

c. Sirius and XM Radio merged in 2007. Later SiriusXM satellites were all launched to GEO orbit. Why was the tundra orbit abandoned after the merger?

9.4 Explain the nature of the Van Allen radiation belts and why they present a danger to the communication systems on board satellites.

a. Why are LEO satellites generally restricted to orbital altitudes below 1200 km? What is the minimum safe orbital altitude for a MEO satellite?

b. The first experimental telecommunication satellites Telstar I and II were launched into elliptical MEO orbits with altitudes that varied between a perigee at 952 km and apogee at 5933 km. The electronics on the satellite failed after a few months. Explain why this happened.

c. GEO satellites must travel through radiation belts to reach their final orbit at an altitude of 35 786 km. What precautions can be implemented to minimize damage to the onboard communication systems?

9.5 Earth station antennas for satellites in NGSO must either be near omnidirectional or have the ability to track satellites across the sky.

a. Explain why phased array antennas are preferred over omnidirectional antennas for fixed installations.

b. What causes the cost of a phased array antenna to be higher than a reflector antenna?

c. Why is it difficult to employ a phased array antenna in a handheld device like a satellite phone?

9.6 Double hop GEO satellite links (two GEO satellite links in series) are not often used for real time communication because of the long round trip delay. An intersatellite link (ISL) that connects one GEO satellite to another using a microwave or optical link can reduce the time delay. This question examines how much the delay can be reduced by the use of an ISL when two earth stations are located on opposite sides of the earth.

In this scenario, three earth stations are located on the equator at longitude 60°W in Brazil, 30°E in Africa, and 120°E in Indonesia. GEO satellites are conveniently located at longitudes 15°W and 75°E. We will denote the satellite at 15°W as GEO #1 and the satellite at 75°E as GEO #2.

a. Calculate the path length from each satellite to each earth station.

b. Calculate the round trip delay time for a signal that is transmitted from the Brazil earth station via GEO #1, the earth station in Africa, GEO #2, the earth

station in Indonesia, and back to the earth station in Brazil, reversing the route of the outbound signal.

c. Calculate the distance between the two GEO satellites.

d. An ISL is established between the two GEO satellites. Calculate the new delay time for the scenario in part (b) when the ISL is used. How much time is saved by using the ISL?

e. A land line exists between a drone operator's location in the United States and the earth station in Brazil. The delay time for the land line is 50 ms. A second land line connects a drone control station to the earth station in Indonesia with a delay of 10 ms. How long does it take for the drone operator to receive acknowledgment that the drone has received a control instruction when (i) the double satellite hop link is used and (ii) the ISL is used?

9.7 This question examines some aspects of multiple access in VSAT networks.

a. Explain what MESH and STAR architectures are in a VSAT network.

b. Give two advantages and disadvantages of the MESH and STAR architectures.

c. What are the most commonly used multiple access schemes in satellite communication systems?

d. Explain how a random access technique like ALOHA can be used in a VSAT network.

e. What is the major disadvantage of random access when used for traffic?

f. Why has a time division multiplexing (TDM) approach been adopted for most downlink applications for digital VSAT and internet applications to small terminals?

9.8 A DTH-TV system needs to select a receiving antenna to use with its system.

a. Calculate, and set down in tabular form, the gain in dB, the 3 dB beamwidth, and the 1 dB beamwidth (in degrees) of antennas with the following diameters: 0.4, 0.6, 0.8, and 1.0 m. Assume a receive frequency of 12.2 GHz, an aperture efficiency of 70%, and that the 1 dB beamwidth in degrees is half of the 3 dB beamwidth.

b. If users are able to point their antennas to within $\pm 0.5^\circ$ and require a minimum gain of 30 dB, what antenna diameter range is available to the users?

c. Given this acceptable range of antenna diameters, which one of these antenna diameters would you choose, stating your reasons?

References

Allnutt, J.E. (2011). *Satellite-to-Ground Radiowave Propagation*, 2e. London, UK: The IET Chapter 2.

ANSI (1992). ANSI/IEEE C95.1-1992, *IEEE standard for safety levels with respect to human exposure to radio frequency electromagnetic fields, 3 kHz to 300 GHz,* IEEE.

Arstechnica (2015). https://arstechnica.com/science/2015/06/onewebs-constellation-of-700-low-altitude-satellites-will-be-built-by-airbus (accessed 9 September 2018).

Benedetto, J.M. (1998). Economy class ion-defying ICs in orbit. *IEEE Spectrum* 35 (3): 36–41.

Cellular (2008). https://en.wikipedia.org/wiki/Cellular_network (accessed 3 September 2018).

Chiavacci, P. (1999). The influence of phased-array antenna systems on LEO satellite constellations. *Microwave Journal* 42 (5): 282–290.
Clarke, A.C. (1945). Satellite communications systems. *Wireless World* 11 (2): 305–308.
COMSAT (1994). https://en.wikipedia.org/wiki/COMSAT (accessed 7 October 2018).
Draim, J.E. (1998). Optimization of the Ellipso™ and Ellipso 2G™ Personal Communications System. In: *Mission Design & Implementation of Satellite Constellations*, Space Technology Proceedings, vol. 1 (ed. van der Have). Dordrecht: Springer.
Ellipso (1998). www.ellipso.com (accessed November 1998, no longer accessible.) (see Draim, 1998 for more details on Ellipso).
ESA (2011). http://www.esa.int/Our_Activities/Space_Engineering_Technology/Radiation_satellites_unseen_enemy/(print) (accessed 1 September 2018).
Evans, J.V. (1997). Satellite systems for personal communications. *IEEE Antennas and Propagation Magazine* 39 (3): 7–20.
FCC (2018a). http://www.fcc.gov/oet/dockets (accessed 6 September 2018).
FCC (2018b). http://www.fcc.gov/oet/rfsafety (accessed 6 September 2018).
Foster, K.R. and Moulder, J.E. (2000). Are mobile phones safe? *IEEE Spectrum* 37: 23–28.
Globalstar (2018). www.globalstar.com (accessed 1 September 2018).
Iridium (2013). Appendix 1, Iridium Next Engineering Statement licensing.fcc.gov/myibfs/download.do?attachment_key=1031348 (accessed 7 October 2018).
Iridium (2018). www.iridium.com (accessed 1 September 2018).
Iridiumnext (2013). www.iridiumnext.com (accessed 4 September 2018).
ITU-R (1986). Recommendation P. 435-5, (2014) Prediction of sky-wave field strength between 150 and 1600 kHz. ITU-R, 2003 https://www.itu.int/en/ITU-R/space/workshops/cyprus-2014/Documents/Presentations/Hazem%20Moakkit%20-%20O3b.pdf (accessed 3 September 2018).
Morgan, G.D. and Morgan, W.L. (1993). *Principles of Communications Satellites*. New York, NY: Wiley.
Mursala, K. and Ulich, T. (1998). A new method to determine the solar cycle length. *Geophysical Research Letters* 25: 1837–1840.
NASA (1994a). https://www.nasa.gov/mission_pages/explorer/fast-facts.html (accessed 14 July 2014).
NASA (1994b). https://www.nasa.gov/mission_pages/chandra/main/index.html (accessed 16 July 2014).
NY Times (1989). http://www.nytimes.com/1989/04/25/science/what-s-next-for-star-wars-brilliant-pebbles.html?pagewanted=all (accessed 4 August 2018).
Orbcomm (2018). www.orbcomm.com (accessed 1 September 2018).
Samsung (2018). https://www.wired.com/2015/08/samsung-looks-join-satellite-internet-space-race (accessed 10 September 2018).
Satellite Today (2017a). http://www.satellitetoday.com/technology/2017/07/18/boeing-exec-surviving-slowdown-geo? (accessed 18 July 2018).
Satellite Today (2017b). http://interactive.satellitetoday.com/via/may-june-2017/future-invesstments-invsat-what-you-need-to-know/ (accessed 22 July 2018).
Schuss, T.J., Upton, J., Myers, B. et al. (1999). The IRIDIUM main mission antenna concept. *IEEE Transactions on Antennas and Propagation* 47 (3): 416–424.
SDI (2018). https://en.wikipedia.com/feature/a-new-hope-for-space-based-radar (accessed 3 August 2018).

Skybridge (1997). https://www.itu.int/newsarchive/press/WRC97/SkyBridge.html (accessed 12 July 2018).
Skyrocket (2018a). http://space.skyrocket.de/doc_sdat/intelsat-1.htm (accessed 10 June 2018).
Skyrocket (2018b). http://space.skyrocket.de/doc_sdat/o3b.htm (accessed 9 September 2018).
Skyrocket (2018c). http://space.skyrocket.de/doc_sdat/microsat-2.htm (accessed 10 September 2018).
Spaceflightnow (2001). https://spaceflightnow.com/atlas/ac156/010617ico.html (accessed 2 August 2018).
Spaceflightnow (2015). https://spaceflightnow.com/2015/05/01/next-round-of-u-s-optical-spy-satellites-to-start-launching-in-2018 (accessed 20 July 2018).
Spacenewsmag (2016). http://www.spacenewsmag.com/feature/a-new-hope-for-commercial-space-based-radar (accessed 1 August 2018).
Teledesic (2000). https://en.wikipedia.org/wiki/Teledesic (accessed 1 October 2018).
Teledesic (2017). http://www.astronautix.com/t/teledesic.html (accessed 10 June 2018).
Terada, M.A.B. (1999). Reflector antennas. In: *Wiley Encyclopedia of Electrical and Electronics Engineering*, vol. 18 (ed. J.G. Webster), 360–379. New York: Wiley.
Cold War. (2008). http://www.coldwar.org/articles/80s/SDI-StarWars.asp (accessed 2 August 2018).
Washingtonian (2016). https://www.washingtonian.com/2016/09/14/70-percent-worlds-web-traffic-flows-loudoun-county (accessed 1 October 2018).

10

Direct Broadcast Satellite Television and Radio

Geostationary satellites have carried television program material almost since their inception for commercial service in the late 1960s. The limited bandwidth of undersea cables designed for voice communications prevented their use for video signals, so live television signals could not be transmitted beyond the limits of any continent at that time. AT&T engineered microwave links in the 1950s that allowed video signals to be distributed throughout the United States, and other countries quickly followed suit to establish national television networks. The first time that a geostationary earth orbit (GEO) satellite was used extensively for video transmission was for the Tokyo Olympic Games in 1968, which were broadcast live in the United States using a link through an early Intelsat satellite over the Pacific Ocean.

The growth of cable television (CATV) systems in the United States in the 1970s encouraged the use of domestic North American satellites for distribution of cable TV signals. In the 1970 and 1980s when analog transmission techniques were used, one or two video signals were transmitted over a 36 MHz C-band transponder. The baseband signals were in the National Television Standards Committee (NTSC) color TV format developed in the 1950s, which remained in use for terrestrial broadcasting in the United States until 2009. Cable TV providers wanted more programming for their customers, so satellite distribution moved to compressed digital transmission allowing a single Ku-band transponder on a GEO satellite to send up to 10 digital video signals to thousands of independent cable systems distributed throughout the country. The digital signals were encrypted much more effectively than the earlier analog signals, making unauthorized reception more difficult. Distribution of video signals by satellite to an entire continent is an example of point to multipoint transmission, better known as broadcasting. This is what GEO satellites do best. A large fraction of all the transponders in most of the world's domestic and regional GEO satellite systems are devoted to the distribution of video signals.

Figure 10.1 shows the earth station complex at Virginia Tech, in Blacksburg, Virginia, which is equipped to uplink analog and digital video signals for distribution to educational institutions. The large antenna in the left of the photograph is a steerable 9 m C-band Cassegrain antenna used to transmit analog video frequency modulation television (FM-TV) C-band signals to a domestic satellite. The antenna is equipped with C-band transmitters connected to orthogonal polarization ports on the antenna feed, allowing simultaneous transmission of two video signals to two orthogonally polarized transponders on the satellite. The antenna was used for many years to distribute graduate classes to sixteen locations in and around the Commonwealth of Virginia.

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

Figure 10.1 Virginia Tech earth station. The two Cassegrain antennas in the left of the photograph are a 9 m C-band steerable antenna and a 5 m Ku-band steerable antenna with dual polarization uplink transmitters. The parabolic torus antenna in the center of the picture is a Simulsat antenna that can receive signals from seven satellites simultaneously. At right is a repositionable Ku-band Cassegrain antenna. Source: Photo credit: Tim Pratt. For a color version of this figure please see color plate section.

The smaller antenna in the middle of Figure 10.1 is a 5.5 m Ku-band steerable Cassegrain uplink antenna used to transmit multiple digital compressed video signals. The university moved to compressed digital transmission at Ku-band in the early 1990s when the leasing price of C-band transponders suddenly increased following the failure of a large domestic C-band satellite. The antenna at the right of Figure 10.1 is a *Simulsat*® antenna. The reflector is a parabolic torus antenna aligned with the GEO arc and has seven feeds. Each feed illuminates a section of the reflector, which approximates a paraboloid. The antenna is used by the university's campus cable TV network to receive video signals from seven GEO satellites. A repositionable 5 m Cassegrain antenna is visible in the right of Figure 10.1, and there are several smaller receive only antennas in the background.

Video distribution and direct broadcast satellite television (DBS-TV) have become the major source of revenue for the satellite communications industry. As seen in Figure 1.2 in Chapter 1, the worldwide income from all satellite services in 2016 was US$122B, of which US$98B was earned from direct to home satellite TV. Satellites designed for DBS-TV are among the largest and heaviest commercial geostationary satellites and revenue earned from entertainment television have become a major driver in the satellite communications industry. The dominance of video transmission via GEO satellites will be challenged from 2020 on by the growth in low earth orbit (LEO) satellite constellations with thousands of smaller satellites used for internet access.

In 2001, two direct broadcast satellite (DBS) radio services, Sirius and XM Radio began operation in the United States using S-band frequencies. The satellites provide a

wide range of radio programming, aimed primarily at drivers of road vehicles. Repeaters are used in city areas to overcome the problem of satellite visibility around tall buildings. Sirius and XM merged in 2007 to form Sirius-XM (Sirius_XM_Holdings 2007).

10.1 C-Band and Ku-Band Home Satellite TV

In the early 1980s, the development of low noise *Gallium Arsenide Field Effect Transistor* (*GaAsFET*) amplifiers for C-band low noise amplifiers (LNAs) and improved threshold extension frequency modulation (FM) demodulators for video signal receivers allowed much smaller diameter antennas to be used to receive C-band FM video signals distributed through GEO satellites. A market rapidly developed in the United States for home satellite TV systems using 3 and 3.6 m dish antennas (10 and 12 ft diameter) and set-top receivers that could receive the video signals from domestic GEO satellites. These were particularly popular in rural areas where terrestrial television broadcasting offered at best three channels from the major networks. At that time, the signals were not encrypted, so owners of satellite dishes could receive a wide range of television programming free of charge. The cable TV industry in the United States became concerned about the growth of home satellite TV receiving systems, and tried to have Congress pass laws that would ban their use. Congress did not pass such laws, but instead told the industry to scramble (encrypt) their signals and to charge customers for the descrambling information, and then passed laws that made the unauthorized use of descrambling equipment illegal. An estimated 4 or 5 million C-band and Ku-band FM satellite TV systems were sold in the United States by the time Ku-band DBS-TV arrived in the 1990s, using digital transmission and 0.5 m dishes, and offering more channels than the earlier system at a comparable price (Satellite Television 2018; History of Satellite TV 2016).

DBS-TV originally started in Europe and the United States in the 1980s using analog FM transmission in Ku-band. Initially, satellite TV was much more successful in Europe than in the United States, possibly because there were fewer alternative sources of TV programming in Europe. Most European countries offered only a handful of broadcast TV channels, and cable service has never been as widespread in Europe as in the United States. Nevertheless, at least one European satellite based direct broadcast TV system failed during the 1980s, and two satellites built for a US company intending to enter the DBS-TV field were sold to a European company. The market for DBS-TV systems grew slowly in the 1980s, and then very rapidly after the introduction of high capacity digital DBS-TV satellites in the 1990s.

10.2 Digital DBS-TV

In the 1990s, digital video transmission became feasible, and several systems were developed in the United States in the 12.2–12.7 GHz band allocated to DBS-TV services. The development of low cost Ku-band antennas and receivers, and high speed digital integrated circuits specifically for DBS television that incorporate quadrature phase shift keying (QPSK) demodulation, error control, decryption, and Moving Picture Experts Group (MPEG) compression made DBS-TV practical. The digital signal processing is incorporated in a single integrated circuit that implements the *digital video broadcast*

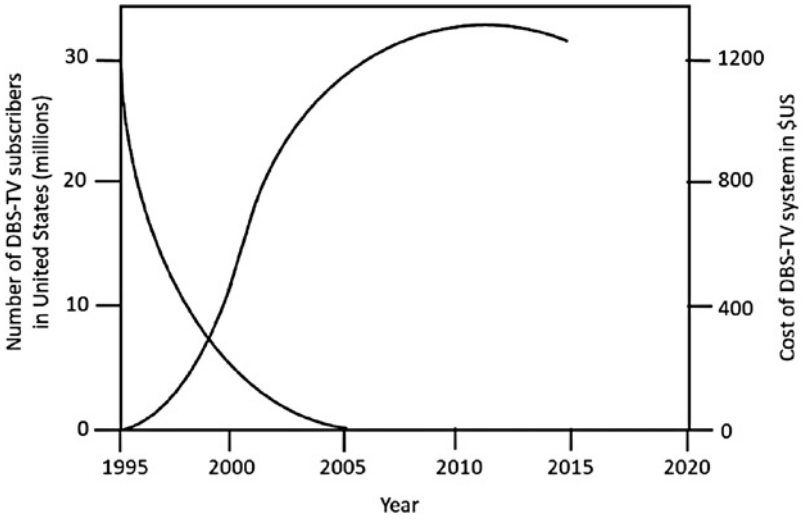


Figure 10.2 Growth in subscribers to US DBS-TV services. Growth flattened out by 2015 as internet TV service became available. The cost to buy and install a DBS-TV receiving system was initially high, but by 2005 both DirecTV and Dish Network would install a system at no cost to customers who signed a two year contract.

standard (DVB) used by most DBS-TV systems. The large volumes in which DBS-TV receivers have been manufactured have allowed the cost of a receiving system to be reduced steadily since the start of DBS-TV service.

Figure 10.2 shows the rapid growth in subscribers to DBS-TV systems that took place between the start of DirecTV in 1995 and Dish Network in 1996, and 2007 when growth started to slow down (SBCA 2013). Figure 10.2 also shows how the cost of a typical home DBS-TV installation decreased during this period. By year 2005, installation of a DBS-TV receiving system was offered free to US customers of Dish Network and DirecTV on signing a two year contract. By 2017 the number of subscribers to DirecTV and Dish network started to drop as internet TV services offered an alternative to both cable and satellite TV.

DirecTV, a fully digital DBS-TV system was developed by a consortium of companies led by Hughes Electronics Corporation, owned at that time by General Motors, and began limited service in 1994 with a single GEO satellite at 101°W longitude. The first satellite, called DBS-1, was launched in December 1993, and was followed by two more satellites, DBS-2 and DBS-3, in 1994 and 1995. A fourth satellite was added in 1999, and a fifth satellite was launched in 2000 with a transmit antenna capable of providing spot beams, using locations of 101°W and 109°W. After several mergers and acquisitions, DirecTV was owned by AT&T in 2014. By 2015, DirecTV owned or operated 18 GEO satellites at six orbital locations; five of the early satellites had been de-orbited (DirecTV 2018).

DirecTV spent about US$1B to develop their system and needed 2 million customers to break even. That number was quickly passed and the *Wall Street Journal* described DirecTV as "one of the most successful business ventures of the (twentieth) century." The early DBS-TV satellites served the entire United States from one GEO location, using relatively broad beams. In 2000, new legislation allowed DBS satellite operators to

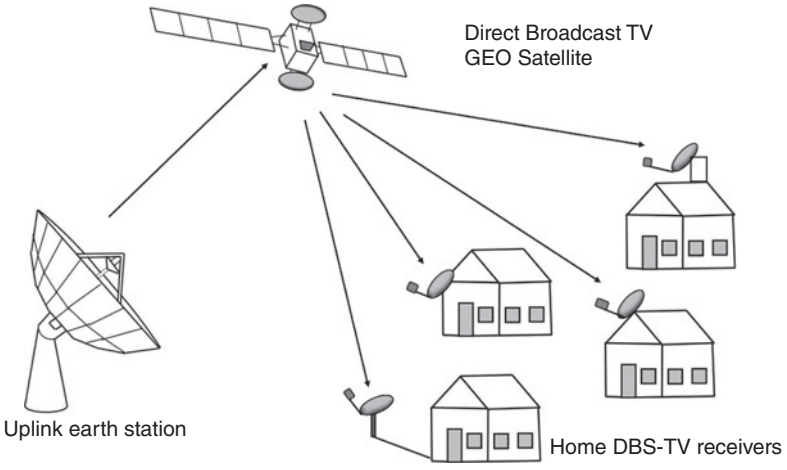


Figure 10.3a Illustration of a Ku-band DBS-TV system. The uplink earth station typically has a large antenna and several powerful transmitters to provide high CNR in the satellite transponders. The DBS-TV home antennas can be mounted on a pole outside the home or on the exterior of the building. A clear line of sight in the direction of the satellite is required.

compete with cable television companies by supplying news from local TV stations to specific regions. This made spot beams serving only a part of the United States very desirable, and later generations of DirecTV and Echostar satellites incorporate large transmit antennas that can generate up to 72 spot beams on centers of population in the United States.

The Echostar Communications Corporation started service with its Dish Network in March 1996 with a single satellite at 119°W. In 2015 there were 24 Echostar satellites in orbit, at longitudes between 61.5°W and 119°W. Echostar has remained a single commercial entity for more than 20 years, in contrast to DirecTV's many different owners (Echostar 2018).

Figure 10.3a shows the general concept of a DBS-TV system. Figure 10.3b shows more detail of the signal processing that is needed to send video and audio signals from a studio or a recording through a GEO satellite to the thousands (or millions) of DTH-TV receiving stations.

Figure 10.3b shows the signal processing required to send a television program from a studio or a prerecorded data base to a millions of home DBS-TV receivers. The blocks in the figure are the topics of following sections in this chapter. The input to the transmitter consists of digital video and audio signals, and associated data, from many television programs. The video and audio signals are compressed using MPEG-2 or MPEG-4 compression, and formed into packets and frames up to 64 000 bits in length that can contain many multiplexed channels. Error correction encoding is applied to the bit stream, which is then encrypted to prevent unauthorized access. Modulation is QPSK or eight phase shift keying (8-PSK) at an intermediate frequency (IF), under the control of the transmitting end of the link, and can be changed by sending a control packet to the uplink station and all receiving stations. The forward error correction (FEC) rate can also be changed if needed. The IF signal is sent to the transmitter for up conversion and transmission to the satellite, along with other TV programs. The receiving end of the link mirrors the transmitting end in reverse, with modulation and

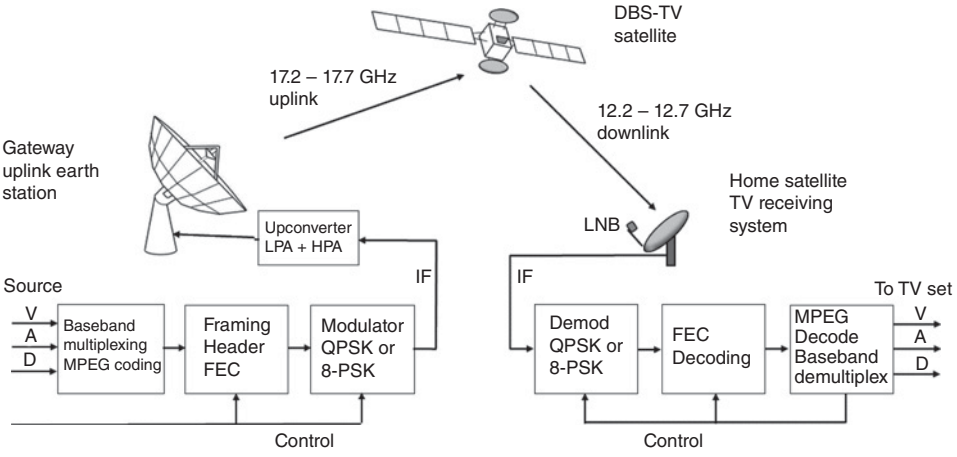


Figure 10.3b Simplified diagram of the signal processing in a DTH-TV link. Video (V), audio (A), and data (D) bit streams are compressed using MPEG-2 or MPEG-4 compression techniques and formed into packets and frames. Forward error correction (FEC) and modulation method can be changed on command by the control line. The data stream is encrypted to prevent unauthorized access. The modulated signal is sent to the transmitter at an IF frequency, often 700 MHz. The home receiver uses a low noise block converter to down convert the received signals to IF and sends them to the satellite receiver, which reverses the operations carried out in the transmitter.

FEC rate set by a controller in the receiver that decodes control packets sent from the uplink transmitter. The output of the home satellite receiver consists of video, audio, and data that contains the selected TV programs and related information and a data stream that is used to send information to the controller in the receiver.

DirecTV grew its customer base in the United States very rapidly through 2010 when the number of subscribers flattened out to twenty million in 2012 and held steady through 2015 with a small drop by the end of 2017. Echostar had fourteen million subscribers in the United States by 2011 and kept its market share through 2015, giving a total of 32.5 million DBS-TV subscribers in the United States in late 2017. The drop in DBS-TV subscribers is attributed to the growth of internet TV systems (DBS-TV growth 2018). In Europe, SES is the major provider of DTH-TV. SES was founded as Société Européenne de Satellites (European Society for Satellites) in 1985 at the Chateau de Betzdorf, Luxembourg, a formal royal hunting lodge. SES owned or operated 50 DBS-TV satellites in 2018 (SES 2018).

Worldwide, India had the highest number of DBS-TV subscribers in 2017 at 67 million (India DBS-TV 2017). Other countries with significant DBS-TV systems include Russia and Brazil with an estimated 17M subscribers each by 2020. An estimated 138 counties had satellite TV service in 2017. SES claims to reach 156 million households with video transmissions, either via DBS-TV or cable systems (SES 2018). DirecTV serves 10M customers outside the United States.

10.2.1 Examples of DBS-TV Satellites Serving the United States

Table 10.1 summarizes the major parameters of two of the DBS-TV satellites serving US customers in 2016.

Table 10.1 DirecTV-14 and Echostar 16 satellite specifications

DBS-TV satellite	DirecTV 14	Echostar 16
Location in geostationary orbit	99°W longitude	61°W longitude
Launch date	December 2014	November 2012
Satellite manufacturer	Space Systems Loral	Space Systems Loral
Launch Vehicle	Arianne 5	Proton M
Type designation	SSL 1300	SSL 1300
Frequency band	Ka-band, R-band	Ku-band (12.2–12.7 GHz)
Transponders	76 active Ka-band, 18 R-band	32
Bandwidth	15 MHz	
Solar power system at beginning of life	20 kW	16 kW
Station keeping thrusters	Plasma Thrusters	Plasma Thrusters
Mass at Launch	6305 kg	6650 kg
Antenna beams, EIRP	Conus: 48.8–56.6 dBW	Conus coverage
	Spot 57 dBW	67 spot beams, 56–59 dBW
Polarization	LHCP and RHCP	LHCP and RHCP
Modulation, Coding	Conus beam 8-PSK, 2/3 rate FEC	Conus beam 8-PSK, 2/3 rate FEC
	Spot beams 8-PSK, 2/3 or 5/6 rate FEC	Spot beams 8-PSK, 2/3 or 3/4 rate FEC

The 12.2–12.7 GHz band was set aside for exclusive use by DBS-TV satellites in geostationary orbit so that high power transponders could be used on specially designed DBS-TV satellites, extended later to include the 11.8–12.2 GHz band. A typical first generation DBS-TV satellite with continental coverage carried up to 32 transponders. DBS-TV satellites are typically large and heavy, generally use a three-axis stabilized design, and have large solar sails to generate up to 20 kW of DC power for the transponders and housekeeping.

Figure 10.4 show a large GEO direct broadcast television satellite under test prior to launch.

10.2.2 DBS-TV Receiving Antennas

The small receiving antenna has a wide beam, typically four degrees for a 0.45 m dish and three degrees for a 0.6 m dish, which forces wide spacing of DBS-TV satellites to avoid interference at the receiving antenna by the signals from adjacent DBS-TV satellites. A 9° spacing in the GEO arc has been adopted by the United States, which restricts the number of DBS-TV satellites that can be placed in geostationary orbit to serve the United States. The spacing was later reduced to 4.5°. In the 1990s the United States Federal Communications Commission (US FCC) successfully auctioned spectrum and

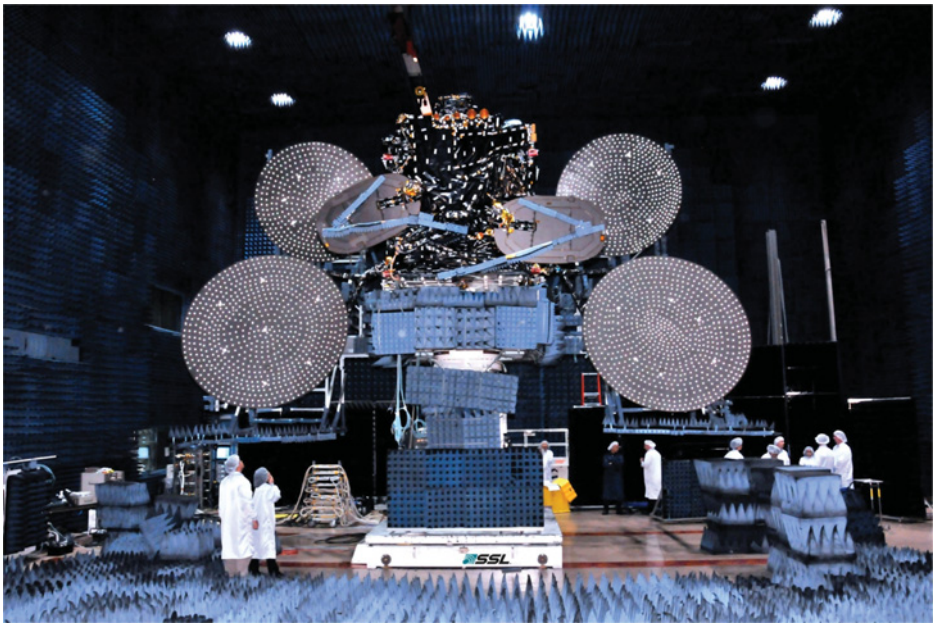


Figure 10.4 A large GEO direct broadcast television satellite under test prior to launch. The four large reflectors create conus beams and spot beams. Source: Image courtesy of SLS, © SLS 2018. For a color version of this figure please see color plate section.

orbital locations for DBS-TV satellites, raising hundreds of millions of dollars from companies that saw a profitable commercial venture.

A typical DBS-TV satellite carries up to 32 high power transponders each covering part of the 11.8–12.7 GHz broadcast satellite service (BSS) band, with bandwidths between 15 and 36 MHz. Spot beams with 15 MHz bandwidth are in widespread use. The satellites at each orbit location transmit in opposite hands of circular polarization (CP), allowing two signals to be overlaid. Signals with opposite hands of circular polarization are orthogonal, and DBS-TV earth station antennas can separate signals with opposite hands of circular polarization and send them to two separate LNAs.

10.2.3 DBS-TV Satellite Antennas

DBS-TV providers often have a cluster of two to four satellites at a single orbital location, with spacing of a fraction of a degree. Together the cluster covers the full DBS-TV band in two orthogonal polarizations, requiring more complex receivers with an orthogonal mode transducer (OMT) and two low noise block converters (LNBs) so that both hands of circular polarization can be received at the same time by using two LNBs and two receivers. Customers wanting to receive signals from more than one orbital location need an antenna with multiple feeds. Reception from two satellites spaced 9° apart in GEO can be achieved with a larger antenna, 0.45 m × 0.6 m (18 in. × 24 in.) that produces two beams separated by the appropriate angle. A long focal length reflector is needed, which makes the dish curvature relatively low when multiple feeds are used. Figures 10.5a–c show examples of DBS-TV receiving antennas. Three feeds can be seen in the photograph of a Dish Network antenna in Figure 10.5b, and a DirecTV antenna

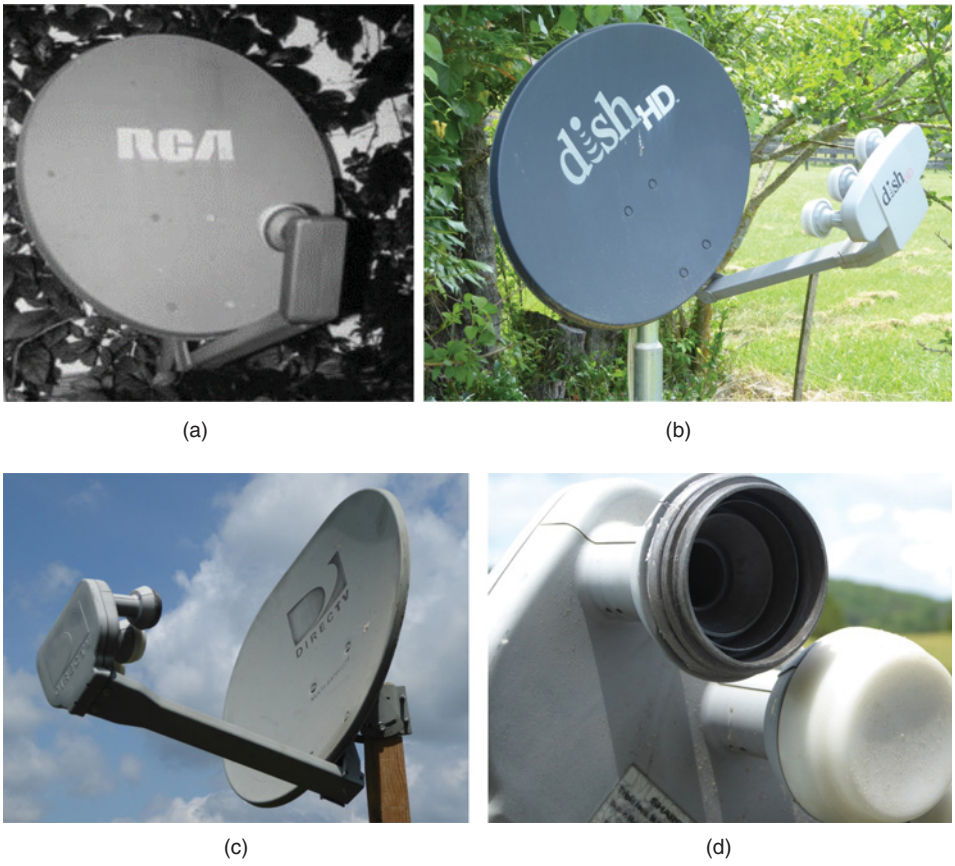


Figure 10.5 Examples of DBS-TV receiving antennas at the author's home (TP) in Blacksburg, Virginia. (a) Early DirecTV antenna from 1996 with a single feed. (b) Dish Network antenna with three feeds, circa 2016. (c) DirecTV antenna with three feeds circa 2013. (d) Corrugated horn of one of the feeds of antenna in (c) with protective cover removed. This is a conical version of the scalar feed illustrated in Figures 10.6a and 10.6b. Note that the single feed dish in (a) has a circular aperture whereas the antennas in (b) and (c) have elliptical reflectors. The wider dimension in the horizontal plane is needed to accommodate the multiple feeds. Source: Photo credit: Tim Pratt. For a color version of this figure please see color plate section.

in Figure 10.5c. Additional feeds can be added for other orbital locations and frequency bands, for example, to receive signals from Ka-band and R-band transponders.

DBS-TV receiving antennas are typically an offset parabolic reflector design with the feed below the antenna aperture. The offset feed design eliminates blockage of the aperture by the feed, which occurs in symmetrical reflector antenna designs, and improves the aperture efficiency of the antenna, and therefore increases its gain. Offset fed parabolic reflectors have a beam squint effect in the plane of symmetry when operated in opposite hands of circular polarizations. For the 0.45–0.6 m diameter antennas widely used for DBS-TV reception in the United States, the left hand circularly polarized (LHCP) and right hand circularly polarized (RHCP) beams are squinted about 0.25° from the antenna's boresight. The 3 dB beamwidth of the antenna is typically between 3° and 4°, so the squint effect does not cause significant loss of gain.

An early single channel DirecTV receiving antenna mounted on the wall of a house is shown in Figure 10.5a, circa 1996. A Dish Network antenna with three feeds mounted on a post is shown in Figure 10.5b circa 2016. Figure 10.5c shows a DirecTV antenna circa 2013. Figure 10.5d shows the corrugated horn of one of the feeds in the antenna in (c) with its protective cover removed. Appendix B provides a review of microwave antenna types and properties.

10.2.4 DBS-TV Receivers

Figure 10.6a shows a block diagram of a single channel Ku-band LNB. In this version, the set top receiver sends a DC voltage along the coaxial cable to select which of the LNAs in the LNB is switched on, to determine whether the RHCP or LHCP signal is received. A circularly polarized wave is composed of two orthogonal E-field waves in phase quadrature. The dielectric plate polarizer oriented at 45° to the probes in the waveguide slows down the E field component of the circularly polarized wave in the plane of the plate by a quarter of a wavelength relative to the E field component at right angles to the plate, introducing a 90° phase shift between the two components. At the output end of the polarizer the two E fields are at right angles and add to create a linearly polarized wave at 45° to the polarizer's plane. In the example shown in Figure 10.6a, the LHCP signal becomes a vertically polarized wave that is picked up by the vertical probe in the waveguide and fed to LNA 1, and the RHCP signal becomes a horizontally polarized wave that is picked up by the horizontal probe in the waveguide and fed to LNA 2. The radio frequency (RF) gain of the two LNAs is typically 45–55 dB.

In a more advanced LNB (often called a low noise converter, LNC) instead of using switched LNAs, which feed a common amplifier and downconverter chain, there are two identical chains with different local oscillator frequencies creating two different bands of IF signal. For example, a local oscillator at 11.3 GHz will create an IF signal in a band 900–1400 MHz. A second local oscillator at 10.2 GHz will create a second IF signal in the band 2000–2500 MHz. The two IF signals can be combined and sent over the same coaxial cable to the set top receiver. When multiple feeds are used in the receiving antenna,

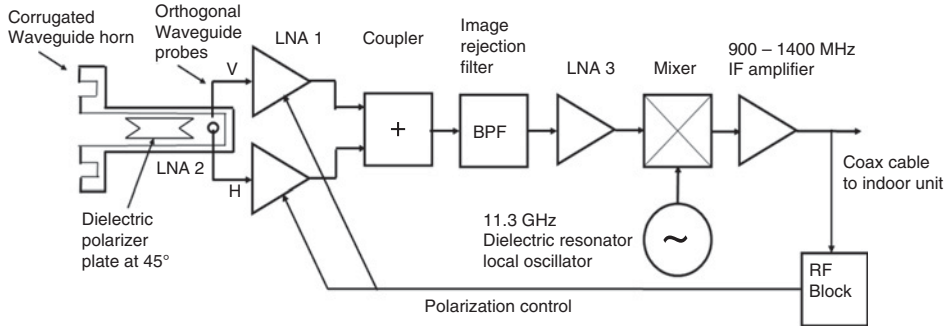


Figure 10.6a Single channel Ku-band LNB block diagram for satellite TV reception. Input is two circularly polarized signals in the 12.2–12.7 GHz band. The polarizer plate converts the LHCP signal to a vertical E-field linearly polarized wave and the RHCP signal to an H-field wave. The orthogonal probes in the circular waveguide extract the V and H polarized waves and feed two LNAs. A command signal sent via the coaxial cable switches on either LNA 1 or LNA 2. The RF block is an inductor to prevent the IF signal entering the polarization line.

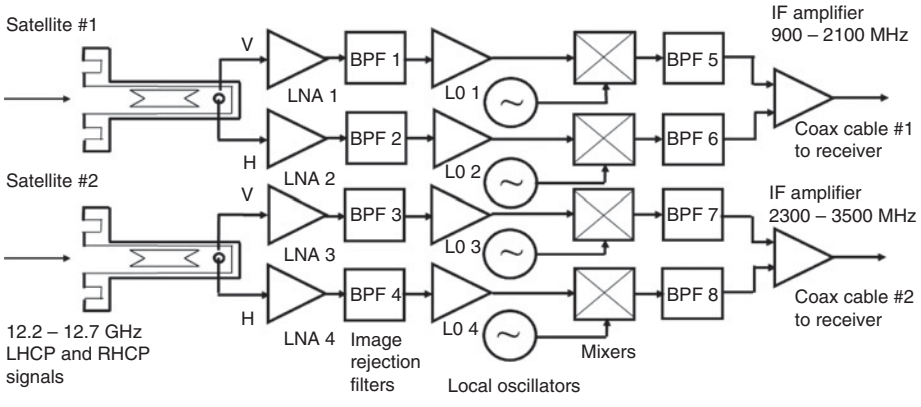


Figure 10.6b Four channel Ku-band LNC block diagram for simultaneous reception of two orthogonal circularly polarized signals in 12.2–12.7 GHz band from two satellites. Two coaxial cables are used to connect the LNC to the satellite receiver and are combined into a single line close to the receiver. Line amplifiers for the bands 900–2100 MHz and 2300–3500 MHz can be inserted to extend the length of the cables. The local oscillator and BPF frequencies in the satellite #2 LNC can be the same as for the satellite #1 LNC if the indoor receiver has two separate inputs.

Local oscillator and filter frequencies

LO 1	11.3 GHz	BPF 1	12.2–12.7 GHz	BPF 5	900–1400 MHz
LO 2	10.6 GHz	BPF 2	12.2–12.7 GHz	BPF 6	1600–2100 MHz
LO 3	9.9 GHz	BPF 3	12.2–12.7 GHz	BPF 7	2300–2800 MHz
LO 4	9.2 GHz	BPF 4	12.2–12.7 GHz	BPF 8	3000–3500 MHz

more than one coaxial cable can be used and there can be IF signals in bands from 900 to 3500 MHz.

Figure 10.6b shows a block diagram of a four channel LNC with two antenna feeds that can receive LHCP and RHCP signals from two satellites simultaneously. The output of the LNC is conveyed to the indoor unit by two coaxial cables; one cable carries IF signals from satellite #1 in bands 900–1400 MHz and 1600–2100 MHz, the second cable carries IF signals from satellite #2 in bands 2300–2800 MHz and 3000–3500 MHz. The indoor unit receiver has four *tuners,* which select specific transponder outputs within the IF band and downconvert each of the four IF signals to a second IF frequency, for example, 700 MHz, at which the signals are digitized with an analog to digital converter (ADC) and processed digitally in the rest of the receiver. The TV viewer can record three programs for later viewing while watching a fourth program in real time.

The corrugated horn is a high efficiency feed for the offset reflector antenna producing a circularly symmetric field pattern. The short circular waveguide that contains the polarizer ensures that losses before the first LNA are kept to a minimum. The remainder of the LNB is built on microstrip, a widely used printed circuit technique for microwave frequencies. The image rejection filter is a series of short parallel bars that produce the required bandpass frequency response. The dielectric resonator local oscillator in a typical home satellite LNB does not have good frequency or phase stability, but is adequate for reception of QPSK and 8-PSK signals. Professional quality LNBs require better phase stability to receive 16-APSK (amplitude phase shift keying) and 32-APSK signals so use

a crystal oscillator and multiplier configuration for the local oscillator, at much higher cost.

The entire front end of the receiver is located at the antenna feed in the form of an LNB or LNC to minimize loss of signal and hence to maintain the lowest possible system noise temperature. The housings for the LNBs can be seen in Figure 10.5a–c immediately behind the feeds. The high gain LNB can drive 100 m of coaxial cable without any reduction in signal quality. Where longer cable runs are needed, amplifiers for the 900–3500 MHz bands can be used to boost the signal strength. The set top box accepts the entire 500 MHz band from each LNB and separates out the individual transponder IF frequencies. Any one of these frequencies (and the corresponding polarization) can be selected on demand by the user.

Antennas with multiple feeds and multiple LNBs use different local oscillator frequencies to generate 500 MHz wide IF signals within the 900–3500 MHz frequency band. Figure 10.6b illustrates a four channel LNC that receives Ku-band signals from two satellites using an antenna with two feeds. There are four identical channels in the LNC, with different local oscillator frequencies to create four different IF frequencies in a band extending from 900 to 3500 MHz. Attenuation in the coaxial cable is greater at the higher IF frequencies requiring line amplifiers when a long run of cable is needed. Separating the signals into two bands avoids the need for a single wideband amplifier and allows the insertion of a single amplifier for the higher frequency band when adequate signal is available in the lower band. The set top box can have up to four tuners to simultaneously select these signals and extract their content. Although that makes it possible to watch up to four TV programs at once (displayed as four smaller pictures on the TV screen) the main use is for recording programs that are transmitted at the same time, for viewing at a later time. This has the advantage that advertisements can be skipped.

10.2.5 DBS-TV Set Top Box

Figure 10.7 shows a simplified block diagram of the satellite receiver (indoor unit, set top box) for a single channel Ku-band DBS-TV receiving system with switching between

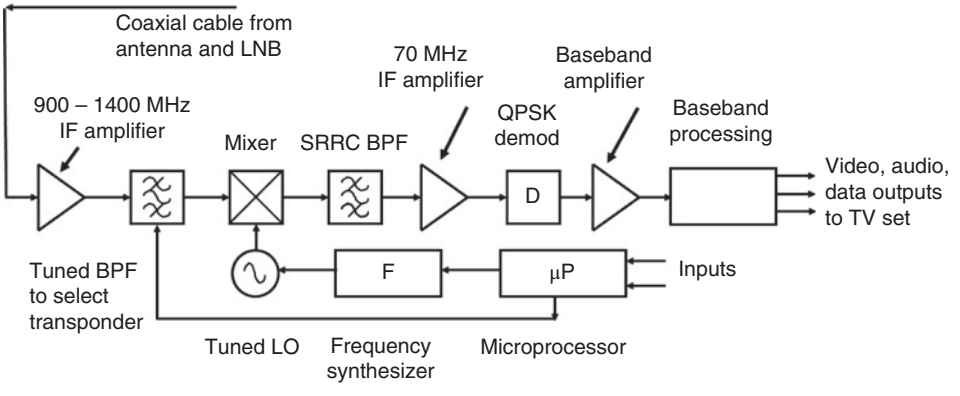


Figure 10.7 Simplified diagram of a single channel satellite receiver for Ku-band DBS-TV system. The receiver is controlled by inputs from the user with a remote control or from the receiver's front panel. The baseband processor extracts digital video, audio, and data from the received DVB-S formatted frames as described in Section 10.3.

polarizations. The LNB supplies a 500 MHz wide signal in the band 900–1400 MHz from the receiving antenna, which is pointed at a single satellite. The first IF signal from the LNB contains up to 16 transponder outputs from one of the two polarizations transmitted by the satellite. The IF signal is amplified and a specific channel is selected by the tuned BPF and then down converted to the second IF of 70 MHz. The SRRC BPF is a square root raised cosine filter with roll off factor α matched to the transmitted signal ($\alpha = 0.35$ for DVB-S format signals).

To select a particular TV program, the viewer enters a desired program channel number, for example, channel #200, into the set top box microprocessor using an IR remote control, or the front panel of the receiver. The program channel number is converted via a stored look-up table in the receiver to a first IF channel frequency and local oscillator setting that is input to the frequency synthesizer. The signal from the required transponder is then selected by the receiver by setting the correct polarization at the antenna in the case of a single channel receiver and tuning the set-top local oscillator and tunable BPF frequency. The QPSK signal is then demodulated. The result is a multiplexed bit stream, typically at a bit rate of 22 Mbps for half rate FEC coding and QPSK modulation and a 30 MHz bandwidth transponder, which contains the bits for channel #200 and other video signals. The bit stream is encrypted and is in the form of frames that contain error control coding bits and data bits. The bit stream is processed to correct errors, de-interleaved, and decrypted. A digital demultiplexer then extracts the bits for the wanted channel – #200 in this example – sends them to an MPEG-2 decoder, and finally outputs digital video and audio signals to drive the TV set. Data signals are extracted and used to update the tables that relate RF signals to channel numbers of TV programs and hold the program guide.

The look-up table in the receiver that relates channel numbers to frequencies, polarizations, and instructions for the time division multiplexing (TDM) demultiplexer is downloaded from the satellite on a regular schedule. This allows the service provider to change the transponder that carries a particular signal, and to alter the mix of signals on a given transponder as required, without the customer being aware of the changes. The satellite is also used to address individual receivers and to load another look-up table that specifies which channels the user is authorized to receive. If the user fails to pay his or her bill to the service provider, the receiver will eventually be instructed to show only a message that it has been disconnected for failure to make timely payment for the service. This process involves a smart card, which identifies each receiving system and enables decryption of the satellite signals. The high level of protection applied to the DBS-TV signals is intended to prevent unauthorized reception by users who have not paid monthly fees.

Pay-per-view channels are handled differently from broadcast channels. A customer wishing to buy a movie or a sporting event selects the desired channel and authorizes the system to make a charge. Most DBS-TV receive terminals have no uplink capability, and the customer must therefore use a terrestrial telephone circuit or the internet to order pay-per-view programming. The cost of the pay-per-view event is then added to the customer's monthly bill. A connection between the DBS-TV receiver and the public switched telephone network (PSTN) at the customer's home, or a connection to the internet, allows the use of a satellite TV remote control to order pay-per-view services. When connected to the PSTN or the internet, DBS-TV receivers typically send information to the satellite TV provider. Such information might include the pattern of channels that the customer selects and watches, which is valuable data for advertisers.

10.3 DVB-S and DVB-S2 Standards

Between 1991 and 1997 the European Telecommunication Standards Institute (ETSI) developed a set of standards for digital television broadcasting, covering satellite, cable, and terrestrial broadcasts (ETSI 1997, 2003, 2009, 2015). Digital video broadcast (DVB) standards are maintained by the DVB Project, an international industry consortium with more than 270 members, and are published by a Joint Technical Committee (JTC) of ETSI. The DVB standard defines the baseband processing, including video and audio compression, FEC, and framing of digital TV signals, and also the available modulations. The original DVB-S standard was first published in December 1994 and finalized in 1997. The DVB-S standard was used in slightly different forms by DirecTV and Dish Network in the United States, and in many other countries worldwide for the transmission of standard definition (SD) digital television in the first generation of digital DBS-TV systems. The DVB-S2 standard was published in 2003, with various updates in succeeding years. The DVB-S2 standard allows for multiple modulations and FEC rates, and was designed for high definition (HD) digital television and also to provide optimum performance with systems that have two-way capability. The capacity of a transponder can be increased when using the DSB-S2 standard by 30% relative to the earlier DVB-S standard. One target of DVB-S2 is internet access by satellite, with the concept that specialized integrated circuits could be developed for use in digital television receivers and also in internet access devices to provide access to both services (ETSI 2015). The DBV-S2 standard is backward compatible with the DVB-S (the -S stands for satellite) standard, but achieves the greatest advantage when a return channel is available from the receiving station that can report the carrier to noise ratio (CNR), allowing the transmitting station to modify the modulation and coding rate of the signal in real time (*adaptive coding and modulation*, ACM).

10.3.1 Television Broadcasting

Terrestrial television broadcasting in the United States changed from analog to digital in 2009. The Advanced Television Standards Committee (ATSC) established the parameters of the signals, with the specific requirement that the RF bandwidth must not exceed 6 MHz so that ATSC TV signals could be delivered in the same 6 MHz channels used for the earlier NTSC TV transmissions (ATSC Standards 2018). There are significant differences between ATSC and DVB-S signals that make receiving equipment incompatible. All TV sets built after 2008 for use in the United States incorporate a receiver for ATSC signals, so an external receiver (a set-top box) is required for reception of satellite TV. Other countries use different digital TV standards (Alencar 2009).

The DVB-S and DVB-S2 standard differs significantly from the DVB standards for terrestrial and cable television distribution. The transponders on many DBS-TV satellites are non-linear, a characteristic of high power traveling wave tube amplifiers (TWTAs), whereas both terrestrial and cable networks can maintain linearity in their distribution systems. The modulation specified in the DVB-S and DVB-S2 standards for DBS-TV are all some form of PSK. QPSK (4-PSK) and 8-PSK are the most popular choices for DBS-TV, with 16-APSK and 32-APSK as options where the transponder is linearized. The advantage of QPSK and 8-PSK is that the signal has constant amplitude and a circular constellation. Operating the transponder close to saturation with small output backoff

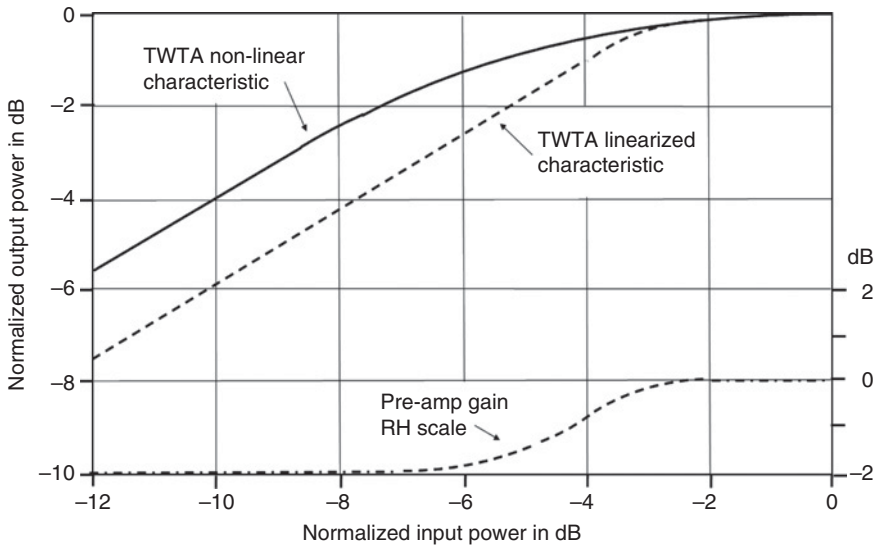


Figure 10.8 Non-linear characteristic of typical TWTA and linearization with a compensating pre-amplifier. The TWTA is linear until the output reaches a power level 2.5 dB below the saturated power. Saturation of the transponder occurs at a normalized value of 0 dB in the figure. The pre-amplifier gain compensates for the TWTA non-linearity between output levels of −2.5 dB and −0.5 dB below saturation. This linearizes the TWTA characteristic for operation up to 1.0 dB below saturation. The TWTA gain is normalized to 0 dB in the figure.

causes some distortion of the RF waveform, but is feasible with a single RF signal as is commonly the case with DBS-TV. If more than one signal is transmitted to the transponder, more backoff is needed to avoid intermodulation problems unless the TWTA in the transponder is linearized. With 16-APSK and 32-APSK the signals have multiple amplitudes. Non-linearity in the transponder high power amplifier (HPA) compresses the larger amplitude signals and reduces the separation between low and high voltage states in the constellation, leading to higher bit error rate (BER).

Figure 10.8 shows the characteristic of a typical non-linear transponder. The transponder is quasi-linear up to an output power 2.5 dB below saturation. Linearization of the transponder can be achieved by compensating for transponder non-linearity with a pre-amplifier that has increasing gain as input power is increased, as illustrated in Figure 10.8. The dynamic range of the compensating amplifier is small, less than 3 dB, but extends the quasi-linear operating range of the transponder TWTA to a point 1 dB below its saturated output. When the linearized transponder is operated with 1 dB backoff, there is minimal intermodulation between multiple carriers or compression peaks in the signal amplitude. The pre-amplifier can be located on the satellite, or at the uplink earth station.

The DVB-C and DVB-T standards (-C stands for cable and -T stands for terrestrial) assume linear amplifiers in the distribution network and employ 64-QAM (quadrature amplitude modulation) and higher order QAM modulations for cable systems, with *coded orthogonal frequency division multiplexing* (COFDM) for terrestrial broadcasting. Most of Europe and many other counties use orthogonal frequency division multiplexing (OFDM) for terrestrial television broadcasting. The ATSC standard for terrestrial

broadcasting in the United States uses a different format, eight level *vestigial sideband* (VSB) amplitude modulation to meet the requirement to confine the RF signal to the 6 MHz channel bandwidth established for the earlier NTSC standard (Alencar 2009).

10.3.2 Baseband Compression of Video and Audio Signals

Digital television signals have very high bit rates at the output of a video camera and must be compressed to reduce the bit rate to a manageable rate for transmission over radio links. A color video camera is basically three cameras in one, delivering three digital output streams of the observed scene in three colors – for example, red, blue, and green. A standard definition TV signal has $1280 \times 720 = 923\,600$ pixels per frame and frame rate for DVB-S transmissions is 25 Hz. Hence 23 090 000 pixels (rounded to 23 M) are generated by the video camera each second. If we allocate 8 bits to each pixel in three colors, we need to transmit the video signal at $23 \times 24 = 552$ Mbps. The very high data must be reduced to allow transmission over bandwidth limited links, a process known as *compression*. The MPEG is an industry group formed in the 1980s to develop compression techniques and standards for digital video signals (MPEG 2018). There are many variants among the MPEG video compression standards. The most widely used are MPEG-2 for standard definition TV signals (1280×720 pixels) and MPEG-4 (1920×1080 pixels) for high definition signal, with a later version H264/MPEG-4ADV. MPEG-3 has not been used for video compression; it was overtaken by MPEG-4, but the audio portion is known at MP3 and is widely used to send compressed audio signals at 128 kbps. In contrast to CD recordings, which do not employ compression, MP3 is a lossy compression system and produces lower quality sound than a CD recording.

10.3.3 The DVB-S Standard

The first version of the DVB-S standard was published by ETSI in 1993 followed by revisions through 1997 (ETSI 1997). Both DirecTV and Echostar developed first generation digital satellite TV broadcasting satellites and systems for the United States that closely followed the DVB-S standard. Table 10.2 lists the important features of the DVB-S standard.

10.3.3.1 Modulation and Filtering

The modulation and filtering specified for DVB-S is QPSK with SRRC filters having $\alpha = 0.35$, with no alternatives. The later standard, DVB-S2, allows higher order PSK modulations and α values of 0.35, 0.25, and 0.20.

10.3.3.2 Transponder Bandwidth, Bit Rate, and Error Control

A wide range of transponder bandwidths between 26 and 54 MHz can be used with symbol rates that increase as the bandwidth of the transponder is widened. Half rate FEC was specified in the original DVB-S standard. Some DBS-TV systems changed to 2/3 rate FEC by reconfiguring the software in the satellite TV receivers, increasing bit rates by 8.6 Mbps in a 36 MHz transponder with a requirement for a 2.1 dB increase in CNR to meet the *quasi error free* (QEF) specification. The QEF specification is for one uncorrected error per hour, corresponding to a packet error rate of 10^{-7}, which would be seen on a TV screen as a brief block of the wrong color (*pixelation*) and would likely not be noticed by most TV viewers.

Table 10.2 Summary of DVB-S standard

Modulation	**QPSK**
Pulse shaping	Square root raised cosine with $\alpha = 0.35$
Symbol and bit rates with typical 36 MHz transponder bandwidth	Half rate FEC: $R_s = 27.8$ Msps, $R_b = 25.8$ Mbps 2/3 rate FEC: $R_s = 27.8$ Msps, $R_b = 31.7$ Mbps
Error control coding	Concatenated inner and outer error correction and detection with interleaving
Inner code	Half rate convolutional coding ($k = 7$) with Viterbi decoding
Outer code	Reed-Solomon (204, 188) with 8 byte burst error correction
Interleaving	12×7
Data packet structure	One sync byte followed by 187 data bytes
Energy dispersal	1503 byte pseudo random sequence (eight packets)
Video and audio compression	MPEG 2
Error rate objective	With CNR = 4.5 dB, BER from Viterbi decoder = 2×10^{-4}. Output from R-S decoder: QEF operation of one uncorrected error per hour, equivalent to BER $\approx 10^{-11}$.

Source: ETSI (1997).

Figure 10.9 shows BER versus CNR for DVB-S and DVB-S2 links with a 1.8 dB implementation margin, based on data from ETSI (1997; 2009). In the original DVB-S specification from 1997, the CNR required to meet the QEF specification with QPSK modulation and half rate FEC coding is 5.3 dB with a 1.8 dB implementation margin. The 2007 standard for DVB-S2 transmissions gives a minimum CNR for QPSK with half rate FEC that is about 5 dB lower than the typical performance of DTH satellite television links around year 2000, when a coding gain of 6–7 dB was often quoted, and the CNR to meet the QEF specification was around 7.5–8.5 dB with a 1.8 dB implementation margin. The improvement in BER performance of the DVB-S2 standard over the earlier DVB-S standard was achieved by employing concatenated inner low density parity code (LDPC) coding for error correction and outer Bose-Chaudhuri-Hocquenghem (BCH) coding for burst error correction, with Reed Solomon coding applied to the MPEG packets.

Based on the 1997 DVB-S standard, with a CNR of 2.8 dB, the receiver is operating with a BER at the output of the QPSK demodulator of 3×10^{-2}. That is, one bit in every 33 bits has an error, on average. The inner convolutional FEC decoder has a BER at its output of 4×10^{-4}, on average, corresponding to one bit in 2500 having an error. The Reed-Solomon outer decoder output BER is then 10^{-11}, having corrected all but one of the errors over a period of one hour (ETSI 2009).

When the satellite link suffers a propagation event that reduces the receiver overall CNR to its minimum allowed value, the BER at the output of the Viterbi decoder is 2×10^{-4}, corresponding to an input error rate of 10^{-2}, as seen in Figure 10.9. At a BER of 10^{-2} one bit in every 100 input bits has an error, on average, and the error control scheme works well. With a Viterbi decoder BER of 10^{-4}, one in every 5000 bits is in error, on average. The Reed-Solomon outer code can correct all of these errors resulting in the QEF objective of one uncorrected error per hour. As the receiver CNR falls, the number of errors at the input to the Viterbi decoder increases. Improvements in the Viterbi decoder algorithms used in DBS-TV receivers allowed the BER at the input of the decoder to

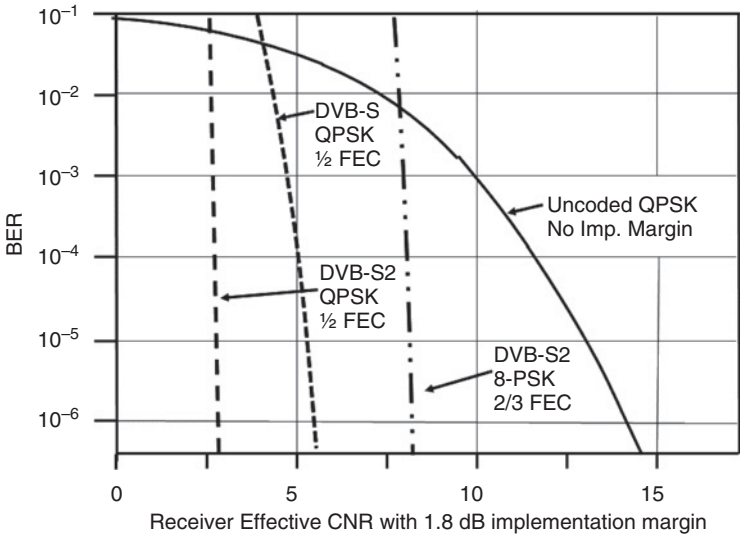
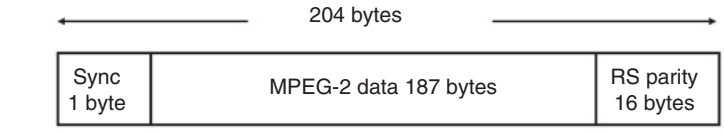


Figure 10.9 Typical bit error rate for DVB-S and DVB-S2 links with 1.8 dB implementation margin. Note that the difference between a very low error rate and a very high error rate is less than a 1 dB change in CNR for the DVB-S links with forward error correction. The TV signal goes from error free to all errors with less than 1 dB change in receiver CNR. The DVB-S2 8-PSK signal with 2/3 rate FEC has a bit rate that is twice that of the QPSK signal with half rate FEC; however, an increase in CNR of 5.6 dB is needed to achieve the same bit error rate.

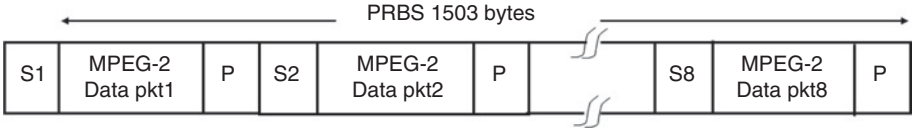
approach 10^{-1}, where the receiver CNR is as low as 1.0 dB. With 1.8 dB implementation margin, half rate coded QPSK signals could meet the QEF requirement with CNR of 2.8 dB. This leads to a much steeper curve for BER when plotted against receiver CNR. Figure 10.9 includes a plot of BER vs CNR for the DVB-S standard, with an implementation margin of 1.8 dB. Many texts on communication theory do not include an implementation margin when presenting BER performance of digital links, leading to unrealizably optimistic prediction of real system performance. The ETSI performance figures are based on simulations carried out at the IF frequencies of the transmitter and receiver, omitting the RF link through the satellite. Non-linear transponder operation increases the BER and a higher implementation margin is needed to achieve QEF operation.

10.3.4 Packets and Frames in the DVB-S Standard

The transmission method for the DVB-S standard is a frame consisting of eight 188 byte MPEG compressed packets of data, called *transport packets* separated by eight synchronization blocks of one byte, as illustrated in Figure 10.10. The 188 byte packet is a standard MPEG-2 packet of data bits, the first byte always being a synchronization byte. A 16 byte parity block is added to the MPEG-2 packet to form a (204, 188) Reed-Solomon (R-S) encoded block in the DVB-S2 standard. MPEG-2 packets contain a mix of video, audio, and data, dominated by the much higher bit rate of the video signal. Data must be sent over the satellite link to update the various tables that are stored by the set-top box, including channel decoding data, program information that can be viewed on the TV screen, and pay-per-view selections.



(a) Reed-Solomon error protected user packet, DVB-S format.



(b) DVB-S 1504 byte frame of eight MPEG packets. Randomization is applied to the packet starting after the first sync byte.

Figure 10.10 Packet and frame structure of DVB-S2 transmissions. (a) The basic packet is an MPEG-2 transport packet consisting of 187 data bytes (8 bits = 1 byte) and one synchronization byte, with 16 parity bytes of (204, 188) Reed-Solomon FEC coding. (b) A frame is made up of eight packets with the first sync byte inverted to mark the start of the frame. 1503 bytes of the frame following Sync1 are randomized with a pseudo random binary sequence (PRBS) for energy dispersal. P, Reed-Solomon 16 byte parity block; Sn, Synchronization byte.

Eight MPEG-2 packets form a frame of 1054 bytes or 8432 bits; all but the first sync byte are multiplied by a pseudo random binary sequence (PRBS) with length 1053 bytes for energy dispersal. This process is called *scrambling* and is necessary because a long string of ones or zeroes in the data stream will cause the transmission of unmodulated carrier that can cause interference to other satellite channels, and may also cause the receiver to lose synchronization. The MPEG-2 transport packet synchronization byte is 47_{HEX}, which is 01000111 binary. The first sync byte in the frame (S1 in Figure 10.10b) is inverted to mark the start of the frame and is not scrambled because it is used by the receiver for carrier recovery and bit synchronization (CRBS). The R-S code is derived from a (255, 239) code with the first 51 bits set to zero and discarded. Frame length is 1053 + 1 bytes = 1054 bytes or 8432 bits.

10.3.5 DVB-S2 Standard

ETSI issued the DVB-S2 specification in 2003 as part of a second generation of digital video systems, building on the success of the original DVB system, with various updates through 2015 (DVB-S2 guide 2015; ETSI 2009). DVB-S2 achieves a 30% increase in bit rate for the same transponder bandwidth and satellite effective isotropically radiated power (EIRP) as DVB-S. When MPEG-4 compression is applied to the TV signals, high definition television (HDTV) signals can be sent over the satellite link using the same transponder bandwidth as for DBV-S, requiring no increase in satellite performance. Table 10.3 lists the major parameters of the DVB-S2 standard. In general, a given DBS-TV link using the DVB-S2 signal format can operate at 2.5 dB lower CNR in the earth station receiver while achieving the same quality of service as the same link using the DVB-S format. DVB-S2 links can operate within 1 dB of the theoretical limit set by the Shannon bound. BER curves for QPSK with rate one half FEC and 8-PSK with 2/3 rate FEC are included in Figure 10.9. When operating at the QEF value of BER = 10^{-11} after

Table 10.3 Summary of DVB-S2 standard for DTH TV without a return link

Modulation	**QPSK, 8-PSK**
Pulse shaping	Square root raised cosine with $\alpha = 0.35$, 0.25 or 0.20
Symbol and bit rates with typical 36 MHz transponder bandwidth, QPSK	Half rate FEC: $R_s = 27.8$ Msps, $R_b = 25.8$ Mbps, $\alpha = 0.35$ 2/3 rate FEC: $R_s = 30.0$ Msps, $R_b = 40.0$ Mbps, $\alpha = 0.2$
Symbol and bit rates with typical 36 MHz transponder bandwidth, 8-PSK	2/3 rate FEC: $R_s = 30.0$ Msps, $R_b = 60.0$ Mbps, $\alpha = 0.2$ 5/6 rate FEC: $R_s = 30.0$ Msps, $R_b = 75.5$ Mbps, $\alpha = 0.2$
Error control coding	Concatenated inner and outer error correction and detection with interleaving of LDPC coding for 8-PSK
Inner code	LDPC coding at rates 1/4, 1/3, 2/5, 1/2, 3/5, 2/3, 3/4, 4/5, 5/6, 8/9, and 9/10
Outer code	BCH coding with 12 bit burst correction capability
Interleaving	$3 \times 21\,600$ for 8-PSK
Data packet structure	187 data bytes followed by 8 bit CRC
Frame structure	Header of 90 bits with (64, 7) Reed-Muller FEC coding. Short frame 16 200 bits, long frame 64 800 bits
Energy dispersal	Two different scrambling sequences, one for header, another for rest of frame.
Video and audio compression	MPEG 2 and MPEG 4
Error rate objective	0.7–1.0 dB from Shannon limit for QEF operation of one uncorrected error per hour, equivalent to BER $\approx 10^{-11}$.

a double layer of error correction has been applied, the CNR at the demodulator input is 1.0 dB and the BER at the output of the QPSK demodulator with half rate FEC is 8×10^{-2}. On average, one bit in every 12 bits is in error. The inner coding of the DVB-S2 standard is LDPC, which ETSI states can operate successfully with QPSK demodulator input CNR as low as 0 dB. The outer code layer is a BCH code with 12 bit burst correction capability. Burst correction is needed because of the high probability of burst errors when bit errors at the demodulator output are occurring at an average rate of one in 12 bits.

10.3.5.1 Modulation and Coding

QPSK and 8-PSK are the preferred modulations for DBS-TV with the DVB-S2 standard. The RF local oscillators used in DBS-TV LNBs do not have sufficient phase stability to support 16-APSK and 32-APSK, where the spacing of symbols in the modulation constellation is much smaller than QPSK and 8-PSK, resulting in a significant increase in BER unless the CNR is increased. Professional point to point links such as electronic news gathering (ENG), which have better quality receivers with more stable local oscillators can make use of the higher order modulations.

Concatenated coding with interleaving of the inner layer is employed when 8-PSK modulation is used. The inner code is LDPC with rates that vary between 1/4 and 9/10. LDPC codes have a small performance advantage over turbo coding at very low CNR and were selected for the DVB-S2 standard to allow operation down to −1.35 dB CNR with 1/4 rate QPSK. This mode lowers the delivered data rate to one half of the more usual

half rate QPSK transmission, and is used only under conditions of severe rain fading to prevent a total loss of signal. In a DBS-TV system, the reduced data rate results in a change to a lower definition TV picture.

Interleaving is employed on the LDPC coded inner layer only for 8-PSK, and is omitted with QPSK modulation. The outer code is BCH, a highly efficient linear block code, with a 12 bit burst error correction capability (Bose and Ray-Chaudhury 1960; Hocquenhem 1959). BCH codes are specified as (n, k, t) where n is the total block length, k is the number of data bits and t is the burst error correction capability of the code. The DVB-S2 standard specifies 21 possible BCH codes; a typical example is (16 200, 16 008, 12) for a short frame with 16 200 data bits and 192 parity bits. Code rate is 0.988 for this example, indicating that the FEC transmission rate is dominated by the choice of LDPC coding rate and affected very little by the BCH encoding (ETSI EN 302 307, Section 5.3 FEC encoding). The ETSI documentation for DVB-S2 contains information on the generation, decoding, and performance of LDPC and BCH codes, topics that are well beyond the scope of this text.

Table 10.4 shows some of the combinations of modulation and coding that are available in the DBS-S2 standard for continuous coding and modulation, as in broadcast satellite television. Spectral efficiency is the number of bits that are transmitted per hertz of RF bandwidth. In Table 10.4, the spectral efficiency value is specified for the data stream for long frames of 64 800 bits, so takes account of overhead bits included in the frame. For QPSK with half rate FEC, the spectral efficiency is slightly below one bit per hertz because of the overhead bits needed to form frames and the outer layer of BCH coding. For broadcast DBS-TV, 8-PSK with 2/3 rate FEC is envisaged as the ModCon of choice when sufficient satellite EIRP and receiving antenna gain are available. The receiver CNR required to achieve the QEF specification of one error per hour is lower with DVB-S2 than with DVB-S; QPSK transmission with half rate FEC is possible with CNR at 2.5 dB with DVB-S2 parameters, rather than 5.0 dB for DVB-S, after allowing for

Table 10.4 Selected modulation and coding rates available with the DVB-S2 standard for DBS-TV, based on simulation by ETSI (2009). A 1.8 dB implementation margin is included in the minimum CNR to meet the QEF objective

Modulation	Code rate	Spectral efficiency (bits/Hz)	CNR with 1.8 dB imp margin (dB) for QEF
QPSK	1/4	0.49	−0.55
QPSK	1/3	0.66	0.56
QPSK	1/2	0.99	2.8
QPSK	2/3	1.32	5.8
QPSK	3/4	1.49	6.4
QPSK	5/6	1.65	7.0
QPSK	9/10	1.79	8.3
8-PSK	2/3	1.99	8.4
8-PSK	3/4	2.23	9.7
8-PSK	5/6	2.48	11.2
8-PSK	9/10	2.68	12.8

a 1.0 dB implementation margin. The improvement is largely the result of using LDPC coding instead of convolutional encoding with Viterbi decoding. The specified QPSK ModCon provides a spectral density very close to 1 bit/Hz of transponder bandwidth. The half rate FEC coded 8-PSK ModCon achieves 2 bits/Hz, but requires 7.6 dB CNR at the input to the DBS-TV digital receiver. A spectral density of 2.5 bits/Hz is possible with 8-PSK modulation and 5/6 rate FEC coding, requiring 10.4 dB CNR at the receiver to meet the QEF specification.

10.3.6 Packets and Frames in the DVB-S2 Standard

The DVB-S2 standard utilizes the same user packet structure as DVB-S, with multiple 188 byte data packets forming a baseband data frame. Data frames are 16 200 bits (short frame) or 64 000 bits (long frame). The transmitted frame starts with a 90 bit header. The format of the header and frame is shown in Figure 10.11 for the specific case of a short frame with fixed coding and modulation using MPEG-2 compression, although the DVB-S2 format allows for many different transmission modes. Information in the header tells the receiver what to expect in the following frame (ETSI 2009). Padding bits, which are discarded by the receiver, are added to the uncoded baseband frame to ensure that it contains 7032 bits. The padding bits are necessary because of the many different ways in which a frame can be constructed with different LDPC coding rates. In the example in Figure 10.11 for a short frame, there are four data frames (k = 4) and

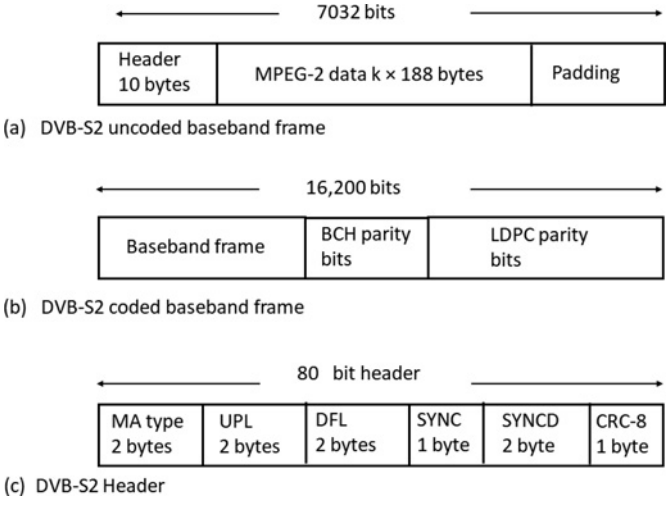


Figure 10.11 Frame structure and header for short frame DVB-S2 transmissions using MPEG-2 compression and fixed coding and modulation. (a) The DVB-S2 uncoded baseband frame consists of a header, *k* blocks of MPEG data, and padding bits to complete a frame of 7032 bits. (b) The DVB-S2 coded frame adds BCH parity bits and LPDC parity bits to form a frame of 16 200 bits. (c) DVB-S2 header structure used for short and long frames. BCH, Bose-Chaudhury-Hocquenhem FEC block code; LPDC, Low density parity code FEC code; MA, Defines input stream type – single or multiple – constant or adaptive coding and modulation, and SRRC roll off factor α; UPL, User packet length in bits, typically $8k \times 188$; DFL, Data field length in bits, in the range 0–58 112; SYNC, User packet sync byte; SYNCD, Used to determine where in the frame the data bits are located; CRC-8, 8-bit cyclic redundancy error check applied to last 9 bytes of header.

936 padding bits. The entire frame, after the header, is scrambled with an appropriate pseudo random sequence. Short frames are used mainly when CNR is very low, with longer frames that have higher efficiency used for DBS-TV transmissions under normal conditions.

The header is transmitted with binary phase shift keying (BPSK) modulation and the first 26 symbols are a start of frame (SOF) sequence $8D2E82_{HEX}$ that is used in the receiver to synchronize the PSK demodulator and check for phase ambiguity. The next 64 symbols are seven data bits encoded with a (64, 7) Reed-Muller code capable of correcting up to 32 errors. The header is scrambled with an 80 bit sequence separate from the scrambling of the remainder of the frame. Five of the seven data bits in the header are used to identify the modulation and coding (ModCon) used in the rest of the frame, and two bits specify the length of the frame, either short (16 200 bits) or long (64 800 bits) (ETSI 2009; 2015). The very heavy protection applied to the header ensures that the header can be decoded correctly at the receiver with $E_b/N_o = -2.5$ dB. An uncorrected error in the header data will result in the entire frame being lost, whereas an uncorrected error in a user packet results in the loss of a block of pixels in a single video frame.

Table 10.4 contains selected coding rates for QPSK and 8-PSK modulations used with constant coding and modulation (CCM) transmissions of direct to home broadcast TV. The coding and modulation can be changed from frame to frame, but this is unlikely in a broadcast system. Reductions in coding rate and a change from 8-PSK to QPSK could be used to combat uplink rain attenuation, which affects all receivers within the satellite footprint.

Figure 10.12 illustrates the spectral efficiency of some of the multiple modulation and coding combinations that are available in the DVB-S2 standard when operating

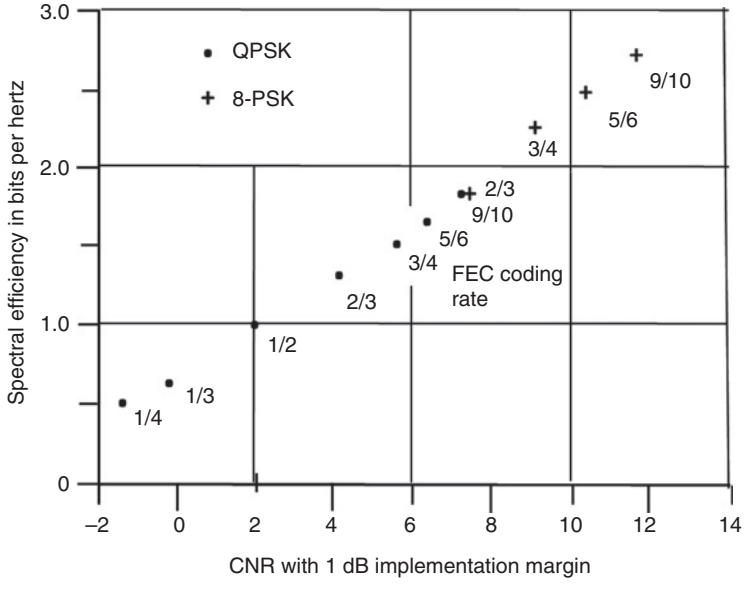


Figure 10.12 Performance for DVB-S2 QPSK and 8-PSK modulation-coding combinations with 1.0 dB implementation margin, SRRC filtering with $\alpha = 0.2$ and a packet error rate of 10^{-7}. The fractions are the FEC coding rate. CNR values are for QEF threshold. Source: Based on data from ETSI (2003).

at threshold CNR. Based on simulations with an ideal receiver, ETSI predicts operation within 1 dB of the Shannon bound for QPSK modulation using the DVB-S2 format (ETSI 2003). This suggests that there is little room for performance improvement by more complex FEC techniques, with only reduction in implementation margin as a feasible strategy to lower the required CNR in the receiver while still meeting the QEF requirement of one error per hour.

10.3.6.1 Adaptive Coding and Modulation

The second generation DVB standard differs from the first generation in one very significant way. Digital video broadcast – return channel satellite (DVB-RCS) specifies a transmission system with a real time return link from the user's receiver to the gateway transmitting station. This allows ACM to be implemented rather than the CCM of the DVB-S standard. DVB-RCS is also a standard that can be employed for internet access by satellite, the next major growth area for satellite communications. Internet access by satellite is the topic of Chapter 11; in this chapter we will concentrate on DBS-TV applications of the DVB-S2 standard using CCM.

Implementation of ACM for internet access requires time division multiple access (TDMA) as the multiple access technique so that a virtual circuit can be established between the gateway and each user terminal. In DBS-TV systems operating in broadcast mode rather than with virtual circuits, ACM can be applied to satellites that have multiple spot beams, based on the reported CNR for receiving stations within the beam. Spot beams are typically 500 km wide, so changes to modulation and coding affect a large number of users, making ACM less effective than with TDMA. However, when the modulation or coding rate are changed to a lower order to increase the link margin, the bit rate is reduced. That causes a loss of definition with video transmissions, which is an undesirable side effect. ACM can be used as an alternative to uplink power control on uplinks, since rain attenuation on an uplink affects all receiving stations. An eventual transition is envisaged to DBS-RCS systems with return links that can deliver both TV programming on demand and internet access. ETSI encourages manufacturers of DVB-S2 integrated circuits and receivers to build this capability into their equipment.

10.3.7 Professional Applications and DVB-S2x

The DVB-S2 standard envisages a different class of professional applications from direct to home satellite TV broadcasting, where the cost of the receiving equipment must be as low as possible since millions of units are required. Under the generic class of professional applications, digital satellite news gathering (DSNG) is an important uses of GEO satellites. The sight of a group of trucks equipped with large antennas at a sporting or a news making event is familiar to all television viewers. The *satellite trucks* are equipped with uplink transmitters that can use the DVB-S2 standard with 16-APSK or 32-APSK modulations to transmit through a GEO transponder back to the home TV station. The larger antennas at each end of the link can provide higher CNR in the receiver at the TV station, and more stable oscillators allow the use of higher order PSK modulations. Details of the DSNG applications of the DVB-S2 standard can be found in ETSI (2009).

DVB-S2x is a proposed improvement on the basic S2 standard that uses SRRC filters with lower alpha values, ACM, and extended ranges of CNR to support higher bit rates

and operation under conditions of high rain attenuation, such as are encountered in Ka-band links. Higher bit rates are possible under clear sky conditions and the addition of some BPSK modulation-coding combinations allows operation at very low CNR (ETSI 2015).

Most satellite trucks are equipped to use Ku-band for transmissions to the home TV broadcast station. A single TV channel requires an RF bandwidth between 2 and 4 MHz, depending on the video definition and MPEG compression applied to the video signal. A portion of the bandwidth of a transponder is leased for the duration of the event, with extra time added for setup. Large news organizations like CNN and ESPN that have a national audience have long term leases on transponders, and satellite trucks standing by for immediate deployment. The trucks are equipped with receivers so that the transmission from the satellite can be monitored, and larger, more sophisticated trucks may be equipped with production facilities. Often, a satellite truck is paired with a production truck so that a complete TV show can be created from a remote location. The trucks typically carry a 12 kW diesel generator to avoid dependence on a local electrical supply, and are useful for emergency communications at the site of a disaster.

Antennas on satellite trucks have diameters limited by the width of the vehicle, from around 1.6 m for vans and 2.4 m for trucks. The antennas must fold down for transport, but be fully steerable in elevation and azimuth. Most trucks are not equipped to transmit while moving; the truck must be parked at a location with a clear view toward the satellite and jacks are lowered to stabilize the antenna pointing. Offset front fed reflector antenna configurations are common, with some Gregorian and Cassegrain antennas employed on larger vehicles. Transmit power is in the hundreds of watts, with redundant transmitters in larger installations. With the advent of more powerful Ka-band satellites such as ViaSat 1, Ka-band systems are available as an alternative to Ku-band. Costs for a satellite truck quoted on internet sites range from US$85 000 for an SUV with single channel capability to over US$1M for a 40 ft truck with a full production suite. Transponder leases in Ku-band are quoted as US$300 per hour for occasional access with a single TV channel to tens of thousands of dollars for monthly lease of a whole transponder (2018 prices).

One problem that appears to be difficult to overcome is the delay introduced into two-way video interviews between a TV studio and a satellite truck. The two-way path delay for a GEO satellite link is around 500 ms, but added to this delay is the processing time to decode the response from the remote location when MPEG-4 coding is employed. With HD TV links, MPEG-4 coding is standard and the decoding delay is around 700 ms, so the two-way link incurs a minimum delay of 1.2 seconds between the instant an interviewer in the studio stops speaking and the start of the response from the remote location arrives. The long delay makes it appear as though the person at the remote location is very slow in reacting to a question from the studio, but it is the 1.2 seconds delay in the two-way transmission process that is the cause. Improvements in the encoders and decoders for MPEG video compression should eventually reduce the delay.

Figure 10.13a shows the antennas at the studio location of TV station WDBJ in Roanoke, Virginia. Figure 10.13b shows a satellite truck operated by the TV station. WDBJ is the premier television broadcast station serving South West Virginia, including Blacksburg, the location of Virginia Tech. WDBJ broadcasts digital TV signals in the ultra high frequency (UHF) band from an antenna located on Poor Mountain, which has an elevation of 1197 m. The signals are also rebroadcast by both Directv and Dish

Figure 10.13a Antennas at the studios of the WDBJ television station in Roanoke, Virginia. The reflector antennas are all used for satellite communications. The center antenna is a Simulsat design capable of receiving signals from multiple satellites. The large antenna at the left of the photograph has a Gregorian configuration with a shaped subreflector. The two antennas on the mast link to WDBJ's broadcast antenna on Poor Mountain. For a color version of this figure please see color plate section.

Network DBS-TV satellites. South West Virginia is in the Appalachian Mountains so there are many locations where the terrestrial broadcast signals are blocked by high terrain. The DBS-TV satellite transmissions provide reliable reception over a wider area than the terrestrial broadcasts.

More details of ENG and satellite trucks can be found by searching the internet.

Figure 10.13b Satellite truck used by the WDBJ television station in Roanoke, Virginia, for outside broadcasts. The main satellite communication antenna is shown in its operating position. Source: Photographs courtesy of WDBJ-TV, © WDBJ-TV 2018. For a color version of this figure please see color plate section.

10.4 DBS-TV System Design

A DBS-TV system must create a received signal power at the small receiving antenna that provides an adequate CNR margin in clear sky conditions. Heavy rain will cause attenuation that exceeds the link margin, so occasional outages will be experienced, especially during the summer months when thunderstorms and heavy rain are more frequent. The CNR margins used in DBS-TV systems are small, to avoid the need for a large receiving antenna. A margin of at least 1 dB is always needed because of scintillations in the atmosphere.

The selection of a CNR margin is a design trade off between the outage level that customers can be expected to tolerate, the maximum allowable diameter of the receiving dish antenna, and the power output from the satellite transponders. Typically Ku-band DBS-TV receiving antennas have dimensions in the 0.45–0.9 m range. With a conus beam and 80 – 200 watts satellite transponders rain attenuation margins of 3–8 dB and outage times totaling 5–40 hours per year are obtained, depending on the receiver's location. However, since most customers don't watch TV for 24 hours per day, they will not be aware of all the outages. Unfortunately, thunderstorms tend to occur more often in the late afternoon and evening, resulting in more outages during prime viewing time. The transmit beams of satellites carrying DBS-TV signals are shaped to deliver more power to those areas that suffer the highest occurrence of heavy rain, such as the states in the southeastern part of the United States, and spot beams directed to those areas can have higher gain and more transmit power. This creates a larger link margin in those areas and helps to keep outages to an acceptable level. Figure 10.14 shows an example

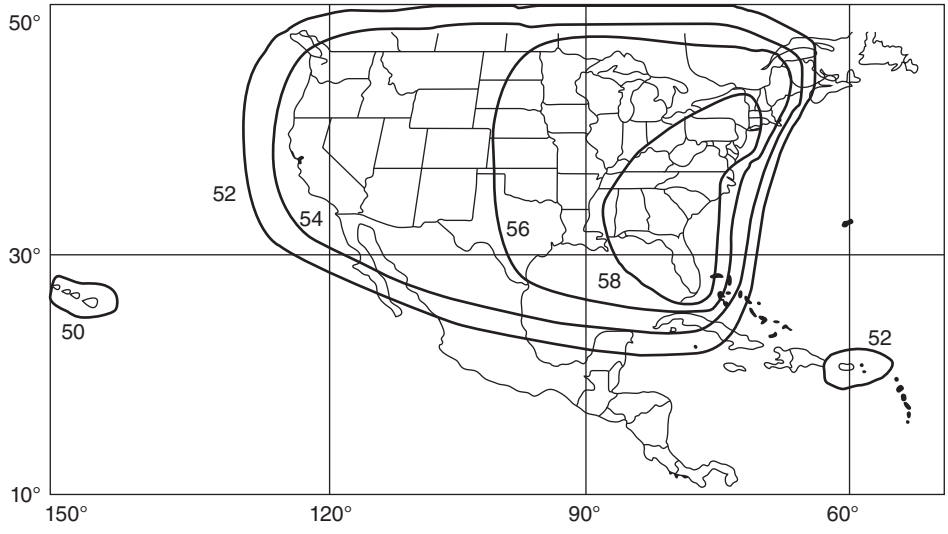


Figure 10.14 Typical conus beam of a DBS-TV satellite serving the United States in Ku-band. Contours are EIRP in dBW, with highest EIRP directed toward the states in the south east of the United States where heavy rain occurs most frequently. Maximum EIRP is typically 60 dBW directed toward Florida. Similar conus beams are transmitted in both LHCP and RHCP with the beams overlaid, occupying approximately half of the available Ku-band frequencies. There are subsidiary beams centered on the Hawaiian Islands and Puerto Rico. Not shown is subsidiary coverage of Alaska.

of satellite antenna EIRP contours over the United States for a typical Conus beam of a DBS-TV satellite. Maximum EIRP of the conus beam is approximately 60 dBW, directed toward Florida. Multiple spot beams are used to provide local TV programming for the service areas of selected cities and conurbations, while the Conus beam provides service throughout the contiguous 48 states. The high gain of the spot beams allows the local program services to be transmitted at a lower transponder output power level, and also permits frequency reuse by spatial beam separation. A dual Gregorian reflector system of the transmitting antenna on the satellite is fed by a complex feed structure that produces the Conus beam contours shown in Figure 10.14. (See Figure B6b in Appendix B for an illustration of this type of satellite antenna.) The contours of the US Conus beam are shaped to cover the 48 lower states with highest EIRP directed to the southern and eastern states that experience the most frequent occurrence of heavy rain. Subsidiary beams are directed toward Hawaii, Puerto Rico, and Alaska carrying regular TV and local channels, because these regions are not served by the Conus beams. Most DBS-TV satellites use separate antennas to generate the RHCP and LHCP conus beams, and a second pair of antennas to generate the spot beams. Four reflector antennas can be seen on the satellite illustrated in Figure 10.4.

Florida, Alabama, and Louisiana, for example, are in rain zones M and N of the United States that have a rainfall rate of 50 mm/hour for five times the number of hours per year that this rain rate occurs in Washington, D.C., and much of the eastern portion of the United States. At Ku-band, a rain rate of 50 mm/hr will cause about 6 dB attenuation on a typical DBS-TV slant path, sufficient to ensure an outage of the DBS-TV signal when the increase in system noise temperature is taken into account. This rainfall rate is exceeded for about five hours per year in Florida and one hour per year in Washington, D.C. The central and western parts of the United States have rainfall rates of 50 mm hour for much less than one hour per year, and therefore do not need such large link margins. See Chapter 7 for detailed maps of rainfall rates for the United States and the world, and for the techniques to convert rainfall statistics into attenuation data that can be used to calculate outage times. Note that DBS-TV coverage for all of the 48 lower states of the United States shown in Figure 10.14 lies inside the −6 dB contour (at 54 dBW) of the satellite antenna beam, giving beam loss values between −4 and −6 dB, so the usual edge of beam loss of 3 dB cannot be applied here. The regions within the – 4 dB to −6 dB contours are those that do not have frequent heavy rainfall.

Some manufacturers of DBS-TV receiving systems offer larger dishes for customers living in high rainfall zones. Increasing the antenna diameter from 0.45 to 0.60 m (24 in.), for example, increases its gain by 2.5 dB. This increase in antenna gain adds directly to the rain fade margin of the receiver, and lowers the outage time in heavy rain. Larger dishes are also needed in Alaska and Hawaii, which are served by subsidiary satellite beams with lower gain.

Figure 11.15 shows an example of some of the spot beams of a typical DBS-TV satellite serving the 48 lower states of the United States. The reflector antennas on a typical DBS-TV GEO satellite have diameters of 2.25 m. At 12.5 GHz, an antenna with a diameter of 2.25 m and an aperture efficiency of 65% has a spot beam gain of 47.5 dB. The 3 dB beamwidth of this spot beam is approximately 0.8°. At a distance of 37 500 km from the satellite, the 3 dB beamwidth of the narrowest spot beam footprint is about 525 km (280 miles). Spot beams A11 and B11 over Florida in Figure 10.15 are two of the narrowest beams with the highest gain to combat attenuation in heavy rain. A transponder output power of 15.5 dBW (36 W) combined with an antenna gain of 47.5 dB produces

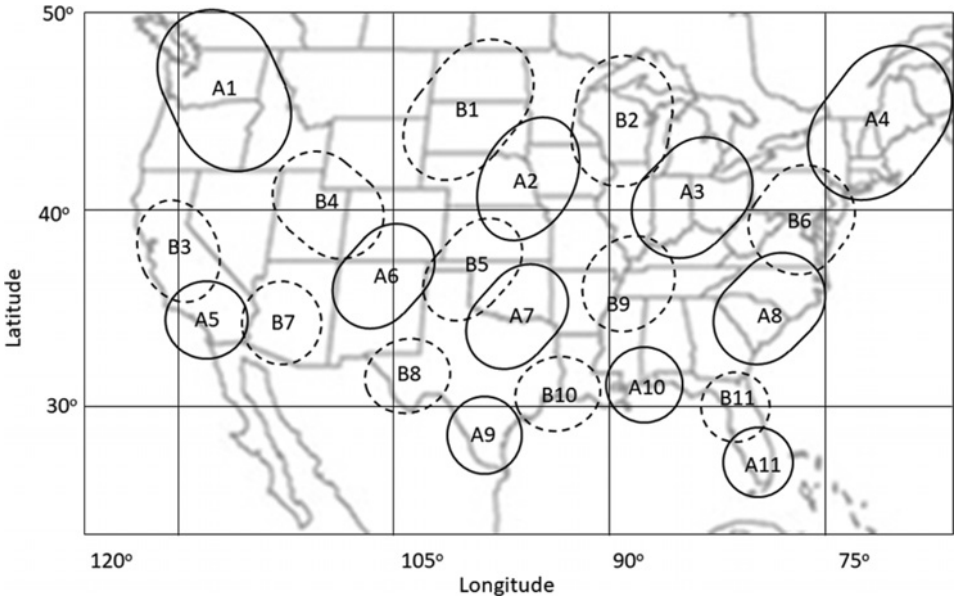


Figure 10.15 Example of spot beams generated by a Ku-band DBS-TV satellite serving the continental United States. Spot beams are transmitted in both LHCP and RHCP with EIRP values between 52 and 63 dBW. DBS-TV satellites may generate as many as 80 spot beams using half of the available Ku-band frequencies, divided into two to five sub-bands. Shown in this diagram are 22 spot beams occupying two sub-bands, A and B, in one polarization. A second set of spot beam in the opposite hand of circular polarization overlays those shown here. Spot beams are used to transmit local television programming, and also some HD material.

an on axis EIRP of 63 dBW. Several transponders operating in different frequency bands may be connected to a given spot beam to provide a large number of TV channels to a specific service area. Other beams are much wider to cover a larger geographical area, and consequently have a lower gain. For example, beams A1 and A4 in Figure 10.15 cover four times the area of the A11 beam, and will therefore have a gain that is 6 dB lower than the A1 beam. Lower gain beams are satisfactory in the northern and western regions of the United States where heavy rain occurs much less frequently than in Florida.

Spot beams typically occupy the frequency bands 12.5–12.7 GHz, carrying primarily local TV stations for the specific geographic area covered by each beam. Several spot beams may overlap to provide additional channels to the densely populated area of the US east coast between Washington, DC and Boston, MA. Other DBS-TV satellites may use the 11.7–12.2 GHz frequency band to deliver additional programming to North America. The 22 spot beams shown in Figure 10.15 in two frequency bands A and B are geographically separated to minimize interference between beams in the same frequency band. A separation of 600 km between the centers of any two beams with the same polarization ensures that mutual interference is below −20 dB. The spot beam contours show in Figure 11.15 represent the −3 dB contours relative to the gain at the center of the beam, not the full extent of the beam coverage. It is possible for receiving locations beyond the −3 dB contour of any beam to have satisfactory reception 4 or 5 dB below the peak EIRP of the spot beam where a large rain attenuation margin is not required. Footprints for conus and spot beams for many GEO satellites can be found at (Satellite

footprints 2018) and (Satellite Downlink Coverage Patterns 2018) along with details of transponder loadings.

10.5 DBS-TV Link Budget for DVB-S and DVB-S2 Receivers

In this discussion, rain attenuation statistics at Ku-band will be used that are representative of many locations in the central and eastern parts of the United States, where typical path attenuation in rain exceeds 3 dB for 0.2% (15 hours) and 6 dB for 0.01% (52 minutes) of an average year. The distribution of the fades is random, with some long fades in the heaviest of thunderstorms and numerous shorter fades in brief periods of heavy rain. DTH-TV services with receiving systems using 0.45 m × 0.6 m diameter antennas aim to provide an availability exceeding 99.7% of an average year, which is an outage time of 0.3% of the year, a total of about 25 hours per year. For much of the United States, this corresponds to a rain attenuation in the slant path of 3 dB and requires a link margin of 5.7 dB when allowance is made for the increase in antenna noise temperature that accompanies 3 dB of rain attenuation. It should be noted that this is a statistical result; there could be some years in which outages occur in some places for longer than 0.3% of the year, causing customers to complain. However, those same customers would be unlikely to notice if the outages were much less than 0.3% of year. Such is the problem of statistics. The analysis that follows is for a receiver located on the −3 dB contour of the Conus beam of a Ku-band GEO satellite transmitting multiple video and audio channels.

10.5.1 Link Budget for DVB-S Terminals with Conus Beam

A representative link budget for a first generation GEO DBS-TV system serving the United States with Conus beams is shown in Table 10.5. The path length of 37 500 km is the typical path length for a receiver in the United States and a satellite at longitude 101°W. The satellite has 32 transponders with 30 MHz bandwidth, 16 transmitting LHCP and 16 transmitting RHCP across the 500 MHz band 12.2–12.7 GHz. Some of the transponders are allocated to the Conus beams and some to spot beams. Saturated output power of the Conus beam transponders in this example is 100 W, but with 1.5 dB backoff the transmitted power level is 71 W. The transponders carry signals in the DVB-S format with an occupied bandwidth of 30 MHz. Standard DSB-S SRRC filtering has $\alpha = 0.35$, so the 3 dB bandwidth of the signal is 22 MHz and the symbol rate is 22 Msps. Using QPSK modulation with half rate FEC, the bit rate is 22 Mbps. The threshold CNR value is set at 5.1 dB, corresponding to a system based on the DVB-S format using QPSK with a realistic implementation margin of 1.6 dB, half rate FEC coding, and a maximum BER of 2×10^{-2} at the output of the QPSK demodulator. An additional allowance of 0.8 dB is made for clear sky atmospheric loss and antenna misalignment, and a further 0.5 dB for miscellaneous losses.

The link budget in Table 10.5 shows that a link margin of 8.0 dB is achieved for a receiver located on the −4 dB contour of the satellite's Conus antenna beam. Earth stations close to the −6 dB contour of the Conus footprint have a link margin of 6.0 dB. A receiver located in the SE states of the United States, within the −2 dB contour of the satellite beam, has a link margin of 10.0 dB. The receiving antenna of the user's DBS-TV system is a high efficiency design with an offset parabolic reflector and a circularly polarized feed. The offset design ensures that the feed system does not block the aperture of

Table 10.5 Link budget for Ku-band DBS-TV receiver with DVB-S signal format

Transponder saturated output power	100 W	20.0 dBW
Output backoff		−1.5 dB
Conus beam on-axis gain		36.0 dB
EIRP on beam axis		54.5 dBW
Path loss at 12.2 GHz	37 500 km path	−205.7 dB
Receiving antenna gain, on axis		34.8 dB
Beam contour loss		−4.0 dB
Atmospheric clear sky loss		−0.4 dB
Receiving antenna mispointing		−0.4 dB
Miscellaneous losses		−0.5 dB
Received power	C	−121.7 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz
System noise temperature, clear sky	110 K	20.4 dBK
Receiver noise bandwidth	22 MHz	73.4 dBHz
Noise power	N	−134.8 dBW
CNR in clear sky		13.1 dB
Link margin over 5.1 dB threshold		8.0 dB
Link availability throughout US		Better than 99.7%

the antenna, which increases its efficiency. The gain of this 0.5 m diameter antenna with an aperture efficiency of 70% is 34.8 dB at 12.5 GHz. Received power under clear sky conditions is −121.7 dBW (−91.7 dBm).

The noise power budget of the link in Table 10.5 is based on a receiver noise bandwidth of 22.0 MHz, an antenna noise temperature of 35 K in clear sky conditions, and a 12 GHz LNA with a noise temperature of 75 K. The result is a noise power of −134.8 dBW referred to the input of the LNA and a clear sky CNR ratio of 13.1 dB. This is 8.0 dB above the DVB-S standard threshold of 5.1 dB assuming a 1.6 dB implementation margin. As noted in Chapter 6, the link margin must be divided between rain attenuation and increase in system noise temperature. Rain attenuation of 4.4 dB increases the sky noise temperature to 181 K and increases noise power in the receiver by 3.6 dB. The result is a reduction in receiver CNR of 8.0 dB, bringing the value close to threshold for a practical DVB-S link using QPSK and half rate FEC.

The link budget in Table 10.5 does not account for interference between the LHCP and RHCP Conus beams, or for uplink CNR in the satellite transponder. The CNR in the transponder can typically be maintained above 30 dB in clear sky conditions, which causes less than 0.1 dB reduction in overall CNR in the earth station receiver. Where beams with opposite hands of circular polarization are overlaid, interference from one beam to another is generally below −25 dB.

10.5.2 Link Budget for DVB-S Terminals Within a Spot Beam

In a typical DBS-TV Ku-band satellite with multiple spot beams, the spot beams have higher gain than the Conus beams, a minimum 3 dB width on the earth's surface of

500 km, and are located at the upper end of the Ku-band frequencies allocated to DBS-TV. The higher gain of the spot beams can be traded for a reduction in transponder output power, or the modulation can be changed to 8-PSK with FEC rate 2/3, increasing spectral efficiency to 2.0 bits per hertz. The alpha value of the SRRC filter can be reduced to 0.25, which increases the bit rate in a 30 MHz transponder to 47 Mbps. With 20 spot beam transponders and 40 spot beams, the satellite can typically transmit 11 live TV programs in each of the 20 spot beam transponders, or a larger number of channels carrying a combination of live and prerecorded programming. This gives a maximum of 220 live channels that can be delivered via the spot beams, which are used primarily to carry local TV stations from within the geographical area served by the spot beam.

10.5.3 Channel Loading

Prerecorded material, which comprises the majority of programming on satellite TV channels, is heavily processed to reduce its bit rate to 1.6 Mbps. Older television programming produced when the NTSC standard was in use had lower definition than programming produced to meet the ATSC standard and can be transmitted in standard definition rather than high definition. Standard definition programming, known as 720p, has a definition of 1280×720 pixels and can be transmitted at a bit rate of 1.6 Mbps using MPEG-2 compression. High definition programming, known as 1080p, has 1920×100 pixels and is transmitted at rates between 3 and 4 Mbps using MPEG-4 compression.

When prerecorded material such as a movie is digitized and processed through MPEG-2 or MPEG-4 compression to achieve a significant reduction in bit rate, digital artifacts appear in the picture, especially when there is a lot of motion in the scene. A digital artifact appears as a freezing of the entire picture for a fraction of a second, caused by overloading of the MPEG processing, or as a block or pixel of the wrong color. The individual artifacts can be removed one by one by a digital artist who works on the recorded material to paint out the effects, and rapid motion in the original scene can be spread over several frames. The final result is a movie or show that can be recorded in digital compressed form for replay over the satellite. Live program material with a lot of motion in the picture can cause the bit rate of an MPEG-2 coded signal to increase above an average value of 2.0 Mbps. Mixing prerecorded and live material in a single transponder helps even out the bursty nature of live material. For example, four live transmissions requiring 8.0 Mbps can be supplemented with eight prerecorded channels requiring 12.8 Mbps to give a bit stream at 20.8 Mbps, delivering 12 different TV programs through one transponder. If needed, padding packets can be added to maintain the bit rate at 22 Mbps; these packets are ignored by the earth station receivers. With 20 transponders allocated to the Conus beams, a total of 180 video and audio channels can be transmitted to all households in the United States. The specific channels carried by the Conus and spot beams of many DBS-TV satellites can be found on the internet (http://www.lyngsat-maps.com/2018).

When excessive rain attenuation occurs in the downlink, the BER of the recovered bit stream in the receiver increases, the impact of the word errors becomes more severe, and larger blocks of the TV picture are corrupted. The receiving system is able to recognize the high error rate and will blank the screen until an acceptable error rate is restored. Thus a rain fade on a DBS-TV link that goes below the receiver threshold is characterized by the initial appearance of small squares of incorrect color on the TV

screen (pixelation), followed by larger block errors and then a blank screen, or a message saying "signal lost." When the rain intensity eases as the storm moves through the slant path, the signal will return above threshold and the picture will reappear on the TV screen.

The user of a DBS-TV system is usually aware of a thunderstorm or very heavy rain in the locality when the signal goes below threshold and the TV screen goes blank. This seems to make loss of the TV picture more acceptable to users, and most DBS-TV customers appear to be satisfied with a nominal availability of 99.7%. The actual availability is undoubtedly higher than 99.7% for most of the customers in the United States, and few complaints seem to arise from the loss of signal in heavy rain.

10.6 Second Generation DBS-TV Satellite Systems Using DVB-S2 Signal Format

The second generation of DBS-TV satellites took advantage of improvements made in satellite antenna technology, baseband signal processing, and modulation techniques, and in some cases an expansion into Ka-band. Multiple narrow spot beams provide up to 10-fold frequency reuse across the continental United States, and improvements in MPEG-2 and MPEG-4 compression reduced the bit rate required for SD and HD video transmission. DVB-S2 baseband processing and 8-PSK modulation increased the spectral density of the transmitted signals from 1.0 bit/Hz with QPSK modulation with half rate FEC coding to 2.0 bits/Hz with 8-PSK modulation with 2/3 rate FEC coding. The adoption of LDPC error correction coding allowed receivers to operate at lower CNR, within 1 dB of the Shannon bound. The QEF threshold for QPSK modulation with half rate FEC coding is 2.0 dB with 1.0 dB implementation margin, an improvement of 2.5 dB over the DBV-S standard. All of the changes increased the capacity of typical second generation DBS-TV satellites to about 10 Gbps.

In 2011, ViaSat launched the VIASAT I satellite, a third generation design with a capacity exceeding 100 Gbps, which uses Ka-band frequencies and has 72 spot beams. Several similar Ka-band satellites followed with even higher capacity, and by 2018 Eutelsat was considering launching a satellite capable of delivering 500 Gbps. These satellites are intended for internet access, rather than video distribution, but can be adapted to either purpose. Chapter 11 gives more details of these high capacity satellites.

An example of a second generation DBS-TV satellite is Echostar 14, launched in 2010 (Echostar 14). Echostar 14 is a large GEO satellite that serves Dish Network customers with Ku-band television programming from 119°W. The satellite has 103 transponders transmitting across the 12.2–12.5 GHz band, with two Conus beams and 51 spot beams. EchoStar 14 was built by Space Systems/Loral, based on the LS-1300 satellite bus, with a launch mass of 6384 kg (14 074 lb). Expected operational lifespan is around 15 years. The launch of EchoStar 14 was conducted by International Launch Services, using a Proton-M carrier rocket with a Briz-M upper stage from the Baikonur Cosmodrome in Kazakhstan, on 20 March 2010.

Echostar 14 transponders are configured to carry DVB-S2 format signals in an occupied bandwidth of 24 MHz with QPSK modulation and 25.4 MHz with 8-PSK modulation. There are 32 such channels available in the Conus beams at the higher RF frequencies, 16 in each hand of circular polarization. The lower part of the RF

band is occupied by the 51 spot beams. There are 103 TWTAs on Echostar 14 with the following saturated output powers: 84 at 150 W, 14 at 70 W, and 5 at 35 W. Two of the 150 W transponders can be phase combined to give a saturated output power of 300 W, producing peak EIRP of approximately 59 dBW per channel in the Conus beam. Combining three 150 W transponders gives a maximum EIRP of 60.7 dBW in the Conus beam, but interference considerations restrict operation at this high EIRP to 21 of the 32 Conus beam channels (Echostar 14 FCC 2009).

10.7 Master Control Station and Uplink

Direct broadcast television satellites are relay devices that provide a very large coverage area serving millions of customers. The many signals that are broadcast by the satellites are collected at several uplink stations and transmitted to the satellites by a group of large antennas with fade margins sufficient to overcome any expected rain fade. It is essential to have more than one uplink station to provide redundancy in case a station is unavailable because of failure or maintenance requirements, and to have interconnection by optical fiber between the stations. If extreme weather affects any one station, another station can immediately take over the transmissions. Echostar, for example, operates six uplink stations across the United States.

The video and audio signals that are uplinked to the DBS-TV satellites are available in prerecorded form, or are collected from other satellites or from fiber optic lines. This is a large operation, which requires substantial resources and a sizable labor force. The location of DBS-TV uplink stations is generally in US rain zone B2, providing a low probability of heavy rain. The statistics of region B2 show that a rain rate of 50 mm/hr is exceeded for only five minutes in a typical year. The major European uplink station for DBS-TV, operated by SES, is located in Luxemburg, which is also in the European rain zone B2.

The uplink station must transmit hundreds of signals to the DBS-TV satellites 24 hours a day, 365 days a year. Many of the signals are prerecorded, either from satellite feeds that are used to distribute new video and audio program material, or from archived material. The uplink stations have high capacity digital storage units, all under computer control, which supply the video and audio signals for each channel. Analog signals must be digitized and compressed before being multiplexed with other signals into the bit streams that are sent to each transponder. More details of the uplink centers operated by DBS-TV companies can be found from their web sites (DirecTV 2018; Echostar 2018; SES 2018).

A simplified block diagram of the transmitting equipment at an uplink station is shown in Figure 10.16. One uplink antenna can typically transmit up to 16 RF channels in LHCP and 16 RF channels in RHCP to one DBS-TV satellite. The input to the baseband processing section in Figure 10.16 is a stream of MPEG encoded video and audio data packets, with added data packets. The upper blocks in the diagram generate frames that are sent to the PSK modulator, which may generate QPSK or 8-PSK modulation. The content of the frames depends on the signal format, which may omit the header, insert a DVB-S header that does not have separate FEC encoding, or insert a DVB-S2 header that is heavily encoded with a (64, 7) Reed-Muller code capable of correcting up to 32 errors. The buffer assembles 8 or 32 MPEG packets to form short or long frames and applies a double layer of FEC coding. The scrambler multiplies the encoded

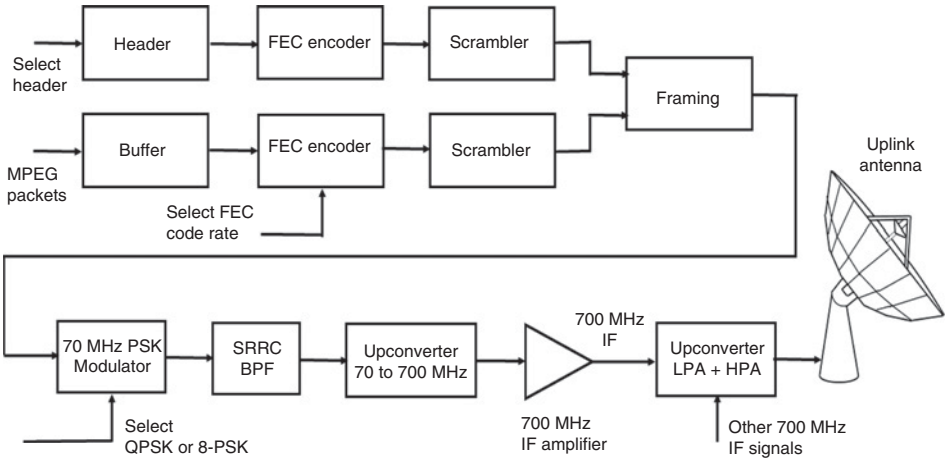


Figure 10.16 Simplified block diagram of a DTH TV transmitting station. The upper blocks in the diagram are at baseband. The frames that are sent to the PSK modulator depend on the signal format, which may omit the header, insert a DVB-S header that does not have separate FEC encoding, or insert a DVB-S2 header that is heavily encoded with a (64, 7) Reed-Muller code capable of correcting up to 32 errors. The buffer assembles eight or 32 MPEG packets to form short or long frames and applies a double layer of FEC coding. The scrambler multiplies the encoded data stream by a pseudo random sequence to provide energy dispersal. In this example, the 70 MHz PSK modulator can operate in QPSK or 8-PSK modes and is followed by a SRRC band pass filter. The SRRC filter must be a digital finite impulse response (FIR) filter to match the bandwidth of the QPSK or 8-PSK signal. The PSK signals are upconverted to 700 MHz and sent to the transmitting antenna.

data stream by a pseudo random sequence to provide energy dispersal. The encoded, compressed, and multiplexed bit stream drives a QPSK or 8-PSK modulator to generate an IF carrier, typically at 70 MHz. The 70 MHz PSK signal is passed through a SRRC filter, upconverted to 700 MHz and sent to the transmitting antenna, which has multiple traveling wave tube HPAs. The HPAs are usually rated at a much higher power than their normal operating output power level, to provide sufficient output back off of the HPA to ensure linear operation. The signals from any number of HPAs are multiplexed together in microwave combiners and sent to the antenna feed for transmission to the satellite.

10.8 Installation of DBS-TV Antennas

Installation of a home satellite TV system offers an interesting challenge to home owners who do not have much knowledge of microwave antennas and satellite communication systems. A DBS-TV system antenna with a diameter of 0.45 m × 0.6 m has a beamwidth of three degrees in the azimuth plane, and needs to be pointed to an accuracy of ±0.3° for optimum reception of the satellite signal. The problem is to provide a simple method for pointing the antenna in azimuth and elevation within about two degrees so that a signal can be received and peaked. Both DirecTV and Dish Network provide on-screen instructions in their set up menus that make the process quite easy.

DBS-TV antennas are typically mounted onto a vertical steel tube with a swiveling clamp. The lower end of the tube has a mounting bracket that is bolted to any

convenient surface that provides a clear view of the southern sky, or planted securely in the ground, and the tube is set vertical using a plumb line or level. The antenna can then be rotated in azimuth. Elevation angle is set by rotating the dish about its horizontal axis using an angle scale marked on the mounting. When the dish is set to the correct azimuth and elevation angles the bolts in the clamps are tightened down and the antenna is permanently set to the correct look angles.

The azimuth and elevation look angles for each specific satellite can also be found on the setup menu of the satellite receiver or by using software that can be downloaded from web sites. The azimuth angle is given relative to magnetic north so that a compass can be used to set an approximate azimuth angle. The elevation angle can be set within one degree by careful adjustment of the clamp, which ensures that the satellite will be within the elevation plane 3 dB beamwidth of the antenna when the azimuth angle is correct.

The procedure used to find the satellite is quite simple. The antenna is rotated in azimuth until a tone is heard from the TV receiver, indicating that a signal is being received. The antenna azimuth angle is adjusted to maximize the loudness of the tone and also the signal strength value (a number between 0 and 100) shown on the screen, which ensures that the satellite is correctly pointed in azimuth. (The procedure typically requires two people because the TV set is rarely visible from the antenna installation point.) Once the satellite signal has been acquired, the azimuth and elevation angles are adjusted alternately to maximize the signal strength, and the clamp is tightened down to hold the antenna at the correct angles. When the above procedure is followed with care, the antenna can be set to the correct azimuth and elevation angles in a few minutes. Professional installers use a signal strength meter that can be connected to the IF output of the antenna's LNB to facilitate installation by one person.

10.9 Satellite Radio Broadcasting

Two companies, Sirius Satellite Inc. and XM Satellite Radio Inc., commenced transmission of digital radio signals from satellites to North America in 2001 and 2002, each offering 50 radio channels for a monthly subscription of US$10–13. Transmissions are in S-band, with Sirius using the band 2320–2332.5 MHz and XM using the band 2332.5–2345 MHz (Sirius Satellite Radio 2017). Neither of the two companies was profitable until 2008 when they merged to form SiriusXM Radio Inc.®. The signal formats of Sirius and XM differ, so the original receivers were not compatible. After the merger, programming on both services was aligned and receivers included chipsets for both services. Customers can purchase subscriptions from Sirius Radio or XM Radio and receive the same radio channels. Echostar and DirecTV include SiriusXM audio channels with some subscription packages. By 2017 the company had a reported 32 million subscribers (Satellite Radio 2017). A now defunct company called 1Worldspace® created a digital audio broadcasting (DAB) system to serve Africa and Asia using L-band frequencies 1467–1492 MHz, but never became profitable and closed in 2009 (1Worldspace 2010). Satellite radio broadcasting also exists in Asia using a standard devised by ETSI (ETSI DAB 2009).

Generically, the system is called Satellite Digital Audio Radio Service (SDARS) or DAB. The target audience in the United States is in automobiles and trucks, which is

where most radio listening occurs in the United States. DAB receivers can be operated indoors, but the antenna may need to be placed near a south facing window. A vehicle equipped with a DAB receiver can receive the same programming anywhere in North America, and is particularly valuable to anyone driving long distances across different states. Vehicles use a roof mounted antenna with near omnidirectional coverage in the visible hemisphere and a typical gain of 3 dB toward the DAB satellites. Although both satellite and cable subscription television has been very successful in the United States, SDARS is the first attempt to create a subscription radio service – in contrast to terrestrial radio broadcasting, which has always been free to the listener, supported by advertising revenue. Originally, the US DAB programming did not carry advertisements, but later up to seven minutes per hour of advertising was allowed on some channels.

The early SDARS satellites had transponder saturated power output of 800 W, but their replacements have a number of transponders operating in parallel with saturated output power up to 2.7 kW. The unusually high power radiated by these satellites compensates for the low gain of near omnidirectional antennas on vehicles. The satellites have one or two 9 m antennas that are unfurled once the satellite is in orbit. Assuming a 50% aperture efficiency, the gain of a 9 m antenna at 2330 MHz is 43.8 dB.

Both systems use terrestrial repeaters in large cities to augment the satellite signal when blockage occurs by tall buildings. XM Satellite Radio Inc., based in Washington, D.C. launched two satellites in 2001 to geostationary orbital locations at 85°W and 115°W longitudes, appropriately named "Rock" and "Roll." In 2004 the satellites developed solar array problems that caused a loss of DC power, so XM launched a second pair of satellites as replacements in 2006. The original satellites remained in orbit as back up spares, although with limited transmission capability (XM Radio 2017).

Sirius Satellite Radio Inc., based in New York City, originally launched three satellites in 2002 into an inclined elliptical geosynchronous orbit (similar to a Molynia orbit) with a period of one sidereal day, crossing the equator at a longitude of 100°W and with apogee over Hudson Bay in Eastern Canada. The satellites were equally spaced around the orbit and each was above the horizon for listeners in the United States for approximately 16 hours in each 24 hours, with two of the three satellites transmitting in separate 4.2 MHz wide bands in the frequency ranges 2320–2324 MHz and 2320–2332.5 MHz. The highly elliptical orbit of the Sirius satellites provided a higher elevation angle than a GEO satellite, which is desirable in cities to minimize blockage by tall buildings, but requires a handoff between satellites. When the original satellites came due for replacement, two new satellites were launched into geostationary orbit in 2011 and 2013 and located at 85.2°W and 116.5°W. Table 10.6 provides some details of the SDARS systems. The XM Radio system adopted many of the characteristics of the Sirius system after the merger of the two SDARSs. Figure 10.17 shows a drawing of a second generation Sirius-XM satellite. Large antennas are needed to create beams that covers the United States at S-band. The antennas are stored against the side of the satellite for launch and unfurl like an umbrella once the satellite is in GEO orbit.

Terrestrial repeaters operate in the same frequency bands, to provide a substitute signal when the satellite line of sight is blocked. Because of the high probability of the satellite signals being blocked by buildings in a city and trees in rural areas, both systems utilize time diversity to overcome short interruptions in signal. The transmissions from one of the two satellites, and from the terrestrial repeaters, are delayed by four seconds.

Table 10.6 US satellite digital audio radio service

Parameter	XM satellite radio	Sirius satellite radio
Number of Satellites	Two in GEO at 85°W and 115°W	Originally three in highly elliptical 24 hour orbit at 100°W. After 2010, two GEO at 85.2°W and 116.5°W
Downlink Frequencies	2320–2332.5 MHz	2332.5–2345.0 MHz
Uplink Frequencies	7050–7075 MHz	7060–7072.5 MHz
Saturated transponder power		
Early satellites	800 W	800 W
Later satellites	1000 W	2.7 kW
Terrestrial Repeaters	600 in 70 cities	Originally 105 in 46 cities Later coordinated with XM
Total number of Audio channels	110	110
Transmission rate before FEC coding	4.2 Mbps	4.5 Mbps
Satellite downlink transmission	TDM-QPSK with half rate FEC	TDM-QPSK with half rate FEC

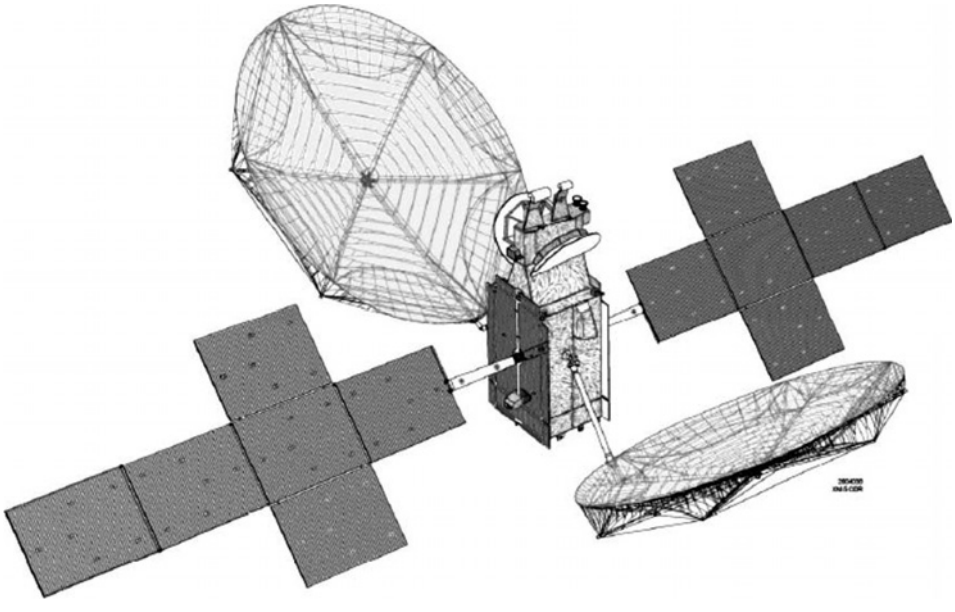


Figure 10.17 Drawing of second generation Sirius-XM satellite. The solar cells fold in over one another to lie flat against the side of the satellite body for launch. The 9 m antennas are stored against the side of the satellite for launch and unfurl like an umbrella once the satellite is in GEO orbit. This antenna has a Gregorian configuration, with a concave subreflector above the feed. Source: FCC SiriusXM filing (2009).

The satellite radio receiver delays the earlier arriving signals to achieve common timing and then selects or combines the signals to achieve the best signal to noise ratio (SNR) (Satellite Radio 2017). Signal transmission formats from the satellites are very similar to those used in DBS-TV: TDM with QPSK modulation is used to send multiple signals as a high speed digital data stream, with half rate FEC coding for error control. The 4 Mbps bit stream from each satellite is used to support voice and music channels, in a ratio of roughly 40 voice to 70 music channels. Audio channel bit rate is 4.0 kbps for voice signals and up to 128 kbps for music. The basic channel bit rate of 4.0 kbps employs heavy compression of the audio signals and several different commercial music compression techniques are used. Some 4 kbps channels are used for control and data transmission.

Two satellites used by both systems transmit the same information with a four second delay between the data streams. Data from both satellites is stored in a memory in the receiver so that when signal is lost from the later satellite because of an obstruction in the satellite path, the receiver can go back four seconds and fill in the missing data from the previously recorded signal of the earlier satellite. This technique allows vehicles to lose connection with both satellites for up to four seconds when transiting underpasses and tall buildings (Satellite Radio 2017).

Terrestrial signals transmitted in the same frequency bands using OFDM techniques provide an additional back up source in city areas where vehicles may be stationary behind tall buildings for much longer than four seconds. Orthogonal frequency division multiplexing is used for terrestrial broadcasting and 5G mobile communications as it has better resistance to multipath than other modulations (OFDM 2017).

Monthly subscription rates in 2018 to SiriusXm® were in the range US$12–US$20 for the regular 110 channels of voice and music, with equipment costs varying from US$30 to US$200. Some automobile manufacturers include SiriusXM® radios in their new cars, and some rental car agencies provide the service in their vehicles.

SiriusXM also offers weather related services for boats and aircraft. Marine subscriptions ranged from US$13 to US$55 and aviation services ranged from US$35 to US$100 per month in 2018; both services provide weather forecasts and near real time Nexrad weather depictions. The Nexrad weather radar output is particularly valuable to pilots of small general aviation aircraft that do not carry an airborne radar, because they must stay well away from thunderstorms. Encounters with strong thunderstorms do not usually end well for small aircraft, as vertical wind speeds of 100 mph are common inside thunderstorms and frequently lead to loss of control of the airplane. Foreflight® provides navigation and weather data that can be displayed on a tablet in the cockpit, including weather radar data from SiriusXM, at a subscription price of US$99 per year and is a preferred option for many pilots of general aviation aircraft (Foreflight.com 2017). Weather radar data and forecasts are also available over the free ADS-B-IN service operated by the Federal Aviation Authority (FAA) in the United States. See Chapter 12 for details of automatic dependent surveillance broadcast (ADS-B).

SiriusXM uses the 7 GHz uplink band allocated to government and military satellite systems, suggesting that some government agencies make use of the DAB satellites to deliver voice and data across the entire United States. The XM Radio uplink antennas are located in Washington, D.C., close to the Pentagon and many government agencies. Sample uplink and downlink budgets are shown in Tables 10.7a and b.

The minimum link margin in the service area is 10.7 dB to achieve the required service availability. Additional margin is useful to overcome attenuation caused by heavy foliage.

Table 10.7a Sample Link Budget for SiriusXM satellite to a vehicle in NE United States

Spacecraft Saturated EIRP	71.0 dBW
Transmitted Frequencies	2322.293 MHz, 2330.207 MHz
Path Loss	191.2 dB
Atmospheric loss allowance at 2.33 GHz	0.5 dB
Receiving antenna gain at elevation angle 40°	3.7 dB
Received power	−117.0 dBW
Receiving system noise temperature 158 K	22.0 dBK
Receiver noise bandwidth 4.5 MHz	66.5 dBHz
Boltzmann's constant	−228.6 dBW/K/Hz
Receiver noise power	−140.1 dB
Downlink CNR in vehicle receiver	23.1 dB
Minimum operating CNR	4.0 dB
Implementation loss in vehicle receiver	1.0 dB
Uplink CNR in clear air (see Table 10.7b)	42.7 dB
Overall CNR in clear air	22.2 dB
Link margin over 4.0 dB threshold	18.2 dB

Source: Data from reference (FCC Sirius-XM FM-5 satellite).

Table 10.7b Sample SiriusXM Uplink Link Budget

Uplink antenna diameter	7.2 m
Uplink transmit frequency	7062.293 MHz, 7070.207 MHz
Antenna aperture efficiency	60.0%
Uplink antenna gain	52.3 dB
HPA saturated output power	20.0 dBW
HPA backoff	3.0 dB
HPA to antenna loss	1.5 dB
Satellite receiving antenna gain 2.2 m diameter	42.4 dB
Path loss	200.9 dB
Transmit antenna on −1.0 dB contour of satellite antenna	1.0 dB
Received power at satellite	91.7 dBW
Receiving system noise temperature 650 K	28.1 dBK
Receiver noise bandwidth 4.5 MHz	66.5 dBHz
Boltzmann's constant	−228.6 dBW/K/Hz
Receiver noise power	−134.0 dBW
Link marging over 4 dB threshhold	18.2 dB

Source: Data from reference (FCC Sirius-XM FM-5 satellite).

10.10 Summary

Satellite broadcasting of television has become a major part of the satellite communications industry. In 2016, DBS-TV and video distribution earned more than half the revenues of the satellite communications industry, worldwide. Most DBS-TV and distribution of video signals is now digital, and DirecTV and Echostar in the United States have been major success stories with a total of 32 million customers by end 2017. SES serves 60M customers in Europe with DBS-TV and video distribution. India has the highest number of DBS-TV users, with an estimated 60M households receiving TV via satellite.

DBS-TV systems operate with small antennas and low cost receiving systems, and offer a very large number of video and audio channels, competing directly with cable TV and exceeding the number of channels available from terrestrial television broadcasting stations. The link budget for a typical DBS-TV signal shows that the link margin is in the 4–8 dB range, which yields a better than 99.7% availability in the United States. Shaping of the transmitted beam and the use of high power spot beams provides higher clear sky CNR ratios in regions where heavy rainfall occurs most often, such as the south east of the United States.

Standard definition digital DBS-TV signals are transmitted by first generation satellites with a Conus footprint, using QPSK modulation and half rate FEC coding in the DVB-S format. The standard RF channels for DVB-TV in Ku-band are 30 MHz wide, with overlaid transmissions in LHCP and RHCP and 15 MHz offset between center frequencies. SRRC filters with $\alpha = 0.35$ are used giving a data rate of 22 Mbps. The second generation of DBS-TV satellites transmit Conus and spot beams using the DBV-S2 format with QPSK or 8-PSK modulation and half rate or 2/3 rate FEC coding and MPEG-2 or MPEG-4 video compression. DBS-TV digital signals make extensive use of error correction techniques in the form of a double layer of error control coding with interleaving. The DVB-S signal format uses an inner layer of convolutional coding and an outer layer of Reed-Solomon linear block coding. The DVB-S2 format employs LDPC coding as the inner layer and BCH coding with burst error correction for the outer layer, and can have SRRC filters with $\alpha = 0.25$ or 0.20 to achieve a spectral efficiency of 2.0–2.5 bits/Hz, giving transponder capacity between 45 and 60 Mbps. The higher bit rates are useful for the transmission of HDTV signals using MPEG-4 compression, which requires up to 5 Mbps for each video channel compared to 2.0 Mbps for standard definition TV with MPEG-2 compression. Ultra HDTV requires even higher bits rates.

The second generation DBS-TV satellites with more than 50 spot beams deliver local TV programming to cities and regions across the United States, and a third generation of GEO satellites with hundreds of spot beams and capacity up to 500 Gbps were being proposed in 2018. These satellites operate in Ka-band where more RF bandwidth is available than in Ku-band, and are intended primarily for internet access, but can also be used for DBS-TV service on demand.

Satellite radio broadcasting commenced in the United States in 2001 from three Sirius satellites in elliptical orbits and two XM satellites in GEO. Sirius and XM merged in 2009 to form SirusXM radio® and replaced the satellites with four in GEO orbit having higher transmit power. The signals are transmitted in S-band at 2.3 GHz with transponder output power up to 2.7 kW, and are aimed primarily at automobiles, which is where most people in the United States listen to the radio. The high satellite transmit power is required because automobiles typically use near omnidirectional antennas with low

gain. Repeaters are used in city areas to overcome signal blockage by tall buildings. SiriusXM had 17M customers by the end of 2017.

Exercises

10.1 Table 10.5 shows a downlink budget for a Ku-band DBS-TV receiver, which receives a digital TV signal in the DVB-S format. The following parameters of the downlink are changed: The EIRP of the satellite transmission is 54.0 dBW, the gain of the receiving antenna is 35.0 dB and the receiving station is located on the −4 dB contour of the satellite transmitting antenna footprint. All other parameters listed in Table 10.5 are unchanged. Calculate the new overall CNR in the earth station receiver when the uplink CNR of the transponder is 30 dB.

10.2 The link in Question #1 suffers interference from a cross polarized signal in the same frequency band at a level of −24 dB. Calculate the new overall CNR in the earth station receiver.

10.3 The link in Question #1 is upgraded to accept signals in the DVB-S2 format, which are transmitted using 8-PSK modulation with 2/3 rate FEC encoding. The earth station receiving antenna is replaced with a larger antenna with 36 dB on axis gain.

- **a.** The earth station receiver bandwidth remains at 22 MHz. What is the symbol rate of the 8-PSK signal with rate 2/3 FEC? If SRRC filters with $\alpha = 0.25$ are used throughout the link, what is the occupied bandwidth of the RF signal?
- **b.** Calculate the clear sky overall CNR for the downlink with an uplink CNR of 30 dB, no interference, and a receiving station on the −3 dB contour of the satellite transmit antenna footprint. The link has an implementation margin of 1.8 dB. Find the downlink margin for 8-PSK modulation with 2/3 rate FEC.
- **c.** The noise power budget of the link in Table 10.5 is based on a receiver noise bandwidth of 22.0 MHz, an antenna noise temperature of 35 K in clear air, and a 12 GHz LNA with a noise temperature of 75 K. Find the downlink rain fade margin. (Note: Requires an iterative solution to obtain noise power increase and downlink power decrease that matches the downlink margin.)

10.4 DTH-TV links using a Conus transmit beam from the satellite typically employ the DVB-S signal format with QPSK modulation and half rate FEC encoding. DBS-TV satellites with spot beams can use the DVB-S2 signal format with 8-PSK and 2/3 rate FEC, which delivers a higher bit rate. Explain how the increase in bit rate is achieved without compromising the link availability. What improvement to the quality of the signal observed on the TV screen is possible with the higher bit rate?

10.5 DBS-TV systems serving the United States typically have satellite antenna coverage patterns with a subsidiary beam that provides service to customers in the Hawaiian Islands. The EIRP of the subsidiary beam is significantly lower than main conus beams that serve the 48 contiguous states. This question examines

the options available to the designer of a DBS-TV system that serves the continental United States and the Hawaiian Islands. The system uses the DVB-S2 standard. The peak EIRP of the conus beams is 60.0 dB and the lowest EIRP is 55.0 dB. The peak EIRP of the Hawaiian beam is 50.0 dB and all of the Islands are within the −1 dB contour of this beam.

a. The standard antenna for customers in the continental United States has a circular aperture with a diameter of 0.6 m. Calculate the gain of this antenna at 12.2 GHz for an aperture efficiency of 70%.

b. Receivers that are located in areas where the conus beam EIRP is 55.0 dB have a link margin of 3 dB over the minimum CNR to achieve the DVB-S2 quality of service when signals are transmitted using 8-PSK modulation with 2/3 rate FEC. If the standard antenna is installed at a Hawaii location on the −1 db contour of the Hawaiian Islands beam, what is the link margin?

c. The standard antenna installation could be used in Hawaii, but the modulation and FEC would have to be changed to meet the quality of service requirement. What combinations of modulation and coding could be used, and what impact would this have on the capacity of the links to Hawaii?

d. A larger antenna is needed for Hawaiian receivers to match the performance of receiving systems on the mainland. What diameter of circular aperture antenna would you recommend?

10.6 What are the major differences between the DVB-S and DVB-S2 standard for DTH-TV transmission? Include the differences between SRRC filter roll-off value, modulation and FEC coding choices, and the range of CNR values that can be used while still meeting QEF objectives.

10.7 The DVB-S standard uses a double layer of convolutional and Reed-Solomon coding for forward error correction on DBS-TV links. The DVB-S2 standard uses BCH and LDPC coding for forward error correction, with additional Reed-Solomon FEC coding of individual packets. What advantages does the DVB-S2 coding method achieve over the earlier DVB-S method?

10.8 Frames in the DVB-S2 standard transmissions start with a 90 bit header that includes a Reed-Muller (64, 7) code capable of correcting up to 32 errors. Why is such a powerful error correcting code used in the frame header? What is the impact of an error in the frame header? Why is this code not used for the remainder of the frame?

10.9 The DVB-S2 standard format is used with 16-APSK and 32-APSK modulation for electronic news gathering (ENG). Satellite trucks transmit signals using this format via GEO satellites to receiving installations at TV broadcast stations that have a large antennas, typically with diameters of 3–4.5 m. Explain why ENG systems can use these higher order modulations while DTH-TV systems cannot. What additional demands are placed on the stability of transmitters and receivers used in ENG links, compared to those in DTH-TV links?

10.10 You have the opportunity to design a satellite digital audio radio system (SDARS) in which the receiving terminals are equipped with self steering

phased array antennas that have a minimum gain of 13.0 dB in the direction of the satellite. Base your design on the link budgets of Tables 10.7a and b.

a. What reduction in satellite transmit power is possible with the phased array receiving antenna compared to the value required in Table 10.7a?

b. Does the higher gain receiving antenna permit any changes to the parameters in Table 10.7b?

c. The phased array receiving antenna has a higher noise temperature than the fixed antenna of the receiver in Table 10.7a. If the steerable antenna increases the system noise temperature from 158 K, as indicated in Table 10.7a, to 193 K what additional power must be transmitted by the satellite to compensate for this change? Give your answer in decibels.

d. The data rate of the time division multiplexed baseband signals transmitted to your SDARS satellite is 4.2 Mbps after half rate FEC encoding, and the system carries 100 digital audio channels with speech and music. The speech and music channels use advanced compression techniques to reduce the bit rate of the digital audio signals.

i) What is the average bit rate of a baseband channel?

ii) The 100 channels are divided between 50 voice channels and 50 music channels. Voice channels use audio compression that reduces the baseband bit rate to 4.8 kbps. What is the average baseband bit rate for a music channel?

e. An improved compression technique is available for music channels that reduces the baseband bit rate to 30 kbps. This allows some music channels to have higher baseband bit rate for better audio quality. How many additional higher quality channels with a baseband bit rate of 120 kbps can be transmitted when 50 music channels at 36 kbps and 50 voice channels at 4.8 kbps must also be sent?

References

1Worldspace (2010). https://en.wikipedia.org/wiki/1worldspace (accessed 9 April 2018).

Alencar, M.S. (ed.) (2009). *Digital Television Systems*. Cambridge, UK: Cambridge University Press.

ATSC Standards (2018). https://www.atsc.org/standards/atsc-standards (accessed 29 March 2018).

Bose, D.C. and Ray-Chaudhury, D.K. (1960). On a class of error correcting binary group codes. *Information and Control* 3: 68–79.

DBS-TV growth (2018). http://www.digitaltvnews.net/?p=30154 (accessed 29 March 2018).

DirecTV (2018). https://en.wikipedia.org/wiki/DirecTV (accessed 29 March 2018).

Echostar (2018). https://en.wikipedia.org/wiki/EchoStar (accessed 29 March 2018).

Echostar 14 FCC (2009). licensing.fcc.gov/myibfs/download.do?attachment_key=-165508 (accessed 14 April 2018).

ETSI (1997). Digital broadcasting systems for television, sound and data services; Framing structure, channel coding and modulation for 11/12 GHz satellite services. ETS EN 300 421. Sophia-Antipolis, France, ETSI, 1997.

ETSI (2003). DVB-S2 standard. http://www.etsi.org/deliver/etsi_en/302300_302399/30230701/01.04.01_60/en_30230701v010401p.pdf (accessed 15 June 2018).

ETSI (2009). Revised DVB-S2 standard. ETSI EN 302 307 V1.2.1 (2009-08) Sophia-Antipolis, France, ETSI, 2009.
ETSI (2015). DVB-S2 guide. *Digital Video Broadcasting (DVB) Implementation guidelines for the second generation system for Broadcasting, Interactive Services, News Gathering and other broadband satellite applications; Part I (DVB-S2)*, DVB Document A171-1, Sophia-Antipolis, France, ETSI, March 2015.
ETSI DAB (2009). https://en.wikipedia.org/wiki/ETSI_Satellite_Digital_Radio (accessed 18 June, 2018).
Foreflight.com (2017). www.foreflight.com (accessed 11 April 2018).
History of Satellite TV (2016). https://itechworld.com.au/blogs/learn/89115078-the-history-of-satellite-tv (accessed 3 April 2018).
Hocquenhem, A. (1959). Code correcteurs d'erreurs. *Chiffres* 2: 147–156.
India DBS-TV (2017). https://en.wikipedia.org/wiki/Direct-to-home_television_in_India. (accessed 18 March 2018).
MPEG (2018). https://en.wikipedia.org/wiki/Moving_Picture_Experts_Group (accessed 18 April 2018).
OFDM (2017). http://www.radio-electronics.com/info/rf-technology-design/ofdm/ofdm-basics-tutorial.php (accessed 11 April 2018).
Satellite Downlink Coverage Patterns (2018). http://www.lyngsat-maps.com (accessed 4 April 2018).
Satellite Footprints (2018). http://www.satbeams.com/footprints (accessed 15 July 2018).
Satellite Radio (2017). http://satelliteradiousa.com/satellite_radio_how_does_it_work.html (accessed 11 April 2018).
Satellite Television (2018). https://en.wikipedia.org/wiki/Satellite_television (accessed 3 April 2018).
SBCA (2013). http://www.sbca.com/receiver-network/history-satellite-providers.htm (accessed 16 June 2018).
SES (2018). www.ses.com.lux (accessed 29 March 2018).
Sirius Satellite Radio (2017). https://en.wikipedia.org/wiki/Sirius_Satellite_Radio (accessed 9 April 2018).
Sirius-XM Holdings (2007). https://en.wikipedia.org/wiki/Sirius_XM_Holdings (accessed 10 April 2018).
XM Radio (2017). https://en.wikipedia.org/wiki/XM_Satellite_Radio (accessed 10 April 2018).

11

Satellite Internet

In the early days of the internet, known then as the world wide web, the only way to connect your home computer to the rest of the world was with a dialup modem that used your home telephone line. The earliest telephone line modems provided two-way bit rates of 2.4 kbps using frequency shift keying (FSK), which later increased to 33 kbps using quadrature amplitude modulation (QAM). Telephone line subscribers who lived within a few miles of a telephone exchange could buy into a digital subscriber line (DSL) with a bit rate of 56 kbps. The development of lower cost home computers in the 1990s and the popularity of the worldwide web created a demand for higher speed connections, so DSL service was improved to provide steadily higher bit rates and cable television companies started to offer internet service over their cable TV lines. Eventually, a number of telecommunications companies began to install optical fibers to the home, with the promise of truly broadband connections and bit rates in the hundreds of megabits per second. However, these services have always been limited to urban and suburban areas where population density is sufficient to ensure that the telecommunication companies could make a profit, given the high cost of installing cables and plant for cable TV or optical fibers. Local area wireless internet access systems were also developed but have not been particularly successful in the United States, although cellular telephone companies have provided internet access within their coverage zones. In many rural areas, well away from towns and cities, satellite internet access, often called satellite broadband, has been the only option. In 2016 it was estimated that 14 million homes in the United States did not have broadband access to the internet, defined at 25 Mbps download and 3 Mbps upload rates (Measuring broadband America 2016).

11.1 History of Satellite Internet Access

If you live in a sparsely populated rural area and do not subscribe to satellite internet access, a telephone line might be your only way to connect to the internet, at a rate no higher than 33 kbps. If cellular telephone service is available, data connections can be purchased providing download bit rates up to 8 Mbps. This is where internet access by satellite has found its largest market: serving homes where there is no other high speed data service. Many countries do not have a well developed terrestrial communications network making it impossible for large numbers of people to access the internet. Low earth orbit (LEO) constellations of satellites can serve those populations and several such systems were in development in 2018. OneWeb and SpaceX were the leading LEO

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

companies, with proposals to create constellations of thousands of satellites serving all parts of the earth's surface.

11.1.1 Early History

Satellite internet access has had a checkered history in the United States. Service began in the United States in 1996 when Hughes Network Systems offered downlinks via geostationary satellite with return link via a telephone modem, with a service called DirectPC. Hughes Network services created the Directv satellite TV service in the mid 1990s and DirectPC was a not very successful offshoot using the Directv satellite transponders for the downlink. Generally, internet access is asymmetric, requiring higher download speed than upload speed since uploads are generally requests for downloads such as video or social media pages. This pattern is common to all internet services. The DirectPC system evolved to a two-way satellite access called Directway in 2001 and HughesNet in 2005, which had 300 000 subscribers by 2006 (HughesNet 2006). The HughesNet service initially offered downlink rates of *up to* 2 Mbps and uplink rates of *up to* 128 kbps. The *up to* wording was important, because the advertised rates were not what customers experienced as *throughput*, especially at peak hours in the evenings. Throughput is the average bit rate that the user actually observes. The system was *oversubscribed*, and throughput could slow down at busy times. Data is delivered in packets, and with many users sharing a common data stream, packets arrive at any individual user less and less often as more and more people try to access the same data stream. The observed data rate falls, and only at quiet times (the early hours of the morning, for example) will the user bit rate come close to the advertised rate. The download bit rate was 2 Mbps as advertised, but the downlink was shared by so many customers at busy times that the time between packet arrivals reduced the throughput to a trickle.

> One of this text's authors (TP) subscribed to HughesNet in the 2000s and recalls trying unsuccessfully to make airline reservations in the evening. The link was so slow that the airline web site would repeatedly time out before the transaction completed, a very frustrating situation. Complaints about poor service and lack of truth in advertising were common, giving satellite internet a bad name.

Oversubscription refers to sharing a common resource, in this case satellite transponders, between too many customers. A 36 MHz satellite TV transponder operating with DVB-S format signals can typically carry a bit rate of 27 Mbps using quadrature phase shift keying (QPSK), half rate forward error correction (FEC), and square root raised cosine (SRRC) filtering with alpha of 0.35. Using time division multiplexing (TDM) the 27 Mbps can be divided into five streams at 5.4 Mbps; after removing headers, the data rate of each stream will be around 5 Mbps. Now suppose we connect a number of customers to this bit stream to sell internet access. Not all the customers will be downloading data at the same time, so providers use a *contention ratio* to determine how many customers to connect to a given bit stream. In terrestrial internet access systems such as DSL and cable, a contention ratio of 20 : 1 is regarded as very good, with little chance of throughput rate dropping at busy times. A contention ratio of 50 : 1 will result in the user bit rate slowing down when too many users try to access the links. At 100 : 1 contention ratio, the throughput will drop to a low rate at busy times. At times with low

activity, say at 3 a.m., the few customers using internet access will likely see download rates approaching 5 Mbps. The temptation for the internet access provider is to oversubscribe the service to increase revenue by raising the contention ratio, which is why there were many complaints about satellite internet in the United States prior to 2011. With a contention ratio of 100 : 1 a single 27 MHz bit stream from one transponder can support 500 users, and a satellite with 48 transponders can service 24 000 users. To provide 300 000 customers with internet access at *up to* 5 MHz and a contention ratio of 100 : 1 would require 13 Ku-band satellites and 624 transponders devoted to internet access. There were fewer than 13 Ku-band satellites devoted to internet access in 2006, so contention ratio was well above 100 : 1.

The capacity of satellites increased as newer satellites incorporated spot beams and Ka-band transponders. Anik F2, launched in 2004 to serve Canada with direct broadcast satellite television (DBS-TV) and digital data services, carried 24 C-band and 32 Ku-band transponders with a footprint oriented toward Canada, and also had 45 Ka-band beams covering all of North America. Total capacity was quoted as 2 Gbps (Anik F2 2018). By 2007, the Wild Blue satellite was providing *satellite broadband* for 250 000 customers in the United States from two Ka-band satellites with 35 spot beams and a reported capacity of 7 Gbps (Wild Blue 2007). Wild Blue was sold to Viasat in 2009.

11.1.2 ViaSat and HughesNet

In 2011 ViaSat Inc. launched the ViaSat 1 Geostationary Satellite Orbit (GSO) satellite and achieved a step function in digital transmission capacity. (In this chapter *GSO* is used instead of geostationary earth orbit (GEO) for geostationary orbit to conform with non-geostationary orbit literature.) As noted in the previous section, Ku-band satellites with regional beams could not provide sufficient capacity to meet a growing demand for satellite internet access. ViaSat satellites and HughesNet Jupiter satellites were designed to use Ka-band frequencies and digital transmission to achieve capacities in the 100–500 Gbps range. ViaSat 1 was launched in 2010 and became operational in 2011. The initial capacity of ViaSat 1 was expected to be 100 Gbps, but was later raised to 120 Gbps and then 140 Gbps after in-orbit testing was completed. The quoted capacity figures represent 100% loading of the satellite, which may not be achieved in a system with thousands of customers accessing the system at random intervals.

The ViaSat 1 satellite is a Ka-band satellite with 72 beams covering the more densely populated areas of the United States. The satellite is operated by Echostar® and originally sold broadband internet access under the trade name Exede® with downlink bit rates up to 15 Mbps and uplink bit rates up to 3 Mbps. The Exede service was later renamed ViaSat®. It does not provide coverage of many states in the western half of the United States but has individual beams for Denver, Colorado, Phoenix, Arizona, and a sequence of beams along the West coast. Echostar uses the Wild Blue satellite to provide satellite internet access to the regions of the United States not covered by ViaSat 1, and moved Wild Blue customers in regions that are covered by Viasat 1 beams to the ViaSat 1 satellite. Bit rates for people living in areas not covered by ViaSat spot beams are lower, but the whole capacity of the Wild Blue satellites is devoted to their region.

Echostar acquired HughesNet in 2009 and in 2012 launched Echostar 17, formerly called Jupiter 1, a satellite similar to ViaSat 1 with 60 Ka-band beams covering the entire continental United States, Alaska, Puerto Rico, and Hawaii. Capacity is quoted as 100 Gbps and the launch on an Ariane V ELV cost $118.5 M. A second satellite, Echostar

19, with 138 beams covering North America and 220 Gbps capacity was launched in December 2016 and entered service in March 2017 (Echostar 19 2017).

Echostar 24 is expected to have a total throughput of 500 Gbps and to launch in 2021; ViaSat expects to launch the first ViaSat-3 in 2020, with an expected capacity of 1000 Gbps to cover the Americas (Echostar 24 2017).

11.2 Geostationary Satellite Internet Access

Figure 11.1 illustrates the concept of a GSO internet access system. A large antenna at a gateway station connects to the internet via a fiber optic line, and sends signals to and receives signals from a large geostationary satellite. The satellite is similar to those used for direct to home television (DTH-TV) transmissions, and typically has many spot beams with multiple frequency reuse. The user terminal is an offset paraboloid reflector antenna, similar to those used for DTH-TV, but often with an elliptical outline and larger than the DTH-TV antenna. Wide dimension (in the plane of the satellite orbit) is typically 0.7–1.0 m.

11.2.1 ViaSat 1 Satellite and Footprint

ViaSat 1 was the first *high throughput* broadband internet access satellite, achieving a 10-fold increase in capacity over earlier GSO satellites. Figures 11.2 a,b show the ViaSat 1 satellite and its coverage map. In 2018, in the regions not covered by ViaSat 1 customers were served by two Wild Blue satellites. There are four Gregorian reflector antennas with subreflectors mounted at the left end of the satellite as seen in Figure 11.2a. The feeds for the antennas are opposite the subreflectors on the body of the satellite. Both the

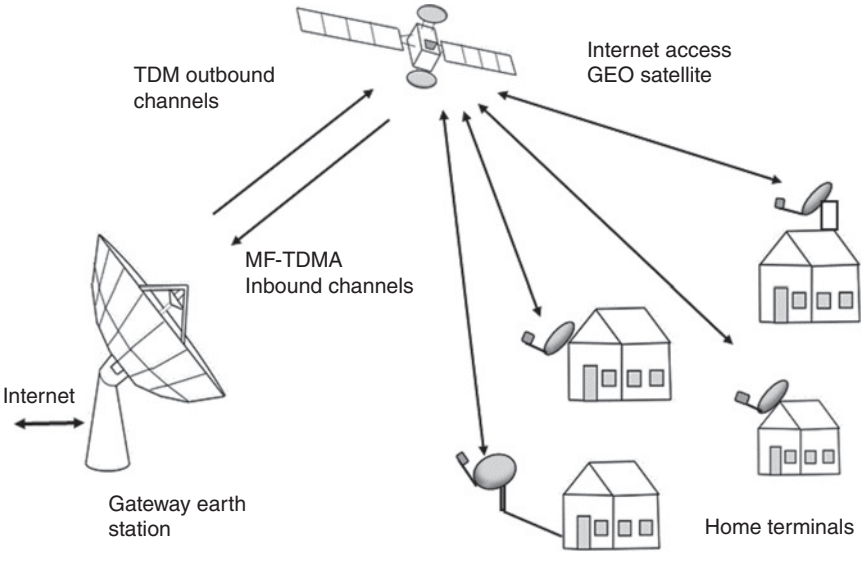
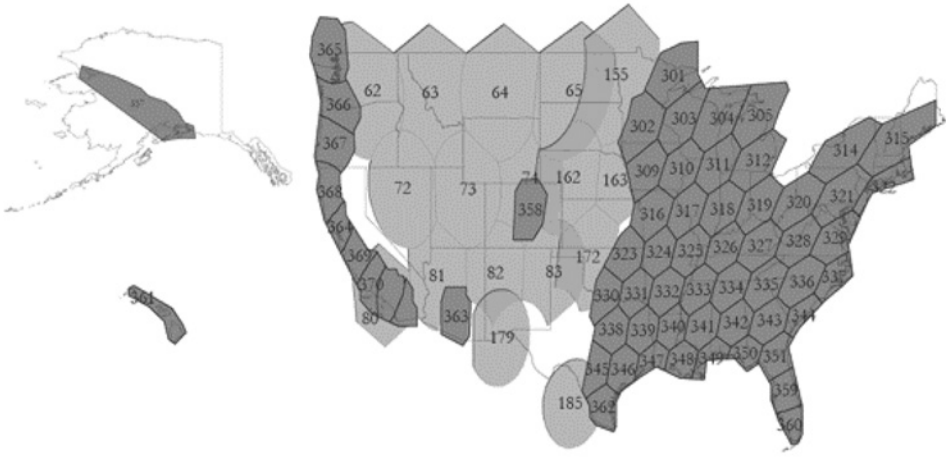


Figure 11.1 Illustration of a GSO internet access system. The user terminals are equipped with offset reflector antennas 0.7–1.0 m diameter.

(a)



(b)

Figure 11.2 (a) Illustration of ViaSat 1. For a color version of this figure please see color plate section. (b) Coverage map. There are 72 beams covering the eastern half of the United States and the west coast. Individual beams cover Denver, Colorado, Phoenix, Arizona, Alaska, and Hawaii. ViaSat 1 beams are numbered in the three hundreds. The other beams that fill in the gaps in ViaSat 1 coverage are provided by two Wild Blue satellites. The area covered by each spot beam is smallest in the SE of the United States where heaviest rainfall occurs and larger further north and over Canada where there is less heavy rainfall. The smallest of the spot beams have the highest downlink EIRP. Images courtesy of ViaSat, © ViaSat Inc 2018.

subreflectors and the main reflectors of the four reflector antennas fold down against the body of the satellite for launch. Two of the four reflector antennas provide links to the users and two provide links with the gateway station. Figure 11.2b shows the footprint of ViaSat I. There are 72 beams covering the eastern half of the United States and the west coast. Individual beams cover Denver, Colorado, Phoenix, Arizona, Alaska, and Hawaii.

Viasat 1 beams are numbered in the three hundreds; additional coverage is provided by two Wild Blue satellites with beams having two digits and numbered in the hundreds.

ViaSat 1 has 53 bent pipe transponders with bandwidths between 200 and 500 MHz. This is a much wider bandwidth than earlier satellites. Through a combination of multiple beams, multiple transponder frequencies, and orthogonal polarizations, ViaSat 1 achieves 18-fold frequency reuse in a 1.5 GHz downlink band made up of a 1 GHz band 18.3–19.3 GHz and a 500 MHz band 19.8–20.3 GHz. In addition, the uplink frequency band 29.7–30.2 GHz is used for downlinks; uplink stations using this band are located in the western part of the United States where there are no downlinks. The effective bandwidth of the satellite is 36 GHz, requiring an average spectral efficiency of 3.33 bits/Hz to create a capacity of 120 Gbps through the use of higher order phase shift keying (PSK) modulations, mainly 16-APSK. Operation with 32-APSK is also possible under clear sky conditions with an increase in capacity to 140 Gbps.

Eutelsat has a similar satellite to Viasat 1 called Ka-Sat with 82 spot beams operating in Ka-band covering Europe and parts of North Africa (Ka-Sat 2012, 2018). The major features of the two satellites are presented in Table 11.1.

All broadband satellite internet access systems establish a virtual circuit between the gateway station and each user. Packets are delivered on the outbound link from the gateway to the user in TDM streams, and on the return link, multi-frequency-time division multiple access (MF-TDMA) is used to send packets from the user to the

Table 11.1 Major features of ViaSat 1 and Ka-Sat satellites

Ka band satellite	ViaSat 1	Ka-SAT
Location in GSO	115°W longitude	9°E longitude
Satellite manufacturer	Space Systems Loral	EADS-Astrium
Launch vehicle	Proton-M	Proton-M
Type designation	SSL 1300	Eurostar E 3000
Frequency band	Ka-band	Ka-band
Transponders	56 Ka-band	82 Ka-band
Bandwidth	200–500 MHz	237 MHz
Capacity	140 Gbps	90 Gbps
Solar power system at beginning of life	16 kW (Estimated)	16 kW
Frequency reuse	18-fold	20-fold
Spot beams	72	82
Station keeping thrusters	Plasma thrusters	Plasma thrusters
Mass at launch	6000 kg	6150 kg
Antenna beam EIRP	52.1–60.7 dBW	≥60 dBW
Polarization	LHCP and RHCP	LHCP and RHCP
Modulation, coding (DVB-S2 standard)	Adaptive 16-APSK, 8-PSK, QPSK, many FEC rates	Adaptive 16-APSK, 8-PSK, QPSK, many FEC rates

LHCP, left hand circularly polarized; RHCP, right hand circularly polarized

gateway station. Adaptive coding and modulation (ACM) is employed to counter rain attenuation, so throughput to an individual user slows down when heavy rain affects the link. The destination of an outbound packet is determined by uplink frequency and polarization; each bent pipe transponder has a specific center frequency and connects to one or more downlink spot beams. The power level of the transponders can be controlled from earth, which allows lightly loaded transponders that do not have fully occupied bandwidth to operate at lower power. Fully loaded transponders can be operated at higher power levels.

11.2.2 Ka-Band Link Performance With Adaptive Coding and Modulation

The DVB-S2 standard can be used for broadband internet access by implementing the digital video broadcast-return channel satellite (DVB-RCS) return channel and thus create a virtual circuit between the gateway station and each user terminal. ACM allows the gateway and user terminal to change the FEC code rate and modulation for an individual terminal to make best use of the receiver carrier to noise ratio (CNR) and to combat fading of the signal caused by rain in the links. In the following example, a typical Ka-band satellite sends a TDM bit stream to multiple terminals in a spot beam at a maximum bit rate of 360 Mbps via the outbound (or forward) channel. The return (or inbound) link uses MF-TDMA to send bursts at 22.5 Mbps to the gateway. The transponders on the example satellite have a bandwidth of 250 MHz allowing two outbound TDM streams, and up to 10 return MF-TDMA return signals in a spot beam. The bit rates for individual users can vary depending on the level of service purchased by the user, but might be 25 Mbps for downloads and 3 Mbps for uploads. A typical link budget is presented in Example 11.1 for this GSO internet access system.

All users may not receive the same throughput in this internet access system. Signal strengths vary through the spot beam, being highest at its center and falling to a nominal level 3 dB below the center value at the edge of the spot beam. The effective isotropically radiated power (EIRP) delivered by the satellite also varies between spot beams, because spot beams in regions with high rainfall have narrow beams and therefore higher transmit antenna gain compared to the broader beams with lower EIRP used for regions with less heavy rain. Within the spot beam, user terminals within the −1 dB contour of the beam have stronger signals than those lying on the −3 dB contour. The ACM system can take advantage of the higher signal strengths to increase the burst rate on some links, allowing shorter packets, and then compensate on lower signal strength links with longer packets, to even out variations in throughput, thus increasing the overall efficiency of the satellite system. This is in contrast to DTH-TV systems that typically deliver the same bit rate to all users in a given beam, whether regional or spot.

Example 11.1 Typical Link Budget for a Ka-Band Broadband Internet Access System

An example of a typical link budget for a Ka-band broadband internet access GSO satellite system is presented in Table 11.2 for a home user terminal operating in clear sky conditions. The downlink operates in the 18.8–19.5 GHz band and the uplink in the 28.6–29.75 GHz band. Mid band frequencies are used in path loss calculations in Table 11.2. The user terminal has a 0.75 m (30 in.) antenna with an assumed efficiency of 70%. The spot beams of the Ka-band satellite have a wide range of EIRP values,

Table 11.2 Typical GSO Ka-band broadband internet access system link budgets for GSO satellite to user terminal (outbound channel) and user terminal to GSO satellite (inbound channel). The link uses the DVB-S2 with RCS standard with 1.8 dB implementation margin

Parameter	Downlink 19.5 GHz to user terminal		Uplink 29.0 GHz to satellite	
EIRP on beam axis (maximum)		60.7 dBW		45.9 dBW
EIRP on beam axis (minimum)		52.1 dBW		45.9 dBW
Signal noise bandwidth	B_N	100 MHz		10 MHz
Path length 38 000 km	L	38 000 km		38 000 km
Burst bit rate in clear sky	$R_{b\,max}$	360 Mbps		22.5 Mbps
Burst bit rate, minimum CNR	$R_{b\,min}$	66.7 Mbps		5 Mbps
Path loss 38 000 km		209.8 dB		213.3 dB
User terminal antenna gain, receiving, on axis	G_r	42.4 dB		
Satellite antenna gain, receiving			G_r	51.0 dB
Pointing loss, receive antenna		0.5 dB		0.5 dB
Atmospheric clear sky loss		0.7 dB		0.7 dB
−1 dB spot beam contour loss		1.0 dB		1.0 dB
Miscellaneous losses		0.5 dB		0.5 dB
Received power (Max EIRP)	C	−109.4 dBW		−119.1 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz		−228.6 dBW/K/Hz
System noise temperature, clear sky	170 K	22.3 dBK	600 K	27.8 dBK
Receiver noise bandwidth	100 MHz	80.0 dB Hz	10 MHz	70.0 dB Hz
Noise power	N	−126.3 dBW		−130.8 dBW
CNR, clear sky, max EIRP spot beam		16.9 dB		11.7 dB
CNR, clear sky, min EIRP spot beam		8.3 dB		11.7 dB

from a maximum of 60.7 dBW to a minimum of 52.1 dBW, on axis. The CNR values in Table 11.2 correspond to a user terminal located on the −1 dB contour of a spot beam with EIRP 60.7 dBW for the maximum EIRP case and a spot beam with EIRP 52.1 dBW for the minimum EIRP case. The user terminal has a low noise amplifier (LNA) with a noise temperature of 120 K and a clear sky antenna temperature of 50 K. The maximum burst rate for the TDM downlink is 360 Mbps using 16-APSK modulation and 9/10 rate FEC in 125 MHz of transponder bandwidth ($\alpha = 0.25$). The uplink operates in a MF-TDMA frame with a burst rate of 22.5 Mbps using 8-PSK modulation and 3/4

rate FEC in 12.5 MHz of transponder bandwidth ($\alpha = 0.25$). A 1.8 dB implementation margin in the user receiver is applied in each case.

11.2.3 Link Performance With Rain Attenuation Between the Satellite and User Terminal

The CNR results in Table 11.2 are optimistic because they do not include the effect of interference from other communication systems in the same frequency bands, the sidelobes of adjacent spot beams at the same frequency, and overlaid beams in the orthogonal polarization. System designs must keep interference to a low level, but total interference C/I ratios of 22 dB are common. The problem may become worse with 12 000 NGSO satellites in LEO sharing the same frequency bands.

Figure 11.3 shows the spectral efficiency of DVB-S2 links for 16-ASPK, 8-PSK, and QPSK modulations for a number of FEC rates. The CNR values include a receiver implementation margin of 1.8 dB and represent the lowest CNR values for which the DVB-S2 quasi error free (QEF) objective of one bit error per hour is maintained. When rain in the slant path causes attenuation of the signal and an increase in sky noise temperature, ACM is used to change the modulation and FEC rate to maintain the QEF objective. In Example 11.1, the clear sky CNR for user terminals within the −1 dB contour of the highest EIRP beam had a clear sky CNR of 16.9 dB. 16-APSK modulation with a FEC rate of 9/10 requires a minimum CNR of 15.0 dB to maintain QEF operation, and can therefore be used when there is no additional attenuation in the slant path. When rain occurs on the downlink to the user terminal, the carrier power is reduced and sky noise temperature increases; the ACM will change the modulation and/or the FEC code rate to compensate for the lower CNR, but the spectral efficiency is reduced and the burst bit

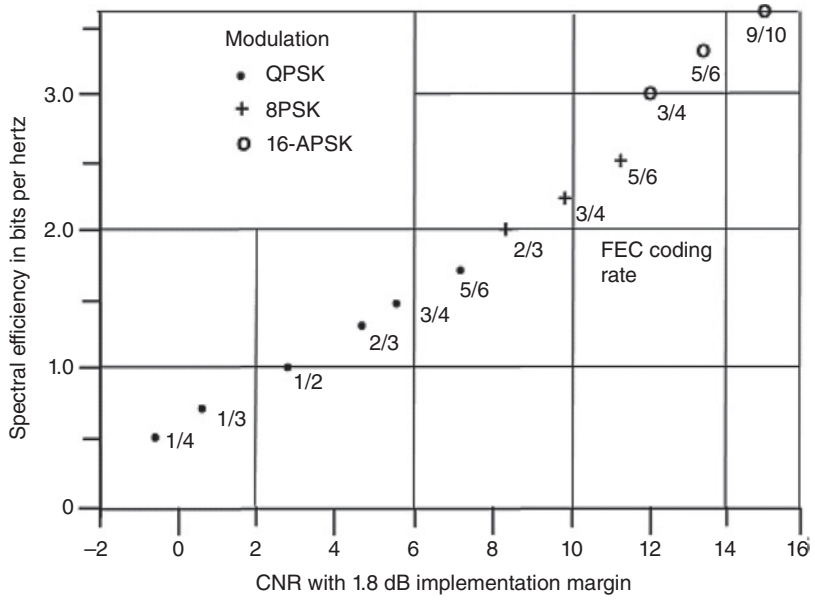


Figure 11.3 Spectral efficiency of DVB-S2 links for several combinations of modulation and FEC code rate. A 1.8 dB implementation margin is included in the CNR values.

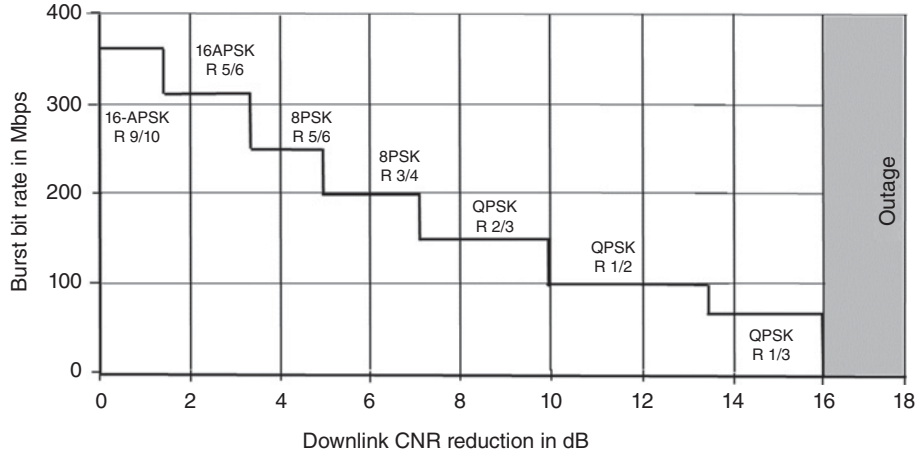


Figure 11.4 Adaptive coding and modulation applied to counter rain attenuation in the downlink from the satellite to the user terminal of Table 11.2 for a terminal on the −1 dB contour of the satellite spot beam. 16-APSK, 8-PSK, and QPSK are modulations. R 9/10, R 5/6, R 3/4, R 2/3, R 1/2, and R 1/3 are forward error correction code rates. The step values for CNR reduction occur when the rain attenuation causes the receiver CNR to reach a level 0.5 dB above the threshold for each ACM setting. The burst bit rate is reduced as each new modulation and code rate is applied to combat rain attenuation. An implementation of 1.8 dB is used in this example.

rate on the link falls. Figure 11.3 does not include all of the possible ACM settings envisaged by ETSI (ETSI 2009). How many ACM combinations are available on a given link depends on the implementation of the receiver application specific integrated circuit (ASIC) and the design of the ACM system.

Figure 11.3 can be used to analyze the performance of the link in Example 11.1 when rain affects the downlink from the satellite to the user terminal, and the uplink from the user terminal to the satellite. This leads to a step-wise succession of ACM changes as the rain attenuation increases, as illustrated in Figure 11.4 for the 19.5 GHz downlink of Table 11.2. In practice, many more modcon steps can be used than are shown in Figure 11.4.

11.2.3.1 Downlink Rain Attenuation for Link in Table 11.2

For terminals in a spot beam that has the maximum EIRP, the CNR within the −1 dB contour in clear sky conditions exceeds 16.9 dB and provides a 1.9 dB margin over the 15.0 dB threshold CNR for 16-APSK with 9/10 rate FEC. When rain attenuation affects the downlink, these terminals utilize the ACM technique of DVB-S2 signals to steadily decrease the FEC code rate. In Figure 11.4 a downward step is implemented when the downlink CNR reaches a level 0.5 dB above the threshold value for a given modulation and FEC rate. As rain starts to affect the downlink, the first step is made at a CNR of 15.5 dB, 0.5 dB above the threshold for 16-APSK with 5/6 FEC rate. The burst bit rate is reduced to 333 Mbps. Further decrease in CNR requires a change to 8-PSK modulation with 5/6 rate FEC when the CNR has fallen to 13.7 dB, 0.5 dB above the 13.2 dB threshold for this ACM combination. The next step occurs with 5.1 dB reduction in CNR down, with a change to 8-PSK with 3/4 rate FEC, and a burst bit rate on the link of 200 Mbps. Further decrease in downlink CNR requires more steps to maintain the QEF objective,

until eventually the limit for QPSK with rate 1/3 FEC is reached. This is at a CNR of 0 dB and the burst bit rate is 66.7 Mbps. Any further decrease in downlink CNR will cause the link to go into outage. The DVB-S2 standard includes more ACM combinations than are shown in Figures 11.3 and 11.4, where a large step in attenuation is chosen to illustrate the ACM process.

Calculating the rain attenuation on the downlink at which each of the above steps occurs is complicated by the increase in sky noise temperature that accompanies the rain attenuation. The iterative technique described in Chapter 4 must be used. For example, the first ACM step from 9/10 rate FEC to 5/6 rate FEC with 16-APSK modulation occurs when the CNR falls from 16.9 dB in clear sky conditions to 15.5 dB in rain, a CNR reduction of 1.4 dB. The system noise temperature of the user terminal receiver in Example 11.1 is 170 K, made up of 120 K for the LNA and 50 K from sky and antenna noise. With 0.7 dB of rain attenuation, the total path attenuation on the downlink is 1.4 dB (clear sky atmospheric attenuation of 0.7 dB plus rain attenuation of 0.7 dB) and the corresponding sky temperature is 74 K. The system noise temperature is 194 K and the increase in noise power is calculated as 10 log (194/170) dB or 0.6 dB. Adding the rain attenuation of 0.7 dB to the 0.6 dB increase in receiver noise power reduces the CNR by 1.3 dB, close enough to the required figure of 1.4 dB.

The process of finding rain attenuation and receiver noise increase that reduces downlink CNR to the step levels continues until the CNR reaches 0 dB when the rain attenuation is 12.2 dB, at which point the link will fail. The lowest burst bit rate is 66.7 Mbps, and the dynamic range over which rain attenuation in the downlink can be overcome is 12.2 dB. The contribution of noise power increase to CNR reduction slows down as attenuation increases. With infinite attenuation, the sky noise temperature would become the medium temperature, assumed to be 270 K in Example 11.1. The maximum feasible rain attenuation of 12.2 dB is sufficient to sustain operation of the 19 GHz downlink for a terminal in the mid-Atlantic region of the United States for 99.7% of an average year with an elevation angle of 20°, or a terminal in a south eastern state where the most frequent heavy rain occurs for 99.4% of an average year (Mitchell et al. 1997).

11.2.3.2 Uplink Rain Attenuation for Example 11.1

The uplink from the user terminal to the satellite operates in a shared MF-TDMA frame with a maximum bit rate of 22.5 Mbps in clear sky conditions using 10 MHz of transponder bandwidth with 8-PSK modulation and 3/4 rate FEC, with CNR of better than 11.7 dB for terminals inside the −1 dB contour of the satellite spot beam. This provides a CNR margin of 1.9 dB over the CNR threshold of 9.8 dB for 8-PSK modulation and 3/4 rate FEC. As rain attenuation affects the uplink, the ACM system will change the modulation to QPSK and steadily decrease the FEC rate to one half, where the CNR threshold is 2.8 dB, giving a CNR margin of 8.7 dB. (The uplink to the satellite is using DVB-RCS format signals with α value of 0.25 and an implementation margin of 1.8 dB.) The burst bit rate will reduce to 10 Mbps, and a further decrease in CNR will cause the modulation the change to binary phase shift keying (BPSK) with half rate FEC, which has a threshold at 0 dB, and a burst rate of 5 Mbps. The ACM process is illustrated in Figure 11.5 for the uplink of Example 11.1 with a relatively coarse step size; as in Figure 11.4, the step is made when rain attenuation causes the receiver CNR to reach a level 0.5 dB above the threshold for that ACM combination.

The dynamic range for CNR reduction, and for rain attenuation is 11.7 dB, giving an average yearly availability of 99% for a 29 GHz uplink in the heaviest rainfall zone in

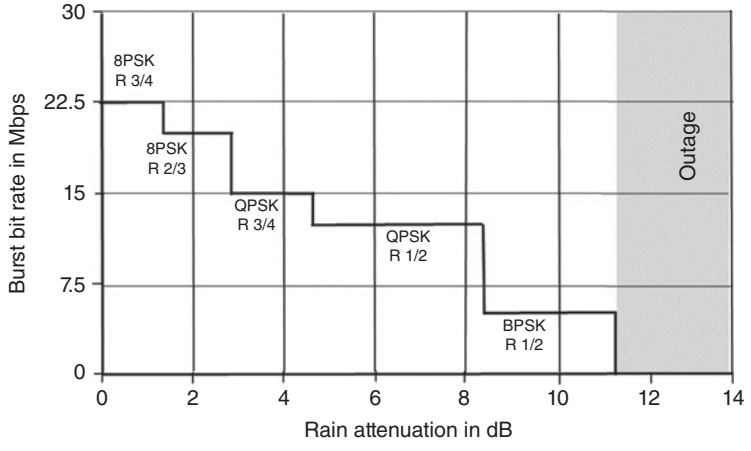


Figure 11.5 Adaptive coding and modulation applied to counter rain attenuation in the uplink from the satellite to the user terminal in Table 11.2. 8-PSK, QPSK, and BPSK are modulations. R 5/6, R 2/3, and R 1/2 are forward error correction code rates. The step values for rain attenuation occur when the rain attenuation causes the satellite receiver CNR to reach a level 0.5 dB above the threshold for each ACM setting in Figure 11.3. The burst bit rate is reduced as each new modulation and code rate is applied to combat rain attenuation. An implementation of 1.8 dB is used in this example.

the United States and 99.5% availability in mid-Atlantic states. There is no change to the system noise temperature of the satellite receiver when rain affects an individual terminal because the spot beam covers a very wide region of the earth's surface, so the reduction in receiver CNR is equal to the rain attenuation on the uplink.

11.2.3.3 Performance of Terminals in Other Locations in Table 11.2

Terminals lying in the maximum EIRP spot beam between EIRP contours of −1 and −3 dB below the maximum have CNR values up to 2 dB below those quoted in Table 11.2. The downlinks at the edge of the spot beam can use 8-PSK with 5/6 rate FEC with a clear sky margin of 3.0 dB and a correspondingly lower burst bit rate of 250 Mbps. Dynamic range for rain attenuation is reduced to 11.2 dB. The uplinks at the edge of the spot beam have a dynamic range of 9.7 dB. These terminals will suffer slightly higher outage durations than the terminals inside the −1 dB contour. A larger antenna can be used to improve performance where the spot beam has a lower EIRP.

User terminals located in spot beams with lower EIRP are where heavy rainfall is less frequent. The downlink CNR of 6.3 dB and the uplink CNR of 5.7 dB support QPSK modulation with 2/3 rate FEC, which has a threshold of 4.9 dB, giving a clear sky CNR margin of 1.4 dB for the downlink and 0.8 dB for the uplink. At the edge of the low EIRP spot beam, the terminals will use QPSK with half rate FEC. Burst bit rates will be correspondingly lower for these terminals, and the dynamic range for CNR reduction for a terminal at the edge of the spot beam is limited to 6.3 dB on downlinks and 5.7 dB on uplinks. Availability should be better than 99% of an average year for terminals in the northern half of the United States and Canada, and the west coast of the United States.

No account has been taken of the uplink and downlink CNR for the satellite to gateway links in the link budget in Table 11.2. It is assumed that the CNR on the gateway to

satellite and satellite to gateway links have CNRs at least 20 dB higher than the links from the satellite to the user terminal. Ka-band gateway antennas are typically 7 m diameter Cassegrain antennas with gain around 61.5 dB at 19.5 GHz and 65 dB at 29 GHz, ensuring the required high CNR is achieved.

The reduction in burst bit rate when rain attenuation affects the links will result in the user throughput being reduced when the links are busy. If the links are lightly loaded, users affected by heavy rain may be able to receive and send longer packets that help to maintain throughput. The virtual circuits established between the gateway and the user allow individual packets to have different modulation and coding, which can be changed by request from a user who is experiencing CNR reduction on the downlink, or by the gateway station when there is increased bit error rate on the uplink.

The heaviest part of a rain fade is usually quite short. Figure 11.6 shows some recordings of rain fades and scintillation made on a 30 GHz link from a beacon on the ATS-6 satellite to a receiving station at the University of Birmingham, UK. The signal was attenuated by more than 6 dB for seven minutes in the first example and four minutes in the second example. With 6 dB attenuation in the downlink at 19.5 GHz, the CNR reduction is 8.8 dB because the system noise temperature of the receiver increases by 2.8 dB. There is no change in the uplink system noise temperature. The ACM system will revert to a

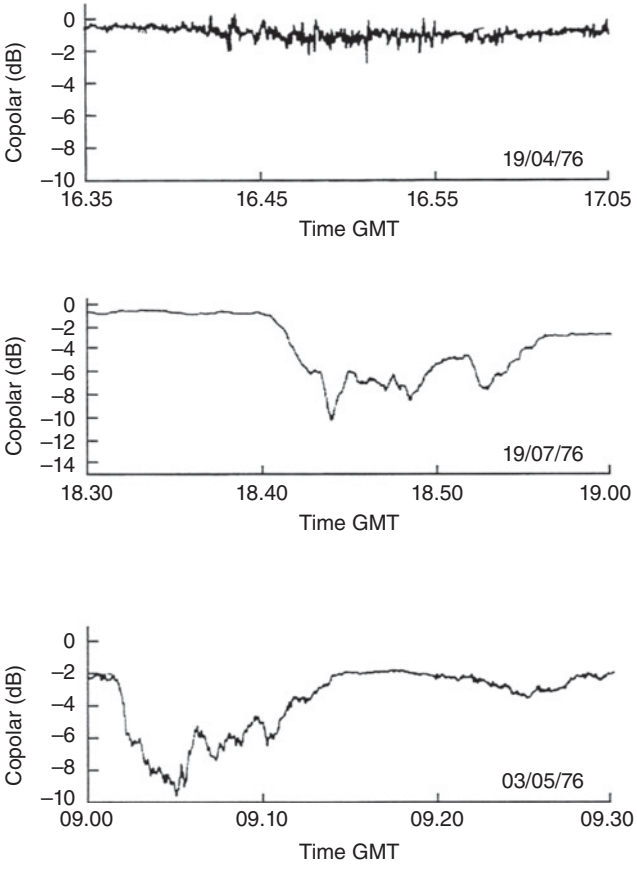


Figure 11.6 Examples of scintillation and rain attenuation on a 30 GHz slant path from the ATS-6 satellite, recorded at the University of Birmingham, UK. Scintillation reached a maximum of ±1 dB in heavy cloud. The rain fade events lasted less than 15 minutes and attenuation exceeded 6 dB for less than 7 minutes.

lower setting, reducing the burst bit rate for the affected terminal by a factor around two when there is 6 dB rain attenuation in the link. If only a few terminals in a spot beam are affected by heavy rain, the system can double the duration of their packets to maintain the same throughput as in clear sky conditions, with a corresponding slight reduction in the duration of packets to and from other users. Figure 11.6 shows a scintillation event in which clouds were present in the slant path and scintillation amplitude reached ±1 dB at 30 GHz. Scintillation is caused by focusing and defocusing of the signal along the slant path with the result that generally there is symmetry about the average signal level. The scintillation rate is sufficiently slow that the ACM system can compensate for changes in signal amplitude that exceed 0.5 dB. Many recordings of propagation events can be found in the literature showing examples of rain attenuation and scintillation on satellite paths at frequencies from 1.5 through 50 GHz (Propagation events 1997; Ippolito 2017).

Some GSO broadband access providers offer professional terminals with larger antennas and higher transmit power, up to 4 W, for the return channel uplink. The larger antenna provides an increase in download burst bit rate and the higher uplink transmit power provides faster uplink speeds and greater protection against rain fades.

11.2.4 European Broadband Satellite Systems

In 2018, Eutelsat offered internet access throughout Europe from the KA-SAT GSO satellite, a Ka-band satellite similar to ViaSat 1 with a capacity of 90 Gbps and 82 spot beams (KA-SAT 2018). Eutelsat is a public company with headquarters in Paris, France, set up in 1977 as an intergovernmental organization (IGO) to develop and operate a satellite-based telecommunications infrastructure for Europe. It is a European equivalent of Intelsat and provides satellite communications services to over 40 countries in Europe and North Africa (Eutelsat 2018). The internet access service is called TooWay™ and offers speeds up to 20 Mbps download and 6 Mbps upload. Customers are limited to downloading 2 GB per month at the lowest tier of service, up to 100 GB at the highest level (TooWay 2018). TooWay terminals use a 0.77 m (30 in.) dish.

11.2.5 Avoiding Oversubscription

Problems of oversubscription in all internet access systems except fiber optic services resulted in action by the Federal Communications Commission (FCC) in the United States, and its counterpart OfCom in the UK, to monitor the performance of companies providing internet access. In 2010, the FCC created the Measuring Broadband America program to accurately measure America's fixed and mobile internet services (Measuring Broadband America 2011). Between 2011 and 2016, the FCC published an annual report, which contained data gathered from measurements made at the homes of 6800 volunteers with connections to 13 ISPs. The results covered fiber to the home (FTTH), cable broadband, mobile long-term-evolution (LTE) and (after 2013) satellite broadband delivery of internet access services. In 2016, the last time a full report on the ISPs was published by the FCC, the commission concluded that satellite internet access systems offered by ViaSat and HughesNet had met their advertised uplink and downlink speeds more than 80% of the time (Measuring Broadband America 2016). Similar data was not published in 2017 or 2018 after the Trump administration took over from the Obama administration and the chairmanship and membership of the commission changed.

The United States Federal Communication Commission has five members drawn from congressional representatives, a chairperson appointed by the president, and a large staff based in Washington, DC. There are three representatives from the party that has the majority in Congress and two from the opposition. In the period 2010 through 2016, the Democratic Party held the majority on the commission and President Obama appointed the chairman. Beginning in 2017, the Republican Party held the majority in Congress and President Trump appointed the chairman. In the 1996 Telecommunications Act, Congress charged the FCC with the task of reporting on the *Deployment of Advanced Telecommunications Capability to All Americans in a Reasonable and Timely Fashion*. The 2018 report of the FCC did not include detailed performance results from the 6800 volunteers who had previously been the basis for earlier FCC reports; instead speed results were based on speed test data acquired via Ookla, the company behind speedtest.net. The two Democratic Party members of the commission attached dissenting opinions to the 2018 report, disagreeing with its findings (Measuring Broadband America 2018).

11.2.6 GSO Broadband Satellite System Capacity

There are a limited number of geostationary orbit locations for internet access satellites. The small antennas of user terminals have relatively wide beams, requiring spacing of satellites sharing the same frequency band to be at least 2° to keep interference from adjacent satellites to an acceptable level. Once the available orbital locations are fully occupied, the only remaining option for GSO satellites is to move up to a higher frequency band. The next available band is at the top edge of Ka-band, where the frequency band 38.0–42.0 GHz is allocated to satellite uplinks and V-band, which covers 40–75 GHz. Satellite downlinks in V-band are available in the frequency bands 42.5–43.5 GHz and 47.2–50.2 GHz. (Some sources identify these frequency ranges as Q-band, and there is some disagreement about where Ka-band ends and V-band begins.) Rain attenuation increases steadily with frequency, and absorption by the oxygen molecules in the air starts to become important between 50 and 65 GHz. At 50 GHz attenuation due to atmospheric oxygen at sea level is 0.5 dB/km, rising rapidly to 15 dB/km at 60 GHz. (Figure 7.11 in Chapter 7 plots atmospheric attenuation for oxygen and water vapor over a wide range of frequencies.)

A large GSO satellite making use of the full 4 GHz of bandwidth available for uplinks and downlinks between 38.0 and 50.2 GHz, combined with 50-fold frequency re-use from hundreds of multiple beams and two polarizations, could potentially have an effective bandwidth of 400 GHz. At 4 bits/Hz spectral efficiency using 16 and 32 APSK modulation and minimal FEC in clear sky conditions, the GSO satellite could achieve a capacity of 1600 Gbps. With satellites spaced two degrees apart in GSO from 50°W to 150°W, North America could be served by GSO satellites in Ka-band and V-band providing a total capacity in excess of 100 Tbps. Half of the capacity must be devoted to uplinks and half to downlinks, and then split between outbound links and inbound links. This represents an upper limit on the ability of GSO satellites to provide internet access to the North America continent. By comparison, optical fiber lines to individual homes (FTTH) can provide bit rates close to 1 Gbps, so all the GSO satellites put together can serve only the equivalent of 120 000 homes in North America served by optical fiber at

1 Gbps (2018 data). In 2018 Verizon claimed over 300 000 subscribers to its FioS optical fiber service (Verizon 2018). Satellite internet access is best used to provide service to people living in areas where fiber optic lines and cable TV are unlikely to be available in the future.

If you live in a country without a well-developed telephone or cable TV system, you might have no way to access the internet. About 3 billion people fall into the latter category. O3B Networks Ltd., a wholly owned subsidiary of SES S.A., the major provider of satellite television and satellite data services in Europe and parts of Asia, was formed to create internet access to the Other 3 Billion (O3B 2018). O3B operates 16 Ka-band satellites in medium earth orbit (MEO), providing voice and internet service to ISPs. The system is not able to deliver internet access to individual homes as two 2.4 m dish antennas are required for continuous communication, so a community access system using WiFi or cellular telephone technology is needed. Other internet access systems using constellations with thousands of non-geostationary (NGSO) satellites have been proposed; these are the subject of the following sections.

11.3 NGSO Satellite Systems

11.3.1 Teledesic

Teledesic was founded in 1994 by Bill Gates, the founder of Microsoft, and Craig McGraw, who had made a fortune from a cellular phone company he had created. At that time the internet was just getting started, access was typically by a telephone line modem, and download rates of 25 Mbps were mere dreams. The concept of Teledesic was to wire the world with broadband access from space using a constellation of 840 Ka-band LEO satellites, because the cost of connecting everyone in the world by optical fiber was too great – and still is. The project was estimated to cost US$9B and attracted an initial US$1B from private investors. However, although the concept of Teledesic was sound, as demonstrated by the construction of similar NGSO satellite internet access systems 25 years later, the project was too far ahead of its time (Teledesic 2018).

In 1994 there were no cubesats, launching a satellite cost tens of millions of dollars, and satellite construction was limited to companies that wanted to build a small number of high cost units rather than hundreds or thousands of low cost satellites. Phased array antennas for customer terminals at affordable prices were 20 years away. In 1997 the number of satellites was scaled back to 288 in an attempt to make the system realizable for US$9B, but the dot.com bubble burst in 2001 making it impossible for Teledesic to raise any further capital. The company closed in 2002 with cash in the bank, and returned money to some of its investors. Had the company gone public, it would have had to file for bankruptcy and the investors would have received very little (Griffin 2016).

Teledesic was a brilliant idea, but too far ahead of its time, as the technology to build the system at an acceptable cost was not available in 2000. By 2016, that had changed, and a large number of Teledesic-like systems were proposed. In December 2017 in an address to the British Royal Aeronautical Society, Sir Martin Sweeting, Chairman of Surrey Satellite Technology Ltd., a pioneer of low cost satellite systems, stated that there were more than 150 proposals worldwide for LEO and MEO internet access satellite systems with a total of over 23 000 satellites. Not all of these systems will be built (Sweeting 2017).

In 2016, the US Federal Communication Commission invited proposals for NGSO Broadband satellite systems. The FCC received 11 proposals, and 2 of these systems were

granted licenses to proceed: OneWeb and SpaceX satellite constellation. As mentioned earlier, O3B had already created a LEO constellation of satellites to serve countries that lacked a good terrestrial communication system. Similar systems proposed by Space Norway with 117 LEO satellites and Telesat Canada with 2 elliptical orbit satellites have received initial FCC approval (Space Norway 2017).

All of the NGSO systems have constellations of satellites that can provide access to the internet from anywhere in the world, including the oceans and in the air. Phased array antennas can be mounted on vehicles, aircraft, and ships to track the satellites while the vehicle, ship, or aircraft is in motion. OneWeb's constellation covers polar regions because it has satellites in near polar orbit, and five of the SpaceX satellites are in a 1275 km orbit with an inclination of 81° to provide coverage of both poles (OneWeb FCC filing 2017). All of the NGSO systems have the ability to earn income by selling access to the internet to anyone, anywhere, who can afford the monthly fee. In countries with low GDP per capita, the monthly fee can be made correspondingly lower than in wealthier countries, or even reduced to zero if the satellites are out of view by anyone who can afford to pay for the service. The potential population that can be served by an NGSO system is 7.6 billion (May 2018 estimate of the world's population). This is a very different economic model from the GSO satellite systems that serve North America, where the population is 360 million (World Population 2018). The potential earnings of the NGSO systems makes it possible for their providers to consider spending US$12B to establish the complete system. As should be expected, both OneWeb and SpaceX propose to phase in the service by launching satellites over a period of years, and use the revenue from the early satellites to pay to establish the rest of the constellation. SpaceX hopes to earn US$30B from their satellite system by 2025 (Time 2018).

Figure 11.7 illustrates the concept of a NGSO satellite internet access system. The NGSO satellites cross the sky in a few minutes and must be tracked by both the user's antenna and the gateway station's antennas. Flat panel phased arrays pointing at the zenith are the preferred antenna for NGSO user terminals. The NGSO satellites have phased array antennas producing multiple spot beams that point toward the user's

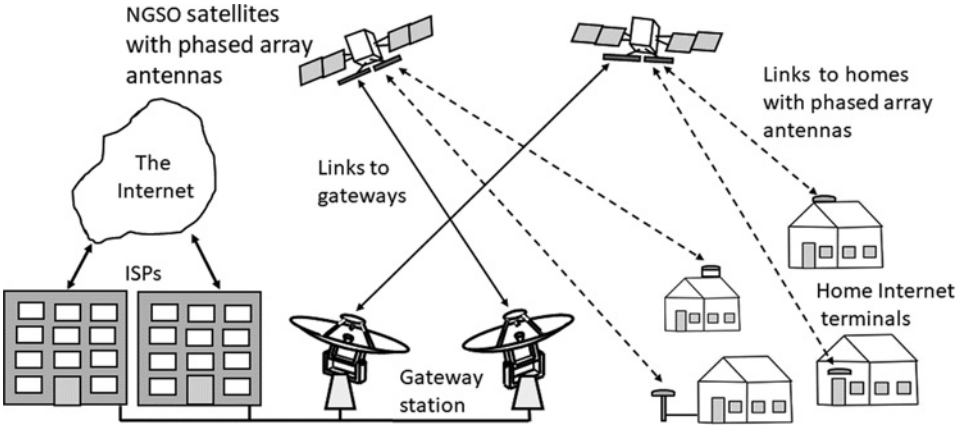


Figure 11.7 Illustration of a NGSO satellite internet access system. The gateway station has many steerable antennas that track the NGSO satellites. The satellites have two phased array antennas: one tracks the gateway stations, the other tracks the user terminals. Note that mounting a flat panel phased array antenna on a roof top is undesirable in regions where snowfall is common in winter. Melting snow on the upward facing antenna will cause outages.

location for a short duration to download and upload data packets. In very low earth orbit (VLEO) systems using V-band where the satellites are as low as 350 km altitude, the spot beams may cover as little as a 9 km circle on the earth's surface. The gateway stations need many tracking antennas to follow NGSO satellites as they fly across the sky. A system with five satellites visible at one time needs at least 12 antennas, 5 to track satellites, 5 to reset to the next satellite position, and at least 2 spares for maintenance and repair. The gateway antennas for V-band VLEO satellites are in the 1.8–2.6 m diameter range and are on X-Y mounts so that satellites can be tracked through zenith (see Figure 11.11).

11.3.2 OneWeb

OneWeb's system proposes 882 satellites in LEO in 18 orbital planes at 1200 km altitude, with links to user terminals in Ku-band, and links between the satellite and gateway stations in Ka-band. Additional satellites are proposed in MEO at an altitude of 8000 km. The estimated cost of the NGSO system is US$3B by the time the full constellation becomes operational (OneWeb 2017). OneWeb was founded by Greg Wyler in 2014 and their NGSO broadband satellite system was originally known as WorldVu. The company's business plan is to reach hundreds of millions of potential users residing in places without (existing) broadband access (OneWeb satellites 2018). The satellites are expected to have a mass of 175–200 kg with a communication capacity of 50 Gbps, and the company was considering adding 1972 satellites to the constellation.

In March 2017, OneWeb filed plans with the FCC to field a constellation of an additional 2000 V-band NGSO satellites. OneWeb's satellites at 1200 km altitude are in near-polar orbit with an inclination around 87.9°. An example frequency plan for OneWeb's V-band system is shown in Table 11.3. Dual circular polarization frequency reuse is employed in all frequency bands. The satellites have on board processing that allows modulation and FEC rates to be different on each of the links, resulting in different bandwidth requirements.

11.3.3 SpaceX Satellite Constellation

In March 2017, SpaceX filed with the FCC plans to launch a constellation of more than 7500 satellites in non-GSO synchronous orbits, known informally as *Starlink*. The planned system has 7518 satellites in VLEO at 340 km altitude using V-band

Table 11.3 OneWeb frequency plan for V-band NGSO constellation

Link	Frequency range	Number of beams	Spatial reuse	Effective bandwidth
Satellite to user terminal	40.0–42.0 GHz	20	5	20.0 GHz
User terminal to satellite	48.2–50.2 GHz	20	5	10.0 GHz
Gateway to satellite	42.5–43.5 GHz 47.2–50.2 GHz 50.4–51.4 GHz	1	1	10.0 GHz
Satellite to Gateway	37.5–40.0 GHz 40.0–42.5 GHz	1	1	10.0 GHz

Table 11.4 SpaceX LEO Ku-/Ka-band satellite constellation with 4425 satellites

Satellite altitude	1150 km	1110 km	1130 km	1275 km	1325 km
Orbital planes	32	32	8	5	6
Satellites per plane	50	50	50	75	75
Inclination	53.0°	53.8°	74.0°	81.0°	70.0°

(40–50 GHz) and 4425 Ku- and Ka-band satellites at 1200 km altitude. SpaceX was founded by Elon Musk, who had previously developed the Tesla electric car, and had also created the Falcon launch vehicle that significantly lowered the cost of launching satellites, an essential component in the creation of two constellations of nearly 12 000 satellites. The SpaceX satellites are expected to have a mass of 384 kg and to be located at multiple altitudes and orbital inclinations, as indicated in Table 11.4. Initial launch is planned to be 1600 satellites with 2825 to follow over several years (SpaceX FCC Ku-Ka-band filing 2016).

11.3.4 NGSO Satellite System Parameters

Several new technological and manufacturing methods had to come together to make the new NGSO proposals successful where Teledesic had failed. To be profitable, the NGSO constellations require a very large number of satellites capable of gigabit communication speeds, and a worldwide market with millions of customers. The satellites have to be built for a fraction of the cost of a large GSO satellite, US$0.5M instead of US$100M, and launch costs to LEO or MEO have to be less than US$1.3 per kg. Ka- and V-band phased array antennas for customer earth terminals are essential to track fast moving LEO satellites across the sky and must be able to switch almost instantaneously between satellites, and still be affordable. Target price for an earth terminal phased array antenna is US$200, although such antennas were costing US$1000 in 2018 (Phased array antennas 2018). It is anticipated that the cost will go down with the manufacture of the very large number of antennas needed worldwide for the OneWeb, SpaceX, and similar broadband NGSO systems.

A constellation of 12 000 satellites costing US$0.5M each with a launch cost of US$0.5M per satellite requires an investment of US$12B. How can these projects succeed where Teledesic failed? Large GSO satellites are very expensive because they must be extremely reliable to operate continuously in space for 15 years. All components must be space qualified, the satellites require complex attitude and orbital control systems because of the low earth gravity at GSO altitude, and must carry enough fuel for injection into GSO and orbital maneuvers over its lifetime. With a typical mass of 6000 kg, launching a large GSO satellite is always expensive. By contrast, between Teledesic's beginnings in 1994 and 2016 when large constellations of LEO satellites became feasible, many small satellites designed for shorter lifetimes, and in particular cubesats, had demonstrated that it was possible to build and launch medium size satellites at much lower cost than traditional GSO communication satellites. Earth's gravity at LEO altitudes is much stronger than at GSO making stabilization of the satellite much easier. With thousands of LEO satellites available, failure of 10 or even a 100 satellites is not the catastrophic disaster of a single large GSO satellite failure. Spare satellites can be kept

in orbit and launching more satellites is far less expensive than launching a replacement GSO satellite.

In the following discussion typical parameters of the proposed NGSO constellations are presented, based on material published in FCC applications and news releases. At the time of writing (2018), only two test satellites had been launched by SpaceX (Microsats A and B 2018). Some of the system parameters are likely to change as more satellites are launched and the NGSO constellations come into use.

The large constellations of LEO satellites proposed by OneWeb and SpaceX ensure that several satellites are visible to users at any time. For example, a constellation of 7500 satellites can have 50 orbital planes with 75 satellites in each plane. Satellites are less than 500 km apart and even at a low altitude of 350 km, five satellites are in view at all times above an elevation angle of 35°. Similar conditions apply to the OneWeb system, with five satellites visible above an elevation angle of 45°. The minimum elevation angle is important because phased array antennas cannot scan from horizon to horizon. A maximum scan angle of ±55° is a practical limit for an array with element spacing of one half wavelength because *grating lobes* can start to appear in the antenna pattern when the scan angle exceeds 60°, and the gain of the antenna falls off quickly at large scan angles. Path length to the satellite varies by a factor less than 1.3, keeping variation in path loss below 2.3 dB, and with a tracking phased array on the satellite transmit power on the downlink can be increased to compensate for the extra path loss as the satellite transmit beam scans away from nadir, and also to compensate for loss of antenna gain when a phased array antenna is scanned to a large angle off axis.

Latency is a major issue in satellite internet access systems. The Viasat and HughesNet GSO systems have an average round trip delay between 600 and 670 ms, according to FCC reports in 2016 (Measuring Broadband America 2016). The delay is caused by the long path length to a GSO satellite, typically 38 000 km, causing a round trip delay of 500 ms, and buffering and processing of the data packets can add another 100 ms. As discussed in Chapter 8, delays in excess of 60 ms cause the TCP/IP internet access protocol to time out, so GSO satellites used for internet access must employ protocol conversion. The round trip delay for a satellite at 1200 km altitude at an elevation angle of 45° is 17.2 ms, and for a VLEO satellite at an elevation angle of 35° the round trip delay is 6.1 ms. This allows transfers of data in TCP/IP format without protocol conversion. However, TCP/IP acceleration is needed to ensure that the data does not make multiple round trips, as can happen on terrestrial circuits. The LEO satellite system behaves in a similar way to terrestrial internet access circuits and enables applications such as gaming that require rapid response, which are not feasible over GSO satellites. Tests with two experimental satellites launched by SpaceX in March 2018 indicated latency of 25 ms (Microsats A and B 2018).

11.3.5 Phased Array Earth Terminal Antennas

The phased array antenna proposed for the OneWeb Ku-band system for user terminals has dimensions 0.36 m by 0.16 m. At an uplink frequency of 18 GHz with a wavelength of 0.0167 m, the array is 21.5 by 9.6 wavelengths, and at 12 GHz downlink frequency where the wavelength is 0.025 m, the array is 14.4 by 6.4 wavelengths. Element spacing in the array should not exceed one half wavelength to ensure that grating lobes are not produced, requiring 817 transmit elements and 377 receive elements. It is possible to reduce the number of elements in a transmitting phased array by a process called *thinning*, so

fewer than 817 transmit elements could be used. See Appendix B for more information on phased array antennas. Assuming an aperture efficiency of 60%, the transmit gain of the antenna is approximately 34 dB and the receive gain 30.6 dB, comparable to the gain of a DBS-TV receiving antenna at Ku-band. Antenna 3 dB beamwidths should be approximately 3.5° by 8° when transmitting and 5.3° by 12° when receiving. Typical uplink transmit power is 1 W (0 dBW), which can be generated by solid state elements in the transmit array with power levels of 1–100 mW. Transmitted power has to be greatest in the center of the array and tapered toward the edges to control sidelobes, as there are stringent rules governing interference into other satellite systems, both GSO and NGSO. The uplinks employ MF-TDMA burst transmissions, so the average transmit power for any element is much less than the burst transmit power. High voltages and high temperatures are major contributing factors in the failure of microwave transmitters, both of which are present in traveling wave tube amplifiers (TWTAs). Solid state devices transmitting a few milliwatts of power do not require high voltages and do not dissipate a lot of heat, and therefore have better reliability than a TWTA. Failure of a few transmitting elements in a phased array does not cause complete failure of the link, making the phased array antenna more fault tolerant.

If we assume a similar design approach for the earth terminal phased array antennas in the 40–50 GHz spectrum of V-band, the antenna dimensions reduce to 0.13 m by 0.06 m. Separate transmit and receive arrays are often used in phased array antennas to simplify the problem of making wide band elements that can operate well at the separate transmit and receive frequencies. For a V-band user terminal, the antenna might have a square profile 0.25 m on a side.

Communication with NGSO satellites is limited to the time they are within the *cone of visibility*, a cone centered on the user terminal with a half angle defined by the minimum permitted elevation angle of the user terminal antenna beam, as illustrated in Figure 11.8 for a minimum elevation angle of 45°. Satellites passing through the vertical axis of the cone of visibility (an overhead pass) are visible for the longest time, whereas satellites passing through the edge of the cone of visibility are visible for a very short

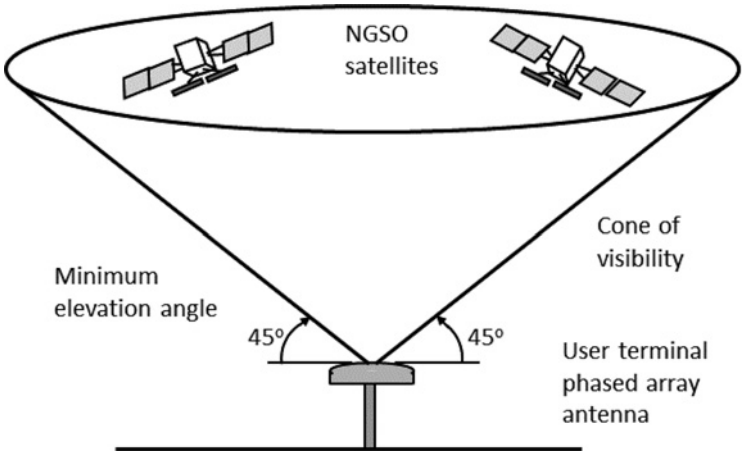


Figure 11.8 Illustration of NGSO link cone of visibility. The minimum permitted elevation angle for the user phased array antenna is 45°, corresponding to a scan angle for the phased array of ±45° in this illustration.

time. However, because the phased array antennas have relatively broad beams, satellites do not disappear from view (which means losing contact) at the edge of the visibility zone, they just have a larger pointing loss. The large number of satellites in the NGSO constellations ensures that at least one satellite is always visible within a visibility cone with a half angle no greater than 45 degrees. This allows the minimum elevation angle of the earth terminal phased array antenna beam to be 45 degrees and ensures that grating lobes do not appear when elements are spaced by a half a wavelength or a slightly greater distance. The antenna axis is vertical and the beam can be scanned to 45 degrees from vertical, so elevation angles always lie between 45 degrees and 90 degrees.

11.3.6 Satellite Phased Array Antennas

The phased array antennas on NGSO satellites are much more complex than those at the earth terminals. The antennas must produce multiple beams that track individual earth terminals as the satellite flies through each station's visibility cone, assumed to have a 45° half angle in the following analysis. VLEO satellites at an altitude of 350 km have a maximum visibility time of 134 seconds (2 minutes 14 seconds) for an overhead pass. Satellites at 1200 km are visible for a maximum of 322 seconds (5 minutes 22 seconds) with an overhead pass. Downlink transmissions use TDM of packets destined for the many user terminals visible to the satellite at any instant of time. The phased array antenna on the satellite switches its downlink beams to the direction of the user for the duration of that terminal's packets, and then switches to another user's direction to repeat the process. The user terminal's phased array antenna tracks a satellite across the sky until a control packet in the downlink data stream instructs the terminal to point its beam to a different satellite. The beam can be steered to new pointing angles in a few milliseconds, so by the time the next satellite starts transmitting to a given terminal, the antenna beams are already tracking that satellite. The satellite requires a second phased array to direct beams toward the gateway station, which must also be tracked as the satellite progresses in its orbit. Optical intersatellites links are used in all low altitude NGSO systems to avoid the need for a very large number of gateway stations. When no gateway station is visible to a satellite, downlink data is sent to an adjacent satellite over the optical link. With 40 satellites in 18 orbital planes, for example, spacing between satellites is 1320 km or less, and transmission time over the optical link does not exceed 4.4 ms. When a satellite is over an ocean, several optical links in series may be needed to reach a satellite that can communicate with a gateway, making latency somewhat longer than the 25 ms quoted for satellites communicating directly with a gateway station. However, this situation applies to only a small number of users on ships or aircraft.

The footprint (the portion of the earth's surface visible to the satellite) of a satellite at 1200 km altitude for earth stations that are limited to elevation angles above 40° is 2120 km (1324 miles) wide and covers a ground area of 4.5 million km^2. The footprint of the V-band satellites is 870 km (543 miles) wide and covers a ground area of 600 000 km^2. The satellite beams are typically 1.5° wide, covering an area of 550 km^2 at Ku-band and 52 km^2 at V-band (SpaceX FCC V-band filing 2016). The V-band beam is notionally circular and has a diameter of 9 km at the earth's surface. This is an extremely small area that can be selected by the satellite. For example, a single city like Roanoke, Virginia, with a population of 100 000 can be selected by a 9 km beam, and Montgomery County, the location of Blacksburg and Virginia Tech requires nine beams to cover the county. By comparison, typical GSO spot beams are 1000 km wide and may cover an entire state.

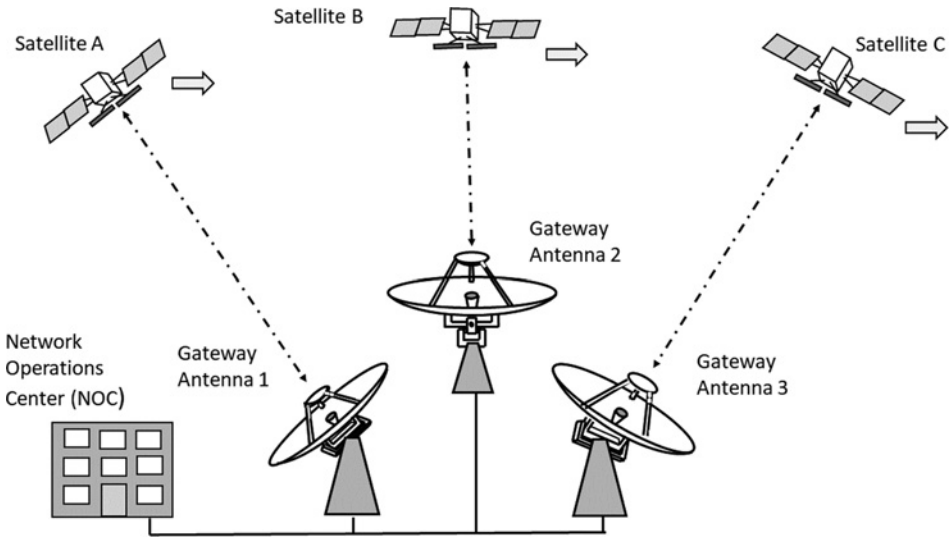


Figure 11.9 Illustration of a gateway station with three NGSO satellites visible to three earth station antennas. Satellite C is about to exit the cone of visibility, and satellite A is about to enter the cone. Communication between satellite A and gateway antenna 1 is established before satellite C leaves the cone of visibility. The network control center sends commands to satellites A and C and antennas 1 and 3 to switch communications from satellite C to satellite A.

Many of the satellites in the NGSO constellation will be silent for a substantial part of each orbit because the satellites are over an ocean, or one of the earth's polar regions with a polar orbiting system. These periods are used to recharge the onboard batteries. In general, polar orbiting satellites are in sunlight for a longer period of each orbit than satellites in inclined orbit, which spend roughly 50% of their time in the earth's shadow. In NGSO constellations, 60% of the electrical power generated by the solar arrays may be devoted to charging the batteries.

Satellites communicate with the user terminals only within the cone of visibility defined in Figure 11.8. Figure 11.9 illustrates a gateway station with three steerable antennas communicating with three NGSO satellites.

The gateway antennas are all fully steerable reflector antennas on X-Y mounts that can track satellites through zenith. (A conventional Az-El mounted antenna has to rotate 180° in azimuth as a satellite passes through zenith and cannot be used.) The gateway antennas can start communications with satellites before they enter the cone of visibility, at an elevation angle below the minimum allowed for the phased array user terminals. In the illustration in Figure 11.9, the network control center must send commands to satellites C and A, through antennas 3 and 1 to switch communications from satellite C to satellite A before satellite C exits the cone of visibility.

The 1 dB beamwidth of the phased array antennas at a user's terminal is typically 1.8°. The maximum time for which a VLEO satellite is within the visibility cone's full angle of 90° degrees is 134 seconds. The minimum time between beam repointing to keep the satellite within the 1 dB beamwidth of the terminal's beam is 1.2 seconds when the satellite passes overhead, and more frequent repointing is desirable to avoid beam pointing losses. The beam of a phased array antenna can be repointed in microseconds, so accurate tracking depends only on how often the antenna receives angle updates.

The operation of a broadband NGSO satellite system is quite different from a GSO internet access system, which has a user terminal with a fixed pointing antenna and constant beam pointing loss of less than 0.5 dB provided the antenna is installed correctly. In the NGSO system, the user terminal's antenna tracks a LEO satellite for anywhere between a few minutes and a few seconds and then switches to a different satellite. Losses caused by finite steps in beam pointing direction can result in a reduction in signal power, and variation in path length can add another 2.2 dB. The additional path loss for a terminal at the edge of the satellite's coverage zone can be compensated for by increasing the power transmitted by the satellite to terminals at the edge of its coverage zone, but this cannot be done so easily with the uplink from the terminal to the satellite, which must usually operate at full power all the time. The additional losses must be built into the uplink and downlink power budgets when assessing link margins.

The downlink phased array antenna on the NGSO satellite that connects to user terminals is connected to a large number of transponders covering the frequency bands used by the satellite. For example, if the satellite downlinks use 4 GHz of V-band bandwidth with 20 beams, each of which transmits at 1 W in a 200 MHz bandwidth, the array transmits a total of 20 W. Because the downlink uses TDM, each of the 20 beams connects to only one terminal at any time, so transmitted power cannot exceed 20 W. A phased array producing a 1.5° spot beam needs 10 000 elements spaced approximately one half wavelength apart, which requires an average of 2 mW per element. Instead of using 10 000 transmitting elements, transmitting phased arrays are often built as a collection of blocks several wavelengths on a side, with a single RF amplifier driving the elements in the block via a distribution network. Because the transmit power of blocks is greatest in the center of the array and tapers toward its edges, the central blocks might each transmit 100 mW, and the blocks at the edge of the array might transmit 10 mW. The RF amplifiers are generating much less power than the TWTA power amplifiers on GSO satellites, which are rated at hundreds of watts.

11.3.7 Avoiding Interference With GSO Satellites and Terrestrial Users

Many of the frequency bands allocated to NGSO internet access systems are shared by other satellites systems in GSO, and also by terrestrial microwave communication links. The most serious potential interference arises when a NGSO satellite passes across the GSO orbit, as seen from a user terminal or a gateway station. To prevent such interference happening, user terminals and gateway stations may be required to shut down transmissions for the time that the NGSO satellite is within 5° of the GSO arc, as seen from the transmitting terminal. Communication is switched to another satellite as the critical angle is approached. The satellite operations center (SOC) calculates the interference times for all user terminals and gateways for every satellite, and distributes this data to all the terminals and gateways. Software at each earth station that controls uplink transmitter power and antenna pointing retains this information and executes the shut down and a switch to a different satellite if necessary. Similar precautions are needed when the path of one NGSO satellite crosses the location of another NGSO satellite operated by a different entity.

NGSO user terminals that are restricted to transmitting only between elevation angles 45° and 90° may not need to shut down, depending on the latitude of the terminal. For terminals north of latitude 35°N the maximum elevation of a GSO satellite is below 45.2°, the top of the GSO arc viewed from latitude 35°N, and the cone of visibility of

an NGSO satellite barely intersects the GSO arc. As a terminal is moved further south, the cone of visibility intersects more and more of the GSO arc, until at the equator the GSO arc passes across the center of the visibility cone. A similar situation exists in the southern hemisphere. Gateway antennas can be located north or south of latitudes 35°N and 35°S to avoid their beams crossing the GSO arc as they track NGSO satellites. Latitudes above 35° are preferred for gateway terminals because heavy rain is less frequent than at lower latitudes. For example, a gateway station located in Seattle, Washington state, at 47.6°N, operating down to an elevation angle of 45° would never need to shut down its transmitters because the top of the GSO arc is at 32.8°, as viewed from Seattle, and there is a separation of at least 12.2° in elevation angle between the station's transmitting beam and any GSO satellite. Optical intersatellite links can be used to transfer data from satellites in latitudes below 35°N and 35°S to satellites further north or south for downlinking to a gateway station. There are many countries with land mass that lies entirely between latitudes 35°N and 35°S that must shut down transmissions from user terminals and gateway stations when a NGSO satellite is within 5° of the GSO arc. However, by placing the gateway station east or west of the footprint of the NGSO satellite, interference with satellites in geostationary orbit can often be avoided.

Reception of transmissions from NGSO satellites in shared frequency bands into the sidelobes of terrestrial antennas used for GSO communication and microwave links is another potential source of interference. There are local and international limits on the power flux density (pfd) of transmissions from NGSO satellites that limit the maximum EIRP of a satellite in a shared frequency band. This is one reason why the EIRP of OneWeb and SpaceX satellites is limited to about 40 dBW in several frequency bands. For details of the pfd limits for these satellites see, for example, (SpaceX FCC filing 2016). Similar limitations exist for GSO satellites, but because they are at a much higher altitude than NGSO satellites and have much larger beam footprints, the EIRP limits are much higher.

11.4 Link Budgets for NGSO Systems

Representative link budgets are presented in this section for an NGSO system similar to those proposed by OneWeb and SpaceX, for downlink frequencies of 18.0 GHz for Ku-band satellites at 1200 km altitude and 40.0 GHz for V-band satellites at 350 km altitude. Uplinks from user terminals are in the 14 GHz band for satellites at 1200 km altitude and in bands from 38 GHz to 50.5 GHz for uplinks from V-band terminals. The OneWeb and SpaceX systems are not identical, but are sufficiently similar that a single uplink and downlink budget demonstrates the performance that can achieved in a broadband internet access system using NGSO satellites.

11.4.1 Example Downlink Budgets for Ku- and V-band Downlinks to User Terminals

In 2018, low noise amplifiers for V-band were available with noise figures in the 4–5 dB range. This translates to a noise temperature of 408–584 K. Clear sky noise temperature is higher at V-band due to greater absorption of microwave energy by oxygen and water vapor molecules, resulting in a typical clear sky earth terminal receiver system noise temperature in the range 500–700 K. A V-band earth terminal system noise temperature

Table 11.5 Example downlink budgets for Ku- and V-band links between the satellite and the user terminal

Parameter	Ku-band 18 GHz downlink		V-band 40 GHz downlink	
Transponder saturated output power	1 W	0 dBW	2 W	3 dBW
Output backoff (regenerative transponder)		0 dB		0 dB
Satellite antenna 1.5° beam on-axis gain	G_t	40.0 dB		40.0 dB
Satellite EIRP on beam axis		40.0 dBW		43.0 dBW
Path loss	1200 km	179.1 dB	350 km	175.3 dB
User terminal receiving antenna gain	G_r	30.6 dB		30.6 dB
Pointing loss		0.5 dB		0.5 dB
Atmospheric clear sky loss		0.7 dB		1.0 dB
Receiving phased array loss of gain at 50° scan angle		2.0 dB		2.0 dB
Miscellaneous losses		0.5 dB		0.5 dB
Received power	C	−112.2 dBW		−105.7 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz		−228.6 dBW/K/Hz
System noise temperature, clear sky	160 K	22.0 dBK	700 K	28.5 dBK
Receiver noise bandwidth	200 MHz	83.0 dB Hz	200 MHz	83.0 dB Hz
Noise power	N	−123.6 dBW		−117.1 dBW
CNR in clear sky		11.4 dB		11.4 dB

of 700 K in clear sky is used in the sample link budgets. For the Ku-band receiver at the user terminal system noise temperature of 160 K in clear sky is used.

The transmitting phased array on the satellite compensates for loss of gain at the edge of its coverage zone, and also increasing path loss, by transmitting additional power as the beam is scanned away from nadir (pointing directly downward). A constant path length of 1200 km for the Ku-band link and 350 km for the V-band link is used in the downlink budgets.

In the link budget illustrated in Table 11.5, the downlink CNR is 2 dB higher at 13.4 dB for a satellite at zenith where the receiving phased array antenna has 0 dB scan loss.

11.4.2 Effect of Rain Attenuation on Downlinks

Figure 11.3 in Section 11.2 shows the spectral efficiency of the DBV-S2 signal format for operation with an implementation margin of 1.8 dB, and was used to describe the operation of the ACM available with the DVB-S2 standard with return channel, for GSO internet access system. The same procedures can be applied for NGSO satellites to determine the availability of uplinks and downlinks. However, the way in which rain affects LEO satellite links is different from GSO links. LEO satellites move in their orbits at velocities around 7.5 km per second. A typical heavy rainstorm is rarely as wide as 7.5 km so the

satellite passes across the storm in a few seconds. How much attenuation the user terminal experiences, and for how long, depends on the relative location of the rainstorm and the terminal. If the rain is directly over the terminal and falls as a vertical shaft, both the attenuation and the duration of a fade will be much longer than when the rain is distant and the beam from the fast moving satellite crosses the rainstorm in a few seconds.

In clear sky conditions the Ku-band and V-band downlinks have CNR of 11.4 dB. This can support 8-PSK modulation with 2/3 rate FEC using the DVB-S2 signal format with an implementation margin of 1.8 dB and a CNR margin of 3.0 dB. This gives a spectral efficiency of 2.0 bits/Hz and a downlink bit rate of 400 Mbps. The DVB-S2 standard has ACM, which allows the link to operate at reduced CNR by changing the modulation and FEC rate. For the case of a satellite at the lowest elevation angle, when rain attenuation on the downlink approaches the 3.0 dB margin, the earth terminal sends a message to the gateway station requesting a change to QPSK and 3/4 rate FEC, for example, which increases the CNR link margin to 5.9 dB, but decreases the bit rate to 300 Mbps. The CNR margin is 7.9 dB for a satellite at zenith because the gain of the earth terminal is 2 dB greater. When the CNR has fallen to 3 dB for the satellite at the lowest elevation angle, the modulation terminal will request a change to QPSK with 1/2 rate FEC and a spectral efficiency of one bit/Hz, giving a downlink speed of 200 Mbps and a CNR margin dynamic range of 8.4 dB (10.4 dB for a satellite at zenith). If the gain of the receiving terminal phased array is reduced by more than 2 dB at the lowest elevation angle, a switch to a lower spectral efficiency may be needed for a short time.

As discussed in Section 11.2 for the GSO internet access satellites, rain in the downlink path causes an increase in sky noise temperature, resulting in rain attenuation values that are up to 3.1 dB lower than the CNR values with high rain attenuation. The dynamic range of 8.4 dB for CNR on the Ku-band downlink is reduced to 5.3 dB for rain attenuation.

The V-band link has a much higher system noise temperature of 700 K in clear sky, and contributes to a smaller decrease in CNR than is the case for the Ku-band link when rain attenuation occurs on the downlink. For example, with 5.2 dB rain attenuation the system noise temperature of the V-band receiver increases to 888 K and the link margin falls by 6.7 dB giving CNR of 4.7 dB. The ACM limit for QPSK with half rate FEC is reached with a rain attenuation of 7.3 dB. Operation is possible at 90 Mbps with QPSK and 1/4 rate FEC, extending the link margin a further 3 dB. To avoid an outage when rain attenuation becomes severe, it may be possible to switch communications to another LEO satellite if the rain is not directly over the earth terminal, and another satellite is within the range of operating elevation angles.

Under clear sky conditions, which prevail for about 95% of the time for a terminal located in all of the United States and Canada except the southeastern states of the United States, the Ku-band and V-band downlinks can reliably deliver 360 Mbps. Only for the remaining 5% of the time will the bit rate decrease, with outages occurring for no more than 20 hours per year.

11.4.3 User Terminal to Satellite Uplinks at 13.0 and 49.0 GHz

Table 11.6 shows a representative uplink budget for links operating at midband frequencies of 13.0 GHz for the Ku-band link and 49.0 GHz for the V-band link. System noise temperatures on a satellite are always higher than those for earth terminals because the satellite sees hot earth rather than cold sky. A noise figure of 8 dB corresponding to a noise temperature of 1430 K is used in the sample V-band uplink budget and 500 K for

Table 11.6 Uplink budgets for Ku- and V-band user terminals between the user terminal and satellite

Parameter	Ku–band 13.0 GHz		V-band 49.0 GHz	
User terminal transmit power	1 W	0 dBW	2 W	3 dBW
Uplink antenna gain, on-axis	G_t	34.0 dB		34.0 dB
Uplink EIRP on beam axis		34.0 dBW		37.0 dBW
Maximum path length and loss	1600 km	178.8 dB	475 km	180.1 dB
Satellite receiving antenna on axis gain	G_r	40.0 dB		40.0 dB
Pointing loss		0.5 dB		0.5 dB
Atmospheric clear sky loss		0.5 dB		1.5 dB
Transmitting phased array loss of gain at 50° scan angle		2.0 dB		2.0 dB
Satellite receiving phased array loss at 50° scan angle		2.0 dB		2.0 dB
Miscellaneous losses		0.5 dB		0.5 dB
Received power	C	−110.3 dBW		−109.6 dBW
Boltzmann's constant	k	−228.6 dBW/K/Hz		−228.6 dBW/K/Hz
Satellite receiver system noise temperature	500 K	27.0 dBK	1430 K	31.6 dBK
Receiver noise bandwidth	18.2 MHz	72.6 dB Hz	18.2 MHz	72.6 dB Hz
Noise power	N	−129.0 dBW		−124.4 dBW
CNR in clear sky		18.7 dB		14.8 dB

the Ku-band uplink. The earth terminal phased array antennas transmit 1.0 W at Ka-band and 2.0 W at V-band.

The uplinks can employ 8-PSK with 3/4 rate FEC, which achieves 2.2 bits/Hz spectral efficiency provided CNR exceeds 10.0 dB, giving CNR margins of 8.7 dB at Ku-band and 4.8 dB at V-band. The satellite system noise temperature does not change with rain in the uplink, so the CNR margins are also the rain attenuation margins. Outage times for the uplinks are similar to the downlinks, with only the V-band link needing to implement ACM under conditions of heavy rain attenuation. The uplink CNR values in Table 11.6 allow for 4 dB loss of gain for a satellite at the lowest elevation angle when both the earth terminal phased array antenna and the satellite phased array antenna suffer a loss in gain of 2 dB. With a satellite at zenith, the loss of antenna gain is 0 dB and the path loss is 2.5 dB less at Ku-band and 3 dB less at V-band. This gives uplink CNR value of 25.2 dB at Ku-band and 22.1 dB at V-band.

11.4.4 Satellite to Gateway Links

The uplink from the gateway to the satellite should be designed to have a higher CNR than the downlink from the satellite to the user terminal to ensure that the uplink

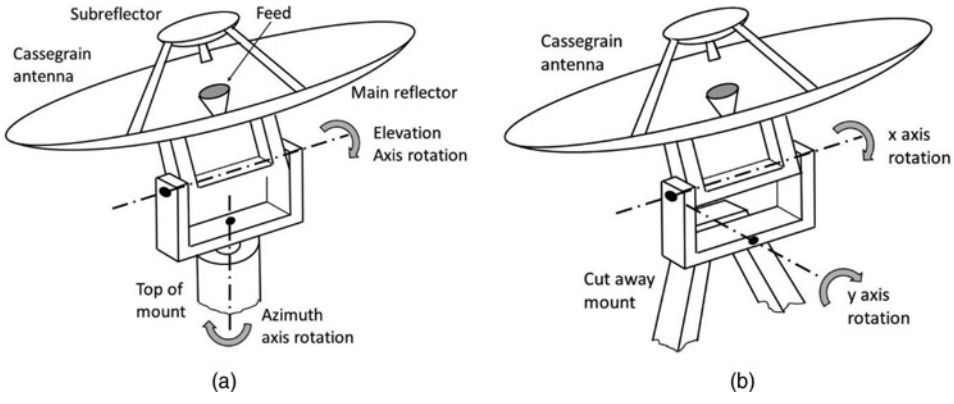


Figure 11.10 Illustration of elevation over azimuth and X-Y antenna mounts. In Figure 11.10a, the Az-El configuration, the elevation axis is horizontal and the azimuth axis is vertical. Typically, the elevation axis is above the azimuth axis. In Figure 11.10b, the X-Y configuration, both *x* and *y* axes are horizontal and orthogonal. (a) Az- El mount (b) X-Y mount.

contributes very few bit errors. With on board processing, the uplink and downlink have separate BERs, which add to give the bit error rate at the user terminal. When rain affects the uplink from the gateway station to the satellite, all the terminals within the satellite footprint are at risk of higher BER, but the users may all be in clear sky conditions and expecting to operate at 360 Mbps. Fast slewing dish antennas at the gateway station with gains of 55–60 dB can be used to overcome the problem. The dishes need to be on X-Y mounts, rather than the Az-El mount commonly used with GSO satellites. An Az-El mount following a satellite that passes overhead must rotate through an angle of 180° when the satellite passes the zenith, resulting in a *cone of silence* around the zenith. The *X–Y* mount does not have this problem and is better suited to LEO satellite systems. Figures 11.10a,b illustrate Az-El and X-Y mounts.

The satellite to gateway downlinks use the 37.5–42.5 GHz band and gateway to satellite uplinks operate in one of three bands: 42.5–43.5, 47.2–50.2, and 50.4–51.4 GHz.

At a midband downlink frequency of 39.5 GHz, a reflector antenna with an aperture efficiency of 70% and a diameter of 2.0 m (6.6 ft) has a gain of 56.8 dB. A similar antenna for the uplink frequency of 47.2 GHz has a gain of 58.3 dB. These antenna gains are 22.8 and 24.3 dB higher than the gain of the user terminal antennas, which is sufficient to maintain the BER on the satellite-gateway links at a very low level even with heavy rain on the link. However, a smaller dish can be used for uplinks if the transmit power is increased, and uplink power control can be used to combat rain attenuation, but the beamwidth of the uplink antenna needs to be kept narrow to avoid interference with GSO satellites.

Gateway stations are typically located in regions where very heavy rainfall rarely occurs. Intersatellite links are used to link satellites serving high rainfall zones to gateways in lower rainfall zones. A cluster of dishes is needed at the gateway station to track up to five visible satellites, with two dishes per satellite so that one can be reset to the correct angles for a satellite appearing at the minimum elevation angle while another is following a different satellite down to the lowest permissible elevation. Additional dishes are needed as spares so that several dishes can be taken out of use for repair or maintenance. The maximum slew rate for a gateway antenna tracking a NGSO satellite is less

than 1° per second, much lower than the slew rate of the antennas used in many aircraft and missile tracking radars.

User terminals in the OneWeb system use MF-TDMA for uplink transmissions, with six terminals sharing a 40 MHz bandwidth transponder that carries a data rate of 38.4 Mbps. With 20 uplink spot beams per satellite, the maximum uplink data rate is 115.2 Mbps. However, Table 11.1 shows that 10 GHz of bandwidth is available for satellite to gateway downlinks. The reason for the difference is that a satellite within visibility of a gateway station may be receiving data from other satellites over its intersatellite optical links that it must forward on its downlink to the gateway station.

The Az-El mount in Figure 11.10a can follow satellites round the horizon, but cannot follow an overhead pass because the azimuth axis must be rotated 180° as the satellite passes overhead. It is used for GSO satellites. The X-Y mount in Figure 11.10b cannot follow satellites round the horizon, but can follow an overhead pass. It is well suited to NGSO satellite that have a limited range of elevation angles well above the horizon.

11.5 Packets and Protocols for NGSO Systems

The topic of data communication via satellite is sufficiently extensive that entire texts are devoted to this subject alone. In this section we present a brief review of some of the methods used to provide internet access over a satellite link. The reader who had a deeper interest in the subject will need to refer to the relevant literature (Zhili 2014; Kota et al. 2004).

A significant advantage of NGSO satellite constellations over GSO satellites is their lower latency, typically 25–37 ms. This allows NGSO satellites to transfer data in many of the formats that are used by terrestrial communication systems – cellular phones, for example. Direct voice connections can be established between cell phones via a NGSO circuit, something that cannot be easily done with a GSO satellite with more than 580 ms latency. Other data formats such as asynchronous transfer mode (ATM) can also be sent over NGSO satellite links.

NGSO constellations have optical links between satellites so that data can be transferred to gateway stations via other satellites in the constellation when none are in view at a particular instant. This feature may offer an advantage to financial organizations that use computer algorithms to trade in distant stock markets, where communication delays of microseconds are important. The Los Angeles–Hong Kong cable called the Pacific Light Cable Network has a length of 13 000 km. Signals travel in optical fibers at 67% of the speed of light due to the dielectric constant of the glass in an optical fiber. A signal traveling 13 000 km in an optical fiber takes about 65 ms to reach its destination. The same signal sent via a constellation of 50 NGSO satellites at 1200 km altitude requires 27 optical links between satellites, assuming optical links operate only between neighboring satellites, and can reach its destination in 48 ms ignoring any delays in the satellites. The 17 ms reduction of propagation delay is significant to such financial companies and may attract their interest in using NGSO satellites.

The success of the world wide web, which evolved into the internet, was the result of the adoption of the TCP/IP protocols. TCP/IP allows any computer or data handling device to be connected to the internet, including personal computers, tablet devices,

cell phones, and the internet of things (IoT). As discussed elsewhere, TCP/IP data packets cannot be transferred over GSO satellite links without protocol conversion because of their high latency, but they can be sent successfully over NGSO satellite links. This makes NGSO systems particularly attractive for internet access and allows the satellite link to appear transparent to the networks at either end of the link.

User terminals for NGSO satellite links must have low cost, comparable to terminals used for DBS-TV. In 1996, a user terminal for Directv in the United States cost US$600, which the subscriber to the service had to pay up front to gain access to the service, in addition to monthly fees for programming. As the number of terminals manufactured for Directv and DishNetwork increased, their cost went down, and by 2005 both companies offered free installation of the terminal provided the customer signed a two year contract, with the cost of the equipment and installation being recovered through the monthly fees. NGSO user terminals will be more expensive than DTH-TV terminals because of the phased array antenna that is required to track NGSO satellites to maintain continuous communication. In 2018, phased array antennas for earth terminals were available from several sources, having been developed primarily for mobile applications on vehicles and aircraft. The projected cost for NGSO applications once volume manufacture got underway was below US$1000 in 2017 and 2018, but above the US$200 objective that would make NGSO satellite services available to the full worldwide audience (Earth terminal phased array antennas 2017, 2018).

As discussed earlier in this chapter, the ETSI standard for DTH-TV, DVB-S2, includes a return channel specification DVB-RCS (digital video broadcast – return channel satellite) that allows DTH-TV terminals to adopt ACM. ACM enables a terminal that is suffering an increase in BER because of rain attenuation to send a message to the gateway station requesting a change in either modulation or FEC coding rate to reduce the BER. To be implemented successfully, ACM requires a virtual circuit between the gateway station and the user terminal. In many DTH-TV systems, the outbound channel is sent to every receiving terminal without the option to modify transmissions for individual receivers, so ACM cannot be used to adapt to downlink attenuation, but could be used to compensate for uplink attenuation. In NGSO internet access systems, a virtual circuit between the gateway and the user terminal is established for every user. That makes the DVB-S2 standard applicable to NGSO satellite systems, and the drafters of the ETSI standard envisaged this application long before the large NGSO systems were proposed. The ASICs at the heart of every DTH-TV receiving system are designed to implement the DVB-S2 standard and can also be used for two-way data communication by implementing the DBS-RCS protocol. Because these ASICs are manufactured in enormous quantities for the DTH-TV market, they have remarkably low cost and are therefore very attractive for NGSO satellite user terminals.

A second attractive feature of the DVB-S2 signal format for internet access is the low BER objective known as QEF transmission of one bit error per hour, equivalent to a BER of 10^{-11}. The low BER is required when digital TV signals are compressed using the MPG-2 and MPEG-4 standards because a single bit error occurring in transmission can cause multiple pixel errors in the recovered video signal. In a NGSO system conveying TCP/IP data packets, the QEF of one bit error per hour means very few automatic repeat requests (ARQs) caused by errors on the satellite link. ARQ requests increase the latency of packet transmissions and need to be kept to a minimum.

The low cost of user terminal ASICs designed for DVB-S2 signals, and low error rate of the DVB-S2 links makes the adoption of DVB-S2 standard likely in NGSO satellite systems. The full details of the standard can be found in the ETSI documentation and

are beyond the scope of this text (ETSI 2009). The DVB-S2 standard is discussed in Chapter 10 for DTH-TV applications; a brief overview of two-way data transfer follows. Data transferred in DVB-S2 links is always in 188 byte packets, referred to as *transport packets*. The 188 byte packet length is related to packets created by MPEG compression of video and audio signals, but can also used for other data transfers. In the video application, the 188 bytes are divided into one byte for synchronization and 187 bytes for payload. Multiple transport packets can be transmitted sequentially in a short frame of eight packets with 16 200 bits after FEC encoding, or a long frame of 32 packets with 64 800 bits. In a TCP/IP internet access application, the payload of the 188 byte transport packets can be filled with TCP/IP packets. Although the DVB-S2 standard allows for a continuous data transmission mode, the advantage of using MPEG transport packets is that the error detection and correction algorithms that produce the QEF of one bit error per hour are tied directly to the transport packet length. Figure 11.11, a repeat of Figure 10.11 in Chapter 10, shows the structure of DVB-S2 short frames of 16 200 bits.

The main difference between DTH-TV and internet access downloads is that DTH-TV is a point to multipoint system delivering identical data streams to thousands of user terminals within the footprint of a GSO satellite. Occasionally, a control packet is addressed to a specific receiver to update information stored there, but for most of the

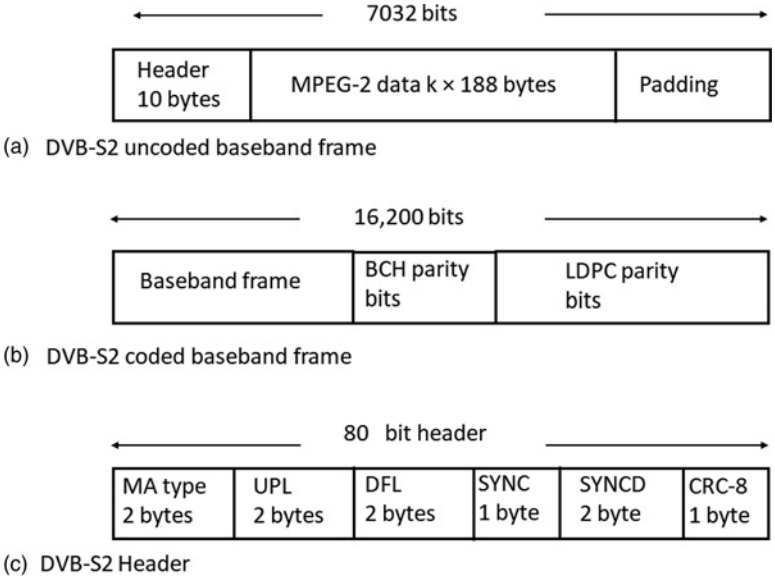


Figure 11.11 Frame structure and header for short frame DVB-S2 transmissions using MPEG-2 compression and fixed coding and modulation. (a) The DVB-S2 uncoded baseband frame consists of a header, *k* blocks of MPEG data, and padding bits to complete a frame of 7032 bits. (b) The DVB-S2 coded frame adds BCH parity bits and LPDC parity bits to form a frame of 16 200 bits. (c) DVB-S2 header structure. BCH: Bose-Chaudhury-Hocquenhem FEC block code; LPDC: Low density parity code FEC code; MA: Defines input stream type – single or multiple – constant or adaptive coding and modulation, and SRRC roll off factor α; UPL: User packet length in bits, typically 8 $k \times 188$; DFL: Data field length in bits, in the range 0–58 112; SYNC: User packet sync byte; SYNCD: Used to determine where in the frame the data bits are located; CRC-8: 8 bit cyclic redundancy error check applied to last 9 bytes of header.

time the transmission is a continuous stream of MPEG packets. Internet access requires a virtual circuit for each user terminal, and is therefore a point to point system. Downloaded packets must contain the address of the recipient, and several header bytes must be included in the transport packet to tell the satellite router where to send the data. When a number of packets are sent sequentially to a single user terminal, the recipient address is not always repeated in every packet.

Uplinks from user terminals can employ the DVB-RCS format for return links to a gateway station. The frame structure for RCS links can be identical to that used in forward links illustrated in Figure 11.11. OneWeb groups six terminals into a single MF-TDMA frame running at 38.6 Mbps, connecting to a 40 MHz bandwidth transponder. If all six terminals are active, throughput could fall to 6.4 Mbps; this is the minimum uplink speed that any user should experience. Download throughput speeds depend on how many user terminals are within the footprint of the NGSO satellite and how many users are active. OneWeb satellites at 1200 km altitude have 20 beams and SpaceX satellites have 50 beams. Using a contention ratio of 100 : 1, there could be 2000 active terminals within the OneWeb satellite footprint and 5000 active users within the SpaceX satellite footprint. If all 50 downlink SpaceX beams are active and delivering data at 360 Mbps, the average throughput for each station will be 7.2 Mbps, assuming equal priority for all stations. However, there may be different priority tiers guaranteeing selected users faster throughput at a higher monthly fee. A spot beam must return to each active earth terminal within 25 ms to meet latency objectives, allowing each terminal a minimum connection time of 500 μs and delivery of 180 kbits every 25 ms.

The preceding discussion has been based on the use of the DVB-S2 standard for broadband satellite systems, with DVB-RCS format for the return channel. A powerful driver is the selection of the DVB-S2 standard is the availability of low cost ASICs that are manufactured by the million for DTH-TV systems. This has been the approach adopted by many GSO broadband satellite systems. In a NGSO system with thousands of satellites and millions of user terminals – possibly billions – a proprietary data communication system could be developed. One hundred million user terminals at US$300 each have a total cost of US$30B, far exceeding the projected cost of the space segment with its thousands of satellites. The largest part of the user terminal cost is likely to be the phased array antenna because of the large number of transmitting and receiving elements that are needed to achieve gain in excess of 30 dB.

The barrier to internet access in rural areas is often the low GDP per head of population. In the United States in 2017, some urban counties had GDP as high as US$60 000 per head while other rural counties had GDP as low as US$11 000 per head (Measuring Broadband America 2018). One solution to providing broadband access for the rural areas is a community antenna system that connects broadband satellites to cellular phones, which are widely available even in poorer communities. NGSO systems have the advantage that they can be adapted to operate with cellular phone protocols for data transmission, and could offer LTE data rates without requiring individual homes to have their own user terminal. Figure 11.12 illustrates the concept of a community antenna broadband satellite internet access system.

A recent article in the British Sunday Telegraph reported that in Uganda and Tanzania, in sub-Saharan Africa, only 20% of the population in rural areas had access to electricity from a utility. However, between 60% and 70% owned cell phones, which were presumably charged from solar sources. The community antenna solution is very attractive for those rural communities (Internet acces in Africa 2019).

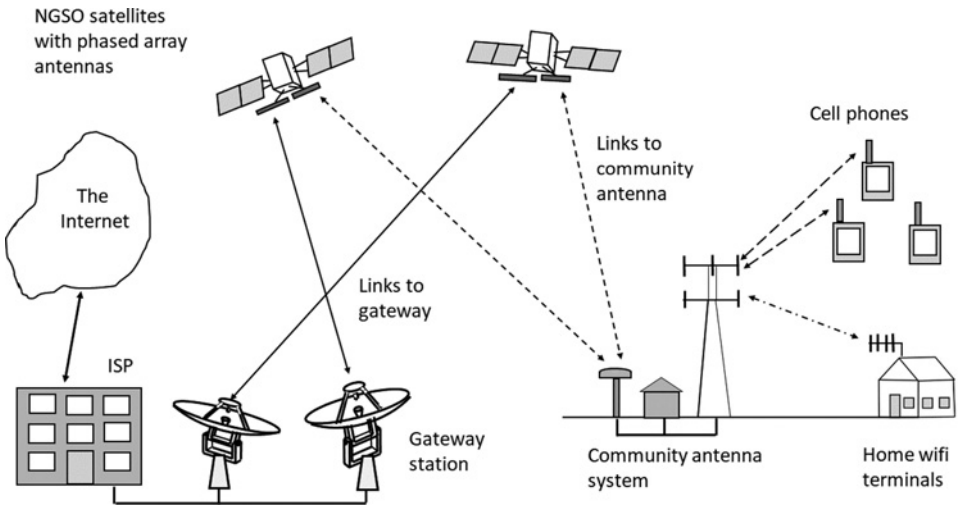


Figure 11.12 Illustration of a community antenna system for internet access that can provide access at a lower cost to individual users. A single user terminal is connected to a local cellular phone system and also a Wi-Fi terminal. Users can link to the local system with a Wi-Fi device, such as a PC or tablet computer, or with a cell phone. Cell phones are much more widely available than PCs in developing countries.

11.6 Gateways, User Terminals, and Onboard Processing Satellites

The transmitters and receivers used in internet access satellite systems are very similar to those used for DTH-TV, with the addition of a transmitter at the user terminal for the return link. GSO broadband satellite systems typically employ an antenna that is slightly larger than a DTH-TV antenna, but the greatest difference is the use of a flat panel phased array at the user terminal by NGSO systems, instead of the offset paraboloid reflector antenna of a DTH-TV installation. The flat panel is oriented in the horizontal plane, can be less than 0.25 m square for VLEO systems, and requires a domed *radome* that sheds rain water. The radome can have a surface that matches its surroundings making it much less obvious than a reflector antenna. The flat antenna is vulnerable to loss of signal because of melting snow that has accumulated on its upward facing surface. For this reason it is preferable to mount the antenna close to the ground in regions where snowfall is common in winter, so that snow on the reflector and feed can easily be removed, rather than in an inaccessible location such as a roof top.

Gateways for GSO broadband satellite systems use large fixed pointing reflector antennas, often located at a teleport, with at least two and sometimes many antennas in one location. The gateway for a NGSO broadband satellite system is quite different because multiple satellites can be in view at the same time, each requiring two reflector antennas that can alternately track the satellites across the sky. Based on the systems discussed in this chapter, there must be at least ten gateway antennas, and when spares are included twelve to fifteen, for each constellation of satellites. Among a group of twelve gateway antennas, five will always be actively tracking satellites and five will be returning to a low elevation angle, to wait for the next satellite to appear. First contact with a LEO or VLEO satellite by a gateway station can be made well below the minimum elevation

angle used for communication, and can be continued after the satellite passes out of the cone of visibility that defines the operating region. This ensures continuity of data processing, as communication must be switched from one gateway antenna to another as satellites pass out of the visibility cone.

Satellites used for GSO internet access are similar to those used for DTH-TV, but have much greater capacity through the use of multiple spot beams and frequency reuse, and are typically among the largest commercial satellites in geostationary orbit, with a mass of 6000 kg or more. GSO broadband satellites have several reflector antennas with multiple feed horns or phased array feeds to generate the multiple beams needed to achieve capacities exceeding 100 Gbps. The beams can be fixed or movable, connected to wide band transponders, with different frequencies and polarizations assigned to beams serving a region of the earth. The destination or origin of data packets determines which beam, which RF frequency, and which polarization will be used for the downlink and uplink. There is a fixed relationship between the uplink channel frequency and the spot beam and down link frequency in the GSO satellite. An RF switch in the transmitter section of the transponder allows the relationship to be changed as traffic demands alter and high power amplifiers (HPAs) fail. Figure 11.13 illustrates a transponder

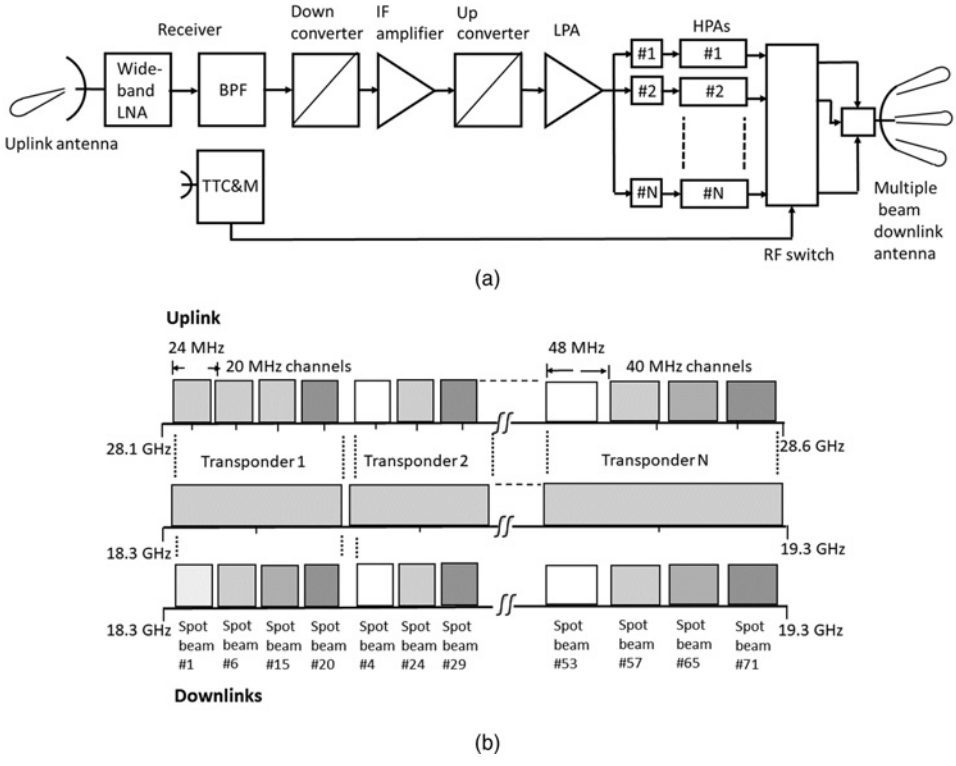


Figure 11.13 (a) Example of transponders for a gateway to user link in a GSO internet access system in the 30/20 GHz band. (b) Frequency plan. Uplink channels are directed to transponders with different bandwidths. The uplink channels are downconverted to downlink frequencies and connect to spot beams through bandpass filters that select the channels. The HPAs of the transponders are linear (bent pipe). Interconnection of uplink channels to spot beam downlink channels can be changed in the RF switch via telecommand.

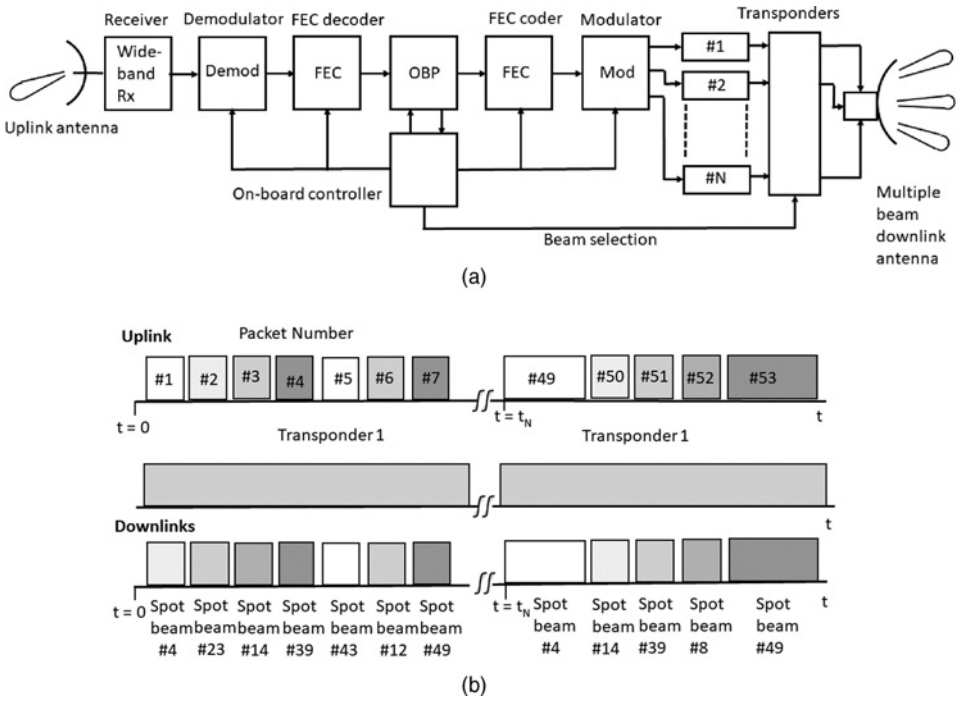


Figure 11.14 (a) Onboard processing transponders for NGSO internet access system gateway to satellite link. Packets received from the uplink are converted to baseband. The onboard processor reads the header information of each packet and instructs the controller which downlink beam to select for that packet. Each wideband transponder connects to multiple downlink spot beams. (b) Illustration of routing of packets by the onboard processor. Starting at time zero, a stream of packets is received from the uplink. The headers indicate that these packets should be sent to transponder #1 and that the downlink spot beams should be directed to the terminal numbers indicated. At time N, transponder #1 is receiving new packets, but terminals #12, #23, and #43 have no packets. Terminals #4 and #49 receive longer packets with more data and terminal #8 now receives a packet. The downlink must return to each terminal that is receiving data within 20 ms to avoid latency issues.

implementation in a gateway to user link through a GSO satellite operating in the 30/20 GHz band, and its frequency plan.

NGSO satellites have onboard processing that creates a router in the sky. Signals from the gateway stations are recovered at baseband so that packet headers can be read to determine the destination of the packet, and hence which downlink beam and time slot should be allocated to the packet. Uplink packets from user terminals are formed into frames that are forwarded to whichever gateway station is in view. Figure 11.14 illustrates an onboard processing transponder for an NGSO satellite. Signals received from the uplink are down converted to an intermediate frequency (IF) frequency, demodulated and error corrected, then passed to the onboard processor where the packet headers are examined. Packet headers contain routing information that specifies the downlink frequency and beam for each packet, and the address of the user terminal. The packets are modulated and error correction coding applied, then up converted to the specified frequency and sent to a transponder. The onboard controller directs the packets to the correct downlink beam for transmission to the user terminal

specified by the address in the packet header. The process can be repeated for every packet, with individual packets being sent to separate user terminals, or a string of packets may be sent to one user to increase that user's throughput.

11.7 Total Capacity of OneWeb and SpaceX Proposed NGSO Constellations

Satellites at 1200 km altitude have a footprint that is 2120 km wide when the minimum elevation angle is set to 40°. Satellites that are more than 1060 km off shore from a land mass can communicate directly only with ships and aircraft and will therefore have relatively few users compared to satellites over land. A very approximate analysis shows that in the southern hemisphere, only 25% of the satellites will be in contact with land based terminals, and in the northern hemisphere, the number is about 75%. Thus over the entire globe, only about half of satellites in polar orbits can be carrying high volume traffic. Satellites north of latitude 60°N and south of 60°S are serving areas with sparse populations and will have little traffic. As a result, only about one third of the total number of satellites in a polar orbit constellation have the opportunity to carry a full load of traffic. For OneWeb's 882 Ku-Ka-band satellites with nominal 50 Gbps capability, the total capacity of the constellation is 14.7 Tb. The situation for VLEO satellites is slightly worse because they must be within 350 km of a land mass to serve land based terminals. Some satellites over oceans or poles may be transferring data via their optical links, but this does not increase the total capacity of the constellation.

SpaceX uses five inclined orbital planes in its constellation of 4425 Ka-/Ku-band satellites, as shown in Table 11.2. One third of the constellation has an orbital inclination of 53°, which keeps 3200 satellites over land for a greater percentage of their orbit than a polar orbit. The remaining three orbital planes have inclinations between 70° and 81° containing a total of 815 satellites that will be lightly loaded for part of each orbit. If the 3200 satellites in 53° orbital planes are loaded to 50% and the remainder to 25%, the total capacity of the constellation with 50 Gbps capacity per satellite is 90 Tb.

Both OneWeb and SpaceX have proposed VLEO constellations in inclined orbits with a total of 6425 satellites. Applying similar analysis, the usable capacity of the VLEO satellites is 160 Tbps, giving a total of 250 Tbps for the two NGSO systems. Assuming that there are 3 billion households worldwide, and using a contention ratio of 100 : 1, the two NGSO constellations could potentially provide every household in the world with roughly 4 Mbps download speed and 500 kbps upload speed. That would be a notable achievement, and would fulfill the original Teledesic concept of wiring the world from space. If more NGSO systems are built and operate successfully, higher speed internet access could be offered to everyone worldwide, perhaps at 8 Mbps down and 1.25 Mbps up, the rates offered by LTE cellular telephone companies in 2018.

11.8 End of Life Disposal of NGSO Satellites

The volume of space between altitudes of 250 and 1500 km contains a large quantity of *space debris*. As of July 2013, more than 170 million debris smaller than 1 cm (0.4 in.), about 670 000 debris 1–10 cm, and around 29 000 larger debris were estimated to be

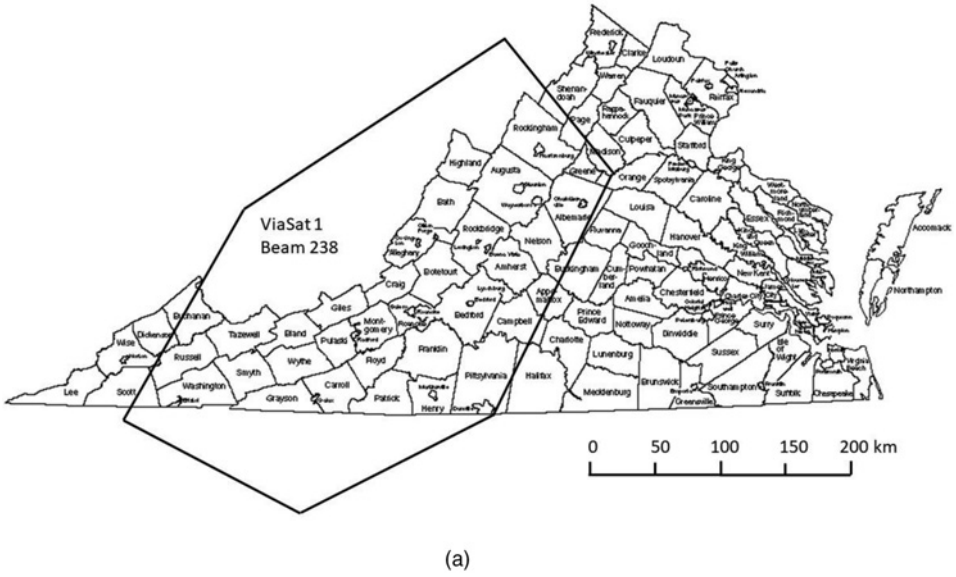
in orbit (Space debris 2018). The debris consists of LEO satellites that have reached the end of their useful life and been shut down, upper stages of launch vehicles that remained in orbit after the payload separated, and thousands of parts of rockets and satellites destroyed by explosions or left by operation of *explosive bolts*. Explosive bolts are bolts with a small explosive charge inside that can be detonated by an electrical pulse. They are used to separate satellites from their launch vehicles, fairings that surround payloads, and anything else that requires two pieces of hardware to be physically separated in space. (Springs and latches can be used for the same purpose without creating space debris, but being mechanical can be less reliable than explosive bolts.) A further contributor to large quantities of space debris is collisions between satellites. Satellites in LEO are traveling at seven kilometers a second and therefore become many small pieces when a collision occurs. Some satellites have been deliberately destroyed by interceptor rockets launched from earth (Space debris 2018).

International regulations require that all NGSO satellites launched into LEO be moved to a low orbit at the outer edge of the earth's atmosphere that will slowly decay, resulting in their return to the earth's atmosphere where they will burn up. Submissions to the US FCC for approval of an NGSO constellation include details of the proposed method of *mission disposal*. See, for example, (SpaceX FCC V-band filing 2016). Satellites in MEO and GSO are moved to a graveyard orbit, usually several hundred kilometers above their operating orbit, before all their maneuvering fuel is exhausted.

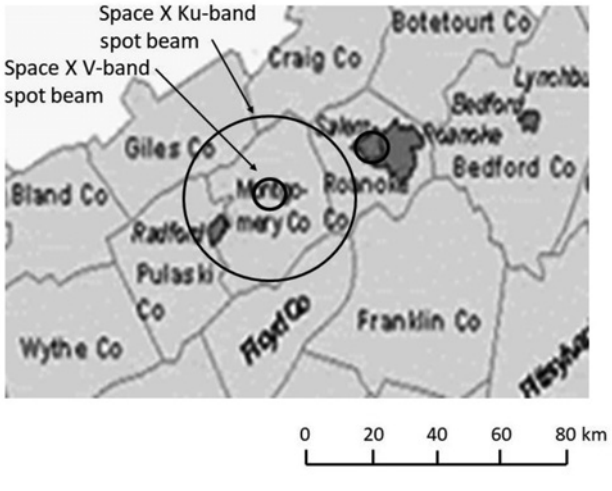
11.9 Comparison of Spot Beam Coverage of GSO and LEO Internet Access Satellites

The spot beam from a GSO satellite is much wider than the spot beam of a LEO satellite, because of the far greater distance between a GSO satellite and earth, compared to the distance between a LEO satellite and earth. Figure 11.15 illustrates this difference for two satellites: ViaSat 1, a large GSO satellite with 72 beams covering parts of the United States, and SpaceX, a LEO satellite constellation, with satellites at 1200 and 350 km altitudes.

Spot beam #385 of the ViaSat 1 satellite is illustrated in Figure 11.15a superimposed on a map of Virginia. This beam covers the western portion of the state of Virginia and is centered on Blacksburg, home to Virginia Tech. The −3 dB contour of the ViaSat 1 Ka-band spot beam is approximately circular and 430 km wide, equal to the north–south dimension of beam #385, but adjacent beams #329 to the east and #327 to the west overlap beam #328 so the east and west boundaries of the beam are inside the −3 dB contour. Figure 11.15b shows the −3 dB contours of the Ku-band and V-band spot beams of the SpaceX satellite, centered on Montgomery County. Blacksburg is approximately in the center of Montgomery county. The Ku-band beam is approximately 52 km wide and the V-band beam is 9 km wide. A second V-band beam is shown in Figure 11.15b over the city of Salem, 40 km to the NE of Blacksburg. Both Blacksburg and Salem have broadband internet access from terrestrial providers, and parts of Blacksburg have fiber optic service at 1 Gbps. Montgomery county, and the adjacent counties in SW Virginia are mountainous, so there are places in Montgomery county, and more so in the surrounding rural counties that do not have terrestrial broadband internet, making satellite internet access the only option.



(a)



(b)

Figure 11.15 Comparison of GSO and LEO spot beam dimensions over southwest Virginia. (a) ViaSat Ka-band beam #385 is centered on Blacksburg, Virginia, and is approximately 430 km in the north-south direction. (b) Spot beams of the SpaceX LEO satellite centered on Montgomery County and the city of Salem, Virginia. The SpaceX Ku-band beam for LEO satellites at 1200 km altitude is 52 km wide and the V-band beam for satellites at 350 km altitude is 9 km wide.

11.10 User Terminal Antennas for Ku-Band, Ka-Band, and V-Band

Figure 11.16 shows examples of antennas for user terminals of satellite internet access systems.

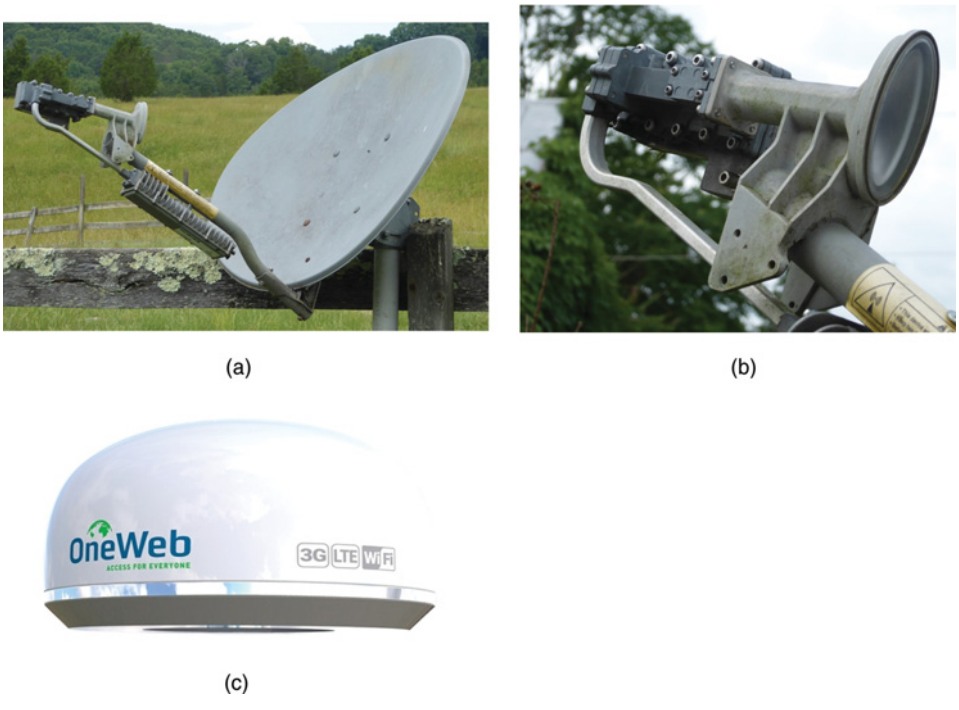


Figure 11.16 Antennas for user terminals and gateways (a) Mid 2000s Hughesnet Ku band user terminal. The HPA is in a finned case below the feed support. (b) Feed horn and LNA for Hughesnet antenna in (a). Note that the elliptical feed is wide in the vertical plane to create a narrow beam in the horizontal plane of the reflector. (c) OneWeb user terminal antenna. For a color version of this figure please see color plate section. Source: Photo credits: (a) and (b) Tim Pratt. (c) Courtesy of OneWeb, © OneWeb 2018.

11.11 Summary

In many developed countries like the United States, the majority of the population live in or close to towns and cities and have access to broadband internet at speeds of at least 5 Mbps download and 1 Mbps upload. However, people living in rural areas do not have broadband access to the internet, including an estimated 14 million in the United States. In developing countries that lack a terrestrial communications infrastructure, large percentages of the population have no access to the internet. Satellite communication systems can provide internet access to underserved areas, using GSO satellites and also NGSO satellites. Historically, internet access by satellite was developed from direct to home TV broadcasting with GSO satellites. Oversubscription caused severe drops in download speeds at busy times prior to 2011 when the first high capacity Ka-band satellite came on line with a capacity of 140 Gbits. Larger GSO satellites with capacity up to 1000 Gbps are planned for launch in the 2020s.

Non-geostationary orbit satellites, and in particular VLEO constellations of thousands of satellites offer the possibility of providing everyone in the world with internet access. This was the vision of Teledesic, created in 1994 with the concept of wiring the world from space, instead of on the ground. Teledesic was too far ahead of the technology

needed to achieve this goal, which eventually became possible toward the end of the 2010 decade with systems developed by SpaceX and OneWeb. Phased array antennas are needed for LEO satellites so that multiple beams from the satellite can track user terminals, and also for user terminals so they can track the fast moving satellites. Producing millions of user terminal phased array antennas at an affordable price is a major challenge.

Internet access systems based on large numbers of VLEO satellites operate in a different way from GSO systems. VLEO satellites must use onboard processing to create a router than directs data packets via a switched beam antenna on the satellite to the user terminals. The multiple steerable beams of the downlink phased array on a VLEO satellite can select a small number of users on the ground within a circle of 9 km diameter, compared to the state-sized downlink beams of a GSO satellite. Gateway stations for LEO systems require larger numbers of reflector antennas to track the LEO satellites as they cross the sky.

Exercises

11.1 Create a link budget for the uplink from a 7 m gateway antenna to the Ka-band satellite at 30.0 GHz and the downlink from the satellite to the gateway at 20.0 GHz using the parameters for antenna gains and system noise temperatures given in Table 11.2. Assume that the gateway 7 m antenna has a 100 W TWTA providing 10 W to the uplink channel in clear sky conditions and employing uplink power control when rain attenuation affects the uplink. The downlink transmitter on the satellite has a transmit power of 10 W. The antennas on the satellite have a diameter of 1.5 m and an aperture efficiency of 65%.
Do the links meet the criterion of CNRs that are at least 20 dB higher than those on the satellite to user links in clear sky conditions? Does ACM need to be implemented when heavy rain affects the gateway to satellite links?

11.2 Construct two graphs similar to Figure 11.4 showing how ACM can combat rain attenuation in the downlinks of the Ku-band and V-band LEO satellites in the link budgets in Table 11.5. Repeat the exercise for the link budgets in Table 11.6.

11.3 Tabulate link budgets for NGSO satellite to gateway, and gateway to satellite links using the frequency bands listed in this section. What are the CNR margins for these links if the gateway earth station is equipped with 1.5 m antennas with 70% aperture efficiency and all transmissions use 2 W transmit power with receiver noise bandwidth of 200 MHz?

References

Anik F2 (2018). https://en.wikipedia.org/wiki/Anik_(satellite) (accessed 12 August 2018).
Earth terminal phased array antennas (2017). http://interactive.satellitetoday.com/via/may-june-2017/phased-array-antennas-can-they-deliver (accessed 12 August 2018).
Earth terminal phased array antennas (2018). https://spacenews.com/internet-for-the-masses-not-a-focus-for-kymeta-phasor (accessed 7 August 2018).

Echostar 19 (2017). http://www.echostarsatelliteservices.com/satellitefleet/fleet.aspx (accessed 19 August 2018).
Echostar 24 (2017). http://www.hughes.com/technologies/hughes-high-throughput-satellite-constellation/echostar-xxiv (accessed 19 August 2018).
ETSI (2009). ETSI EN 302 307 V1.2.1 (2009–08) standard for DVB-S2, European Telecommunications Standards Institute, 650 Route des Lucioles, F-06921 Sophia Antipolis Cedex – FRANCE.
Eutelsat (2018). http://www.eutelsat.com/en/home.html (accessed 13 August 2018).
Griffin, T. (2016). https//25iq.com/2016/07/23/a-dozen-things-i-learned-being-involved-in-one-of-the-most-ambitious-startups-ever-conceived-teledesic/ (accessed 10 August 2018).
HughesNet (2006). http://www.terradaily.com/reports/HughesNet_Broadband_Satellite_Surpasses_300000_Subscribers_999.html (accessed 23 July 2018).
Internet access in Africa (2019). https://en.wikipedia.org/wiki/Internet_in_Africa (accessed 15 May 2019).
Ippolito, L.J. Jr. (2017). *Satellite Communications Systems Engineering: Atmospheric Effects, Satellite Link Design and System Performance*. Hoboken, NJ: Wiley.
Ka-Sat (2012). https://www.itu.int/dms_pub/itu-r/md/12/iturka.band/c/R12-ITURKA.BAND-C-0004!!PDF-E.pdf (accessed 27 August 2018).
Ka-Sat (2018). https://www.eutelsat.com/en/satellites/the-fleet/EUTELSAT-KA-SAT.html;jsessionid=A65F0E02AF1CBD469FDDAD2203AD0DB4 (accessed 13 August 2018).
Kota, S.L., Pahlavan, K., and Leppanen, P.A. (2004). *Broadband Satellite Communications for Internet Access*. New York: Springer Science + Business Media.
Measuring Broadband America (2011). https://www.fcc.gov/reports-research/reports/measuring-broadband-america/measuring-broadband-america-august-2011 (accessed 1 August 2018).
Measuring Broadband America (2016). https://www.fcc.gov/reports-research/reports/measuring-broadband-america/measuring-fixed-broadband-report-2016 (accessed 1 August 2018).
Measuring Broadband America (2018). https://www.fcc.gov/reports-research/reports/international-broadband-data-reports/international-broadband-data-report-4 (accessed 9 August 2018).
Microsats A and B (2018). https://en.wikipedia.org/wiki/Starlink_(satellite_constellation) (accessed 15 August 2018).
Mitchell, W.C., Nguyen, L.N., Dissanayake, A. et al. (1997). https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19980019442.pdf (accessed 20 August 2018).
O3B (2018). https://en.wikipedia.org/wiki/O3b_(satellite) (accessed 12 August 2018).
OneWeb (2017). https://spacenews.com/fcc-gets-five-new-applications-for-non-geostationary-satellite-constellations (accessed 1 August 2018).
Oneweb FCC filing (2017). https://transition.fcc.gov/Daily_Releases/Daily_Business/2017/db0601/DOC-345159A1.pdf (accessed 1 August 2018).
OneWeb satellites (2018). https://en.wikipedia.org/wiki/OneWeb_satellite_constellation (accessed 15 August 2018).
OneWeb FCC filing (2017). http://www.licensing.fcc.gov/myibfs/download.do?attachment_key=1190495 (accessed 15 August 2018).
Phased array antennas (2018). https://www.netbigfuture.com/2018/09/spacex-phase-shifting-array-for-receiving-internet-from-satellites.html (accessed 15 September 2018).

Propagation events (1997). Arbesser-Rastburg, B. https://ieeexplore.ieee.org/document/598408 (accessed 15 August 2018).
Space debris (2018). https://en.wikipedia.org/wiki/Space_debris (accessed 14 August 2018).
Space Norway (2017). https://spacenews.com/fcc-grants-telesat-leo-market-access-despite-viasat-protests (accessed 14 August 2018).
SpaceX FCC Ku-/Ka band filing (2016). https://licensing.fcc.gov/myibfs/download.do?attachment_key=1158350 (accessed 1 August 2018).
SpaceX FCC V-band filing (2016). https://licensing.fcc.gov/myibfs/download.do?attachment_key=1190019 (accessed 1 August 2018).
Sweeting (2017). https://www.aerosociety.com/media/5612/2017-banquet-speech-by-sir-martni-sweeting-group-executive-chairman-sstl.pdf (accessed 23 May 2018).
Teledesic (2018). https://en.wikipedia.org/wiki/Teledesic (accessed 10 August 2018).
Time (2018). http://time.com/4638470/spacex-internet-elon-musk (accessed 11 August 2018).
TooWay (2018). http://www.toowayhome.com (accessed 13 August 2018).
Verizon (2018). https://fios.verizon.com/fios-speeds.html (accessed 1 August 2018).
World population (2018). https://en.wikipedia.org/wiki/World_population (accessed 11 August 2018).
Zhili, S. (2014). *Satellite Networking*. Chichester, London: Wiley.

12

Satellite Navigation and the Global Positioning System

The global positioning satellite (GPS) system has revolutionized navigation and position location. It is now the primary means of navigation for most ships, aircraft, and automobiles, and is widely used in surveying and many other applications. The GPS system, originally called NAVSTAR, was developed as a military navigation system for guiding missiles, ships, and aircraft to their targets. GPS satellites transmit L-band signals that are modulated by several codes. The principles of the GPS system were made public in 1983, and GPS receivers using the C/A (*coarse acquisition*) code became available to the public by 1990. Early GPS receivers cost thousands of dollars, but prices fell quickly once volume production began. The GPS system was declared fully operational with 24 satellites in 1995. The secure high accuracy P-code allows authorized users (mainly military) to achieve positioning accuracy of 1 m. This was the accuracy that the military users wanted for targeting smart bombs and cruise missiles, but such accuracies are also useful for auto-landing aircraft in fog and for docking ships in bad weather (Parkinson and Spilker 1996).

The first commercial use of GPS was in surveying, but by 1995 several companies had produced low cost, handheld GPS receivers for general position location and navigation. The development of integrated circuits (ICs) specifically for GPS receivers and larger volume production quickly brought down the price of a GPS receiver, and the market expanded rapidly. GPS receivers are now a consumer product, and can be found in many cars and all cellular telephones. The GPS system has been successful because it is available everywhere in the word, is free to all users, and provides a direct readout of the present position of a GPS receiver with a typical accuracy of 5 m. The success of GPS is an excellent example of what satellites do best: *broadcasting*. An unlimited number of GPS receivers can operate simultaneously because all that a GPS receiver has to do to locate itself is to receive signals from four GPS satellites. Earlier radio location systems such as Loran could achieve accuracies around 300 m under good conditions, but have been discontinued because of the far greater accuracy and reliability of GPS.

The success of the GPS system has encouraged other countries to develop their own Global Navigation Satellite Systems (GNNSs) using similar technology to GPS. The former USSR developed GLONASS, now operated by Russia, Europe has developed Galileo, China has developed BeiDou, and Japan has a local satellite based location system. A significant driver in the development of these alternative systems is that the US government controls the availability of the GPS system for civilian users and can switch off that section in the event of an international political crisis (such as war).

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

In this chapter, we will concentrate on GPS, with only brief references to alternative satellite navigation systems. All modern satellite navigation systems use the same principles as GPS. GPS and the other GLONASS systems have become of major importance worldwide, with investment running into hundreds of billions of dollars. The literature of GPS is extensive, with many text books devoted to the topic and thousands of technical documents. This chapter provides an overview of the subject; the reader who needs more information on GPS and its European counterpart Galileo can find full specifications of both systems on the internet.

12.1 The Global Positioning System

In 2018, the GPS space segment consisted of 27 satellites in medium earth orbit (MEO) at a nominal altitude of 20 183 km with an orbital inclination of 55°. The satellites are clustered in groups of four, called constellations, with each constellation separated by 60° in longitude, with an additional three satellites for improved coverage in certain parts of the world. The orbital period is approximately one half a sidereal day (11 hours 58 minutes) so the same satellites appear in the same position in the sky each day. The satellites carry station keeping fuel and are maintained in the required orbits by occasional station keeping maneuvers, just like geostationary earth orbit (GEO) satellites. The orbits of the 24 GPS satellites ensure that at anytime, anywhere in the world, a GPS receiver can receive signals from at least four satellites. Up to ten satellites may be visible at times, and at least four satellites are visible all of the time. Replacement satellites are launched as needed, so there may be more than 24 operational GPS satellites at any given time.

Figure 12.1 shows a GPS Block IIF satellite. There have been four generations of GPS satellites since the first launch of a Block II satellite in 1990. All Block II satellites were decommissioned by 2016. Block IIR, Block IIR-M, and Block IIF satellites were operational in 2018 with a total of 31 operating in space. A further generation Block III-IIIF was planned for launch beginning in 2018 (GPS satellites 2018).

The Block IIF satellites weighed 1408 kg (3230 lb) at launch and had a design lifetime of 12 years. DC power of 1900 W was generated by Gallium Arsenide solar cells. The satellites have three axis stabilization and multiple thrusters for orbital corrections. Fuel for the thrusters is hydrazine, with 140 kg (320 lb) at launch. Because GPS is an integral part of the defense of the United States, spare GPS satellites are kept in orbit and more spares are ready for immediate launch. The GPS system is operated by the US Air Force from the GPS master control station (MCS) at Schriever Air Force Base in Colorado Springs, CO, with an alternative MCS at Vandenburg Air Force Base in California (GPS Control segment 2017).

The MCS and a series of subsidiary control stations around the globe continuously monitor all GPS satellites as they come into view and determine the orbit of each satellite. The MCS and other stations calculate ephemeris data for each satellite, atomic clock error, and numerous other parameters needed for the *Navigation Message*. The data are then transmitted to the satellite using a secure S-band link and used to update onboard stored data. There are six GPS control stations located in Hawaii, Colorado, Florida, Ascension Island in the Atlantic Ocean, Diego Garcia in the Indian Ocean, and Kwajalein in the Pacific Ocean. Each satellite carries two rubidium clocks and one cesium clock, with an accuracy better than one part in 10^{-12}, corresponding to a maximum

Figure 12.1 Block IIF GPS satellite. Source: US government. For a color version of this figure please see color plate section.

timing error of less than a tenth of a microsecond over 24 hours. Corrections to the satellite's clock time are sent via an S-band telemetry link from the monitoring stations at regular intervals during each day. The control stations have precise cesium time standards and make continuous measurements of range to all visible satellites. These measurements are performed every 1.5 seconds, and are used to provide updates for the navigation messages. In addition, there are 10 monitoring stations located around the world that report back to the MCS with information on the current status of all GPS satellites (GPS.gov 2018).

The position of a GPS receiver is found by trilateration, which is one of the simplest and most accurate methods of locating an unknown position. In trilateration, the distance of the unknown point from three known points is measured. The intersection of the arcs corresponding to three distances defines the unknown point relative to the known points, since three measurements can be used to solve three equations to give the latitude, longitude, and elevation of the receiver. The distance between a transmitter and a receiver can be found by measuring the time it takes for a pulse of RF energy to travel between the two. The distance is calculated using the velocity of electromagnetic waves in free space, which is assumed to be equal to the velocity of light, c, with $c = 299\,792\,458\ \mathrm{m\,s^{-1}}$. The velocity of light in a vacuum was established by international agreement in 1975 at the 15th Conférence Générale des Poids et Mesures meeting of

the Bureau International des Poids et Mesures (BIPM, International Bureau on Weights and Measures in English) (BIPM 1975). This value for the speed of light is used in all GPS calculations (Interface specification IS-GPS-705B 2011). Time can be measured electronically more accurately than any other physical parameter by the use of atomic clocks, and this is how the GPS position location system can achieve a measurement accuracy of 1 m in a distance of 25 000 km. It is rare in electrical engineering that any parameter is specified to nine significant figures, but GPS is the exception because of the need to measure time extremely accurately. To achieve a position location accuracy of 1 m, timing measurements must have an accuracy better than 3 ns. This is possible with sophisticated system design, modern digital circuitry, and a great deal of averaging.

It is not possible to make timing measurements with this precision with a single transmitted pulse. All GNSS satellites use direct sequence code division multiple access (CDMA) transmissions (spread spectrum) to send a long sequence of pulses in a wide bandwidth that are correlated with an identical sequence generated in the receiver. The spread spectrum signal is buried well below the receiver noise floor, typically with carrier to noise ratio (CNR) around −20 dB for the civil C/A code. The correlation process removes the code sequence from the received signal allowing it to be processed through a narrow bandwidth giving an output with signal to noise ratio (SNR) well above 0 dB and making accurate time measurements possible. The principles of spread spectrum systems are discussed in Chapter 6.

Each satellite radiates a different sequence of bits (known as chips in a spread spectrum system) which starts at a precisely known time. A GPS receiver contains a clock that is synchronized in turn to the clock on each satellite that it is receiving. The receiver measures the time delay of the arrival of the chip sequence, which is proportional to the distance between the satellite and the GPS receiver. The position of each satellite is calculated in the GPS receiver using the ephemeris for the satellite orbits that are broadcast by each satellite in a *navigation message*. Making the calculation for four satellites provides the receiver with sufficient information to determine its position with very good accuracy. Four satellites, rather than three are needed because the clock in the receiver is not inherently accurate enough. The fourth distance measurement provides information from which clock errors in the receiver can be corrected and the receiver clock synchronized to GPS time with an accuracy better than 10 ns.

Originally GPS satellites were designed to transmit two signals at different frequencies, known as L1 and L2. The L2 signal has a frequency of 1227.6 GHZ (120 times the basic GPS frequency of 10.23 MHz) and is modulated with a 10.23 Mbps pseudorandom (PRN) bit sequence called the *P-code* that is used by military positioning systems. The P-code is transmitted in an encrypted form known as the *Y code,* which restricts the use of the P-code to authorized users. The L1 carrier at a frequency of 1575.42 Hz (154 times the basic GPS frequency of 10.23 MHz) is modulated by a 1.023 Mbps PRN sequence called the *C/A code* that is available for public use, and also carries the P-code as a quadrature modulation. The higher bit rate of the P-code provides better measurement accuracy than the 1.023 Mbps C/A code.

C/A stands for *coarse acquisition* and P stands for *precise*. GPS systems using the secure Y code require the C/A code as an intermediate step in making distance measurements with high accuracy. In the early days of the GPS, the accuracy of C/A code receivers was deliberately degraded some of the time by a process called *selective availability* (SA). SA caused variations in the C/A code satellite transmissions that resulted in less accurate calculation of position. SA was discontinued in May 2000 and will not be

applied again. The existence of other GNSS systems with accuracy comparable or better than GPS has rendered selective availability useless.

The GPS system provides two categories of service. The precise positioning service (PPS) receivers track both P-code and C/A code on L1 and L2 frequencies. The PPS is used mainly by military users, since the P-code is encrypted into the Y code before transmission and requires decryption equipment in the receiver. Standard positioning service (SPS) receivers track the C/A code on L1. This is the service that is used by the general public. Both the C/A codes and the P-codes are publicly available, but the P-code cannot be recovered in a GPS receiver without a knowledge of the Y code decryption algorithm. In this discussion we will concentrate on the C/A code and its use in the SPS. The P-code transmissions are used in differential global positioning system (DGPS) systems where the phase of the signal is measured without any knowledge of the P-code itself.

A number of upgrades to later GPS satellites were defined in 2001 (McDonald 2002). These include adding a civil L2 signal called L2C at 1227.6 GHz to make dual frequency transmissions available to civil users so that more accurate estimation of ionospheric delay is possible, transmitting quadrature C/A signals with only one component modulated by the navigation message to improve the ability of the receiver to lock to the signal, and adding a new L5 frequency at 1176.45 MHz with a longer PRN sequence. All of the changes are intended to improve the accuracy of the standard positioning system to 1 m. By 2011, investment by the US government in GPS was estimated to have exceeded US$53B, with about US$1B spent each succeeding year on operations and upgrades (GPS cost 2011).

The former USSR built and operated a global navigation system that is very similar to GPS, known in the West as GLONASS, for GLobal Orbital NAvigation Satellite System, with a total of 30 satellites. Almost everything about GLONASS is similar to GPS except the multiple access technique. GLONASS uses frequency division multiple access (FDMA), with a different transmit frequency at each of 15 satellites. Satellites on opposite sides of the earth share the same frequency, since interference cannot occur. The equivalents of the P-code and C/A code are transmitted by GLONASS satellites in RF channels spaced 562.5 and 437.5 kHz apart centered at 1246 and 1602 MHz (Grewal et al. 2013). A frequency synthesizer that can be tuned to the unique frequency of each satellite is required in a GLONASS receiver. However, GLONASS has somewhat lower accuracy than GPS. The GLONASS system is now operated by Russia and has been upgraded (GLONASS 2017).

The European Union has built a similar satellite navigation system to GPS called Galileo. The program was authorized in 2003 and by December 2017, 22 of the planned 30 active satellites were in orbit, at an estimated program cost of €5 B (Galileo 2017). China has constructed a GNSS called BeiDou that is similar to the Galileo system, and Japan has its own regional position location system (Beidou 2018; Tsui 2000). The reason for multiple GNSS systems is a perceived need by many nations for a precise navigation system without dependence on the GPS system of the United States or the Russian GLONASS system.

12.2 Radio and Satellite Navigation

Prior to the development of radio, navigation was by compass and landmarks on land, and by the sun and stars at sea. Neither technique provides high accuracy, and

shipwrecks caused by inaccurate navigation and foggy weather were a common occurrence. On land, people often got lost in wilderness areas (and still do). Pilots of light aircraft, relying solely on a map and landmarks, would get lost and run out of fuel before they found somewhere to land. With a GPS receiver and a map, it is almost impossible to get lost. GPS receivers are very popular with airplane pilots, owners of seagoing boats, and wilderness hikers, as well as most automobile drivers on out of town trips.

The development of aircraft that could fly above the clouds, and particularly the building of large numbers of bomber aircraft in the 1930s, made radio navigation essential. Military thinking after WWI, and during WWII, placed high reliance on the ability of bomber aircraft to win a war by destroying the weapon manufacturing capability of the enemy. During WWII, the allies sent 1000 bomber aircraft at a time to targets in Germany, causing immense destruction to many cities. The philosophy of mass destruction continued after WWII with the development of nuclear bombs, intercontinental ballistic missiles (ICBMs), and cruise missiles. However, bomber aircraft, ICBMs, and cruise missiles must find their targets, so accurate navigation is an essential part of each of these weapon systems. This demand for accurate targeting of airborne weapons, especially submarine launched ICBMs, led to the development of GPS. However, the majority of GPS users are now civilian, and the worldwide market for GPS equipment was projected to be worth US$75 billion in 2017 (SAI 2015).

Commercial aircraft have historically flown on federal airways using VOR (very high frequency omni range) beacons. The airways are 8 mi wide to allow for the angular accuracy of VOR measurements, which is around 4°. GPS has replaced VOR navigation, allowing aircraft to fly directly from point of origin to destination, but the system of VOR beacons in the United States is likely to remain for many years as a backup to GPS.

The earliest radio navigation systems, developed in the 1930 and 1940s, were simple transmitting stations operating in the mf (AM) radio band. A direction finding antenna and receiver can make an indicator needle point at the transmitter, so that an aircraft or a ship can home in on the beacon. The beacons used by aircraft are called *non-directional beacons* (NDB) and the receiver is called an *automatic direction finder* (ADF). NDBs and an ADF receiver were the main means of radio navigation for aircraft in the 1930 and 1940s, but have several serious disadvantages. If a strong wind is blowing across the path that the aircraft is taking to the NDB, it may fly a curved path instead of a straight line. The FAA has been decommissioning most of the NDB beacons in the United States, but they are still in use in many other countries.

The NDB was largely superseded by the VOR beacon by the 1950s. VOR stands for very high frequency (*VHF*) *omni range*. A VOR transmitter generates a rotating VHF radio beam and also radiates a continuous sine wave signal that is phase referenced to the time that the rotating beam sweeps through north. A VOR receiver on the aircraft synchronizes to the reference signal and measures the angle of the beam relative to north at the time it is received. With two VOR transmitters and a map showing their locations, the aircraft can determine its position. Many VOR stations have DME (*distance measuring equipment*). The aircraft DME equipment transmits a pair of pulses to a transponder at the VOR station and measures the time for the round trip to the VOR and back, which provides a measurement of range to the VOR station. With knowledge of range and angle to a VOR beacon, navigation is possible using a single VOR-DME station.

WWII aircraft needed to navigate to targets over enemy territory where there were no VORs available. *Hyperbolic navigators* were developed in Germany, Britain, and the United States during WWII to provide radio navigation at longer ranges than can be achieved in the VHF band, by using frequencies between 100 kHz and 2 MHz. The low RF frequencies can propagate round the earth's curvature making long range navigation possible. A hyperbolic navigator has three radio transmitters that transmit at the same frequency, or very close frequencies. In the earlier forms, the phase of the radio wave was used, but later systems like LORAN timed the arrivals of pulse transmissions. The receiver compares the phase or time of arrival of the radio signals from two transmitters and calculates the difference in distance between the two transmitters. A line with a constant difference in the distance between two points is a hyperbola with the two transmitters at the foci. A third transmitter provides two more hyperbolas, and their intersection locates the receiver, hence the name hyperbolic navigator.

Instrument landing systems (ILSs) are essential when aircraft must land in conditions of poor visibility. An ILS installation at an airport provides two radio beams that allow the aircraft to fly an approach along a straight line in space to the runway threshold. The *localizer*, a VHF transmitter and antenna at the end of the runway, provides two modulated beams in the horizontal plane. A vertical needle on a *course deviation indicator (CDI)* in the aircraft cockpit shows the aircraft's lateral position relative to a line leading to the runway threshold. A *glide slope* transmitter at the side of the runway transmits another radio beam which points upwards at about three degrees. A horizontal needle on the CDI shows the position of the aircraft relative to the glide slope, with a sensitivity of ± 3 m as the aircraft approaches the ground. The pilot of the aircraft tries to keep both the CDI needles centered, so that the aircraft flies a straight line down the glide slope and arrives at a height of 15 m above the runway threshold (Clarke 1996).

GPS can provide a single navigation system with better accuracy and reliability than all earlier radio navigation aids. It can provide navigation of aircraft directly between airports, instead of indirectly via airways, while providing absolute position readout of latitude and longitude. DGPS can be used instead of ILS to provide the required straight line in the sky for an instrument approach to a runway, and can be linked to an autopilot to provide automatic landing of aircraft in zero visibility conditions. Ships can safely navigate and dock in treacherous waters in bad weather by using DGPS. Eventually, GPS will replace all other means of navigation, although some may be retained as back-up systems in case of failure of the GPS receiver(s) or jamming of the signals.

GPS was preceded by an earlier satellite navigation system called *Transit*, built for the US Navy for ship navigation, which achieved much lower accuracy and became obsolete when GPS was introduced. Transit satellites were in low earth orbits and the system used the Doppler shift observed at the receiver when a beacon signal was transmitted by the satellite. Because of the high velocity of LEO satellites – about 7.5 km/s – their signals are significantly shifted up in frequency when the satellite appears over the horizon with a component of velocity toward the receiver. The Doppler shift falls to zero as the satellite passes the observer, and then becomes negative as the satellite flies away. Observation of the Doppler shift with time, which may need to be as long as 10 minutes, and a knowledge of the satellite orbit, allows calculation of the receiver's position.

Air traffic control (ATC) systems using radar were developed after WWII to ensure separation between aircraft in increasingly busy airspace. The resolution of ATC radars

is limited to 2 km close to airports and 8 km at long ranges, with position updates at 6–10 second intervals. A GPS receiver on an aircraft can determine its position within a few meters, and generate updates every second. The US Federal Aviation Administration (FAA) decided to change to a GPS based ATC system from 2020 using a system known as automatic dependent surveillance-broadcast (ADS-B) that relies on aircraft reporting their position to a network of ground receivers across North America. The program was begun in 2001 under the name NexGen and is expected to be complete by 2025 (FAA NexGen 2018). The ADS-B system is discussed in Section 12.11 of this chapter.

The success of the FAA in ensuring safe travel by commercial airlines in the United States is illustrated by the fact that there were no passenger fatalities between 2010 and 2017, with an estimated 2.5 M passengers traveling by air every day. By comparison, train and bus accidents caused many fatalities during this period, and motor vehicles resulted in 237 851 fatalities between 2010 and 2016 at a cost to society of US$242B (IIHS 2018). Politicians like to claim that government regulations in the United States are overbearing and should be eliminated. Airlines and commercial aircraft in the United States are one of the most heavily regulated sectors of the US economy, but the result is that air travel is safe thanks to the FAA, ATC, and GPS: the dangerous part of traveling by air is driving to and from the airport.

12.3 GPS Position Location Principles

The basic requirement of a satellite navigation system like GPS is that there must be four satellites transmitting suitably coded signals from known positions. Three satellites are required to provide the three distance measurements, and the fourth is used to remove receiver clock error. Figure 12.2 shows the general arrangement of position location with GPS. The three satellites provide distance information when the GPS receiver makes three measurements of *range*, R_i, from the receiver to three known points, that is, GPS satellites. Each distance R_i can be thought of as the radius of a sphere with a GPS satellite at its center. The receiver lies at the intersection of three such spheres, with a satellite at the center of each sphere. Locally, at the receiver, the spheres will appear to be planes since the radii of the spheres are very large. A basic principle of geometry is that the

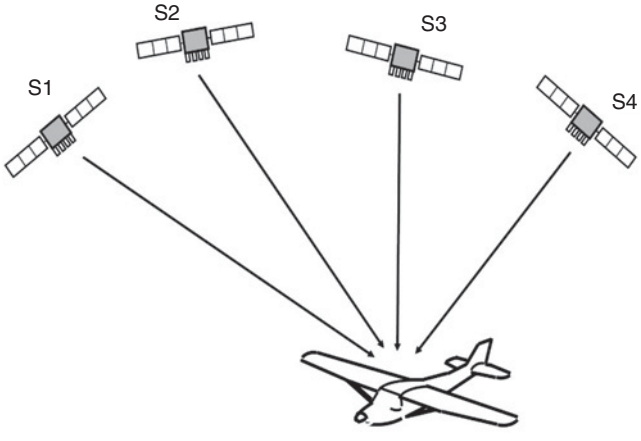


Figure 12.2 Position location by trilateration. The aircraft must receive signals from three GPS satellites to find its location in three dimensions. The fourth GPS satellite signal is required to correct differences between the satellite clock and the internal clock of the GPS receiver.

intersection of three planes completely defines a point. Thus three satellites, through measurement of their distances to the receiver, define the receiver location close to the earth's surface. There is another point in outer space where the three spheres intersect, but it is easily eliminated in the calculation process.

Although the principles by which GPS locates a receiver are very simple, requiring only the accurate measurement of three ranges to three satellites, implementing the measurement with the required accuracy is quite complex. We will look first at the way in which range is measured in a GPS receiver and then consider how to make the measurements. Range is calculated from the time delay incurred by the satellite signal in traveling from the satellite to the GPS receiver, using the known velocity of EM waves in free space. To measure the time delay, we must know the precise instant at which the signal was transmitted, and we must have a clock in the receiver that is synchronized to the clock on the satellite.

GPS satellites each carry three atomic clocks, which are calibrated against time standards in GPS control stations around the world. The result is *GPS time,* a time standard that is available in every GPS satellite. The accuracy of an atomic clock is typically 1 part in 10^{12}. A standard crystal oscillator with a long term accuracy of 1 in 10^5 or 1 in 10^6 is used in low cost civil GPS receivers. However, over the short time period in which GPS location measurements are made, the oscillator is stable to one part in 10^{12}. Atomic clocks with a stability of 6×10^{-12} were available in IC form for less than US$20 by 2018, but have been included mainly in military grade GPS receivers to date. The receiver clock is allowed to have an offset relative to the GPS satellite clocks, so when a time delay measurement is made, the measurement will have an error caused by the *clock offset*. For example, suppose the receiver clock has an offset of 10 ms relative to GPS time. All distance measurements will then have an error of 3000 km. Clearly, we must have a way to remove the time error from the receiver clock before we can make accurate position measurements. C/A code receivers can synchronize their internal clocks to GPS time within 10 ns, corresponding to a distance measurement uncertainty of 3 m. Repeated measurements and integration improve the position location error to below 10 m.

It is surprisingly easy to remove the clock error, and this removal is one of the strengths of GPS. All that is needed is a time measurement from a fourth satellite. We need three time measurements to define the location of the receiver in the three unknown coordinates x, y, and z. When we add a fourth time measurement we can solve the basic position location equations for a fourth unknown – the receiver clock offset error τ (often called *clock bias*). Thus the four unknowns in the calculation of the location of the receiver are x, y, z, and τ.

12.3.1 Position Location in GPS

First, we will define the coordinates of the GPS receiver and the GPS satellites in a rectangular coordinate system with its origin at the center of the earth. This is called the earth centered earth fixed (ECEF) coordinate system, and is part of the WGS-84 description of the earth. WGS-84 is an internationally agreed description of the earth's shape and parameters, derived from observations in many countries (Strang and Borre 1997). GPS receivers use the WGS-84 parameters to calculate the orbits of the GPS satellites with the accuracy required for precise measurement of the range to the satellites. The Z-axis of the coordinate system is directed through the earth's north pole and the X- and Y-axes are in the equatorial plane. The X-axis passes through the Greenwich meridian – the

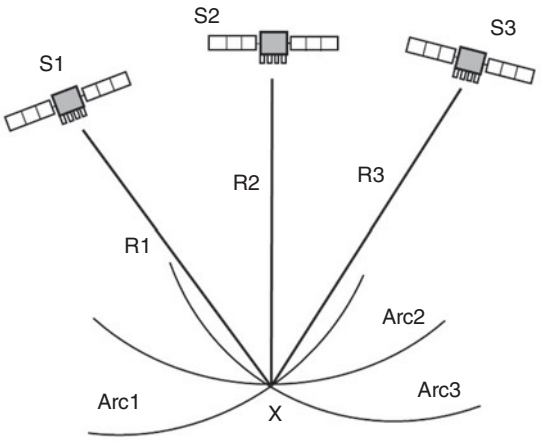


Figure 12.3 Position location by the measurement of the distance from three known points**.** The GPS receiver is at point X. The three arcs lie on the surface of spheres centered on each GPS satellite S1, S2, and S3, and have radii R1, R2, and R3. The intersection of three spheres uniquely defines a point in space – the GPS receiver in this case.

line of zero longitude on the earth's surface, and the Y-axis passes through the 90° east meridian. The ECEF coordinate system rotates with the earth. The receiver coordinates are (U_x, U_y, U_z), and the four satellites have coordinates (X_i, Y_i, Z_i), where $i = 1, 2, 3, 4$. There may be more than four satellite signals available, but we use only four signals in a basic position calculation. The measured distance to satellite number (i) is called a *pseudo range*, PR_i, because it uses the internal clock of the receiver to make a timing measurement that includes errors caused by receiver clock offset. The geometry of a GPS measurement is illustrated in Figure 12.3.

Pseudo range, denoted as PR_i, is measured from the propagation time delay T_i between the satellite (number i) and the GPS receiver, assuming that EM waves travel with velocity c m/s.

$$PR_i = T_i \times c \text{ m} \tag{12.1}$$

The distance R between two points A and B in a rectangular coordinate system is given by

$$R^2 = (x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2 \text{ m}^2 \tag{12.2}$$

The equations that relate pseudo range to time delay are called *ranging equations*:

$$\begin{aligned}
(X_1 - U_x)^2 + (Y_1 - U_y)^2 + (Z_1 - U_z)^2 &= (PR_1 - c)^2 \\
(X_2 - U_x)^2 + (Y_2 - U_y)^2 + (Z_2 - U_z)^2 &= (PR_2 - c)^2 \\
(X_3 - U_x)^2 + (Y_3 - U_y)^2 + (Z_3 - U_z)^2 &= (PR_3 - c)^2 \\
(X_4 - U_x)^2 + (Y_4 - U_y)^2 + (Z_4 - U_z)^2 &= (PR_4 - c)^2
\end{aligned} \tag{12.3}$$

where τ is receiver clock error (offset or bias). With four equations we can solve for four unknowns.

The position of the satellite at the instant it sent the timing signal (which is actually the start of a long sequence of chips) is obtained from ephemeris data transmitted along with the timing signals in the navigation message. Each satellite sends out a data stream that includes ephemeris data for itself and the adjacent satellites. The receiver calculates the coordinates of the satellite relative to the center of the earth (X_i, Y_i, Z_i), at the instant the satellite started to transmit the chip sequence and then solves the four ranging equations

for the four unknowns using standard numerical techniques for the solution of non-linear simultaneous equations. (The equations are non-linear because of the squared terms.)

The four unknowns are the location of the GPS receiver, (U_x, U_y, U_z), relative to the center of the earth and the clock offset τ – called *clock bias* in GPS terminology. The receiver position is then referenced to the surface of the earth, and can be displayed in latitude, longitude, and elevation. Typical accuracy for a GPS receiver using the GPS C/A code is 5 m defined as a 2DRMS error. The term DRMS means the root mean square (RMS) error of the measured position relative to the true position of the receiver. If the measurement errors are Gaussian distributed, as if often the case, 68% of the measured position results will be within a distance of 1DRMS from the true location and 95% of the results will be within 2DRMS of the true location. Accuracy in GPS measurements is usually defined in terms of 2DRMS, in the horizontal or vertical plane.

In practice, the error surface that encloses 68% or 95% of all measurements is not a circle but an ellipse, and the error in any dimension is affected by several *dilution of precision* (DOP) factors. DOP is discussed later in this chapter. Atmospheric propagation effects (tropospheric and ionospheric) cause errors in the timing measurements made by a GPS receiver, leading to position location errors. The atmosphere and the ionosphere introduce timing errors because the propagation velocity of the GPS signals deviates from the assumed free space value. The errors can be largely removed if a number of GPS reference stations are built at precisely known locations. The stations observe the GPS signals and compute the current error in position as calculated from GPS data. This information can then be broadcast to all GPS users by a GEO satellite as a set of corrections to be applied to GPS measurements. The US system is called a *wide area augmentation system* (WAAS) and is one example of a space based augmentation system (SBAS).

A network of 38 WAAS stations in the United States, Canada, and Mexico provide aircraft with improved position measurement accuracy. Using WAAS, accuracies of 2 m can be obtained with C/A code receivers. WAAS also includes an integrity monitoring system to ensure that the GPS signals used by aircraft do not contain errors, which could cause false readings. Aircraft GPS receivers used for approach guidance to airports must perform an internal *receiver autonomous integrity monitoring* (RAIM) check to ensure that the GPS system is working correctly for a safe approach and landing. Alternatively, the receiver can use integrity information sent via the WAAS system, which is required to send a warning of possible errors within 5.6 seconds if a problem is detected with any GPS satellite signal. The International Civil Aviation Organization (ICAO) calls this type of system a *satellite-based augmentation system* (SBAS). Europe and Asia are developing their own SBASs: the Indian Global Positioning System Aided Geo Augmented Navigation (GAGAN), the European Geostationary Navigation Overlay Service (EGNOS), and the Japanese Multi-Functional Satellite Augmentation System (MSAS), respectively.

Similarly, a single reference station at a known location – for example, an airport – can determine the local measurement error in GPS and broadcast this information to GPS users so that greater accuracy can be obtained with a C/A code receiver. This is one (simple) form of *differential GPS* (DGPS). More complex forms of DGPS use a fixed reference station so that phase comparisons can be made by the receiver. With lengthy integration times and a sophisticated phase comparison receiver, DGPS accuracies of 1 cm can be obtained. With DGPS, the receiver computes its position relative to the reference station rather than in latitude and longitude. DGPS is used when a vehicle,

aircraft, or ship needs to be positioned accurately with respect to a fixed point, such as an aircraft with respect to a runway or a ship with respect to a berth, and also in land surveying and agriculture.

12.3.2 GPS Time

The clock bias value τ, which is found as part of the position location calculation process can be added to the GPS receiver clock time to yield a time measurement that is synchronized to the GPS time standard. The crystal oscillator used in the GPS receiver is highly stable over a period of a few seconds, but will have a frequency that changes with temperature and with time. Temperature changes cause the quartz crystal that is the frequency determining element of a crystal oscillator to expand or contract, and this changes the oscillator frequency. Crystals also age, which causes the frequency to change over time. The changes are very small, but sufficient to cause errors in the clock time at the receiver when the clock is not synchronized to a satellite. Calculating the clock bias by solving the ranging equations allows the receiver clock time to be updated every second or two so that the GPS receiver time readout is identical to GPS time.

Every GPS receiver is automatically synchronized to every other GPS receiver anywhere in the world through GPS time. This makes every GPS receiver a super clock which knows time more accurately than any other time standard. Prior to the widespread use of GPS receivers, standard time transmissions were broadcast by the US National Institute of Science and Technology (NIST, formerly the Bureau of Standards). The broadcasts were made in the HF (shortwave) band, and could be received throughout the United States. However, the HF signals propagate over long distance by reflection from the ionosphere, which introduces an uncertain delay into the time of arrival of the signal. The time standard provided by GPS is typically accurate to better than 170 ns, and has been used to synchronize electric power generators across the United States, networks of cell phone towers, and for scientific applications that require synchronized clocks in different locations, and also as a long term frequency standard. GPS time differs from Greenwich Mean Time (GMT or UTC) because UTC is tied to the rotation of the earth. Leap seconds are added to UTC to account for the slowing of the earth's rotation, but not to GPS time. As a result, GPS time differed from UTC by 19 seconds in 2018 (GPS.gov 2018).

12.4 GPS Codes and Frequencies

GPS satellites transmit *pseudo-random sequence* (PRN) codes, also known as *pseudo noise codes*. All satellites transmit a C/A code at the same carrier frequency, 1575.42 MHz, called L1, using binary phase shift keying (BPSK) modulation. The L1 frequency is 154 times the master clock frequency of 10.23 MHz. The C/A code has a clock rate of 1.023 MHz and the C/A code sequence has 1023 chips, so the PRN sequence lasts exactly 1.0 ms. The exact frequencies at a GPS receiver are about 0.005 Hz lower than stated here to allow for relativistic effects caused by the high velocity of the satellites in their orbits (3.865 km/s). (GPS measurements are one of the few examples where relativistic effects must be taken into account, because the clocks are mounted on platforms moving at very high speeds.)

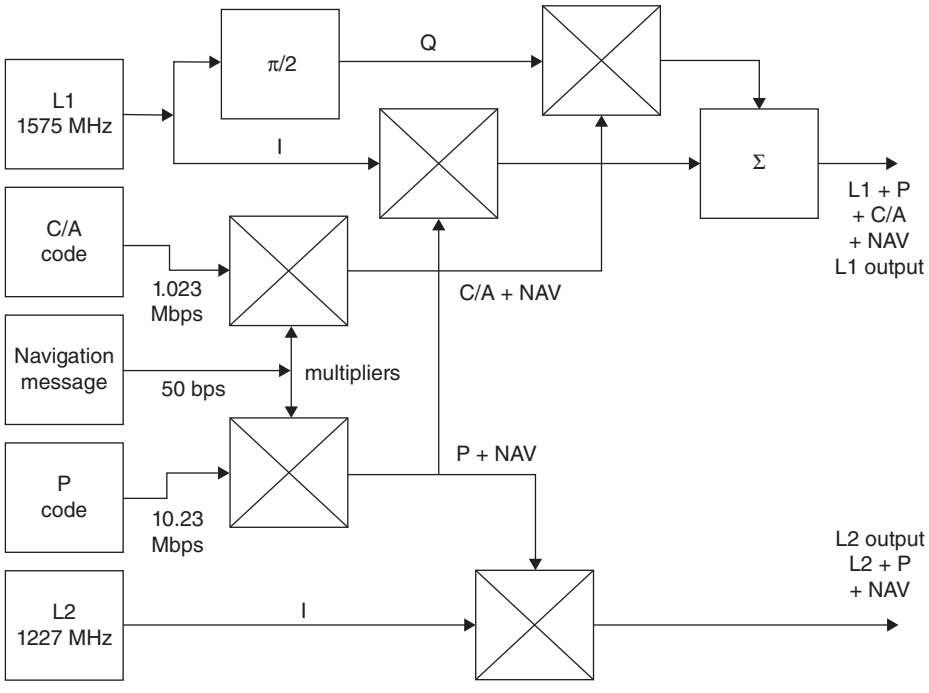


Figure 12.4 Generation of C/A and P-code signals in early Block II GPS satellites. The C/A code and P-code are transmitted on frequency L1 using quadrature BPSK modulation. The P-code is transmitted on the L1 and L2 frequencies. Later satellites transmit additional signals.

The P-code is transmitted using BPSK modulation at the L2 carrier frequency of 1227.6 MHz (120 × 10.23 MHz), and is also transmitted with BPSK modulation on the L1 carrier frequency, in phase quadrature with the C/A code BPSK modulation. Figure 12.4 shows the way in which the L1 and L2 signals are generated on board a GPS satellite.

The C/A and P-code transmissions from all GPS satellites are overlaid in the L1 and L2 frequency bands, making GPS a direct sequence spread spectrum (DSSS) system (see Chapter 6 for details of spread spectrum techniques). The receiver separates signals from individual GPS satellites using knowledge of the unique C/A code that is allocated to each satellite. At most, 12 GPS satellites can be seen by a receiver at any one time, so the coding gain in the spread spectrum receiver must be sufficient to overcome the interference created by 11 unwanted signals while recovering the twelfth wanted signal. However, GPS signals received on earth are very weak, and receiver noise power is much greater than interference from other GPS satellites.

12.4.1 The C/A Code

The C/A codes transmitted by GPS satellites are all 1023 bit *Gold codes*. GPS C/A Gold codes are formed from two 1023 bit *m-sequences,* called G1 and G2, by multiplying together the G1 and G2 sequences with different time offsets. An m-sequence is a maximum length PRN sequence, which is easy to generate with a shift register and feedback

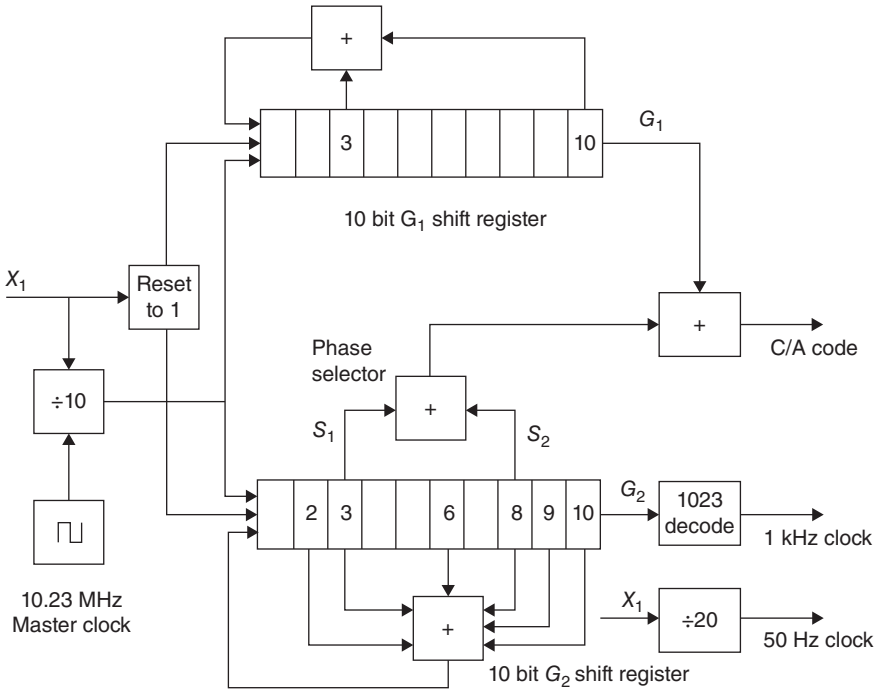


Figure 12.5 C/A code generator. Identical code generators are used on GPS satellites and in GPS receivers. The C/A code is created by combining the outputs of the G1 and G2 shift registers. The tap settings S1 and S2 on shift register G2 define the number of the C/A code. In this example, taps 3 and 8 are selected, which generates code #31. The 10.23 MHz master clock is derived from an atomic clock on the satellite with extremely high stability. All the other frequencies are directly related to the master clock frequency. The clock in the receiver has much lower accuracy than the satellite clock and is synchronized by the use of the fourth GPS satellite signal.

taps. A shift register with n stages can generate a PRN sequence $2^n - 1$ bits in length. The bit pattern is set by the output taps of the G2 shift register, which create a different delay for each sequence. The PRN sequences G1 and G2 are both generated by 10 bit shift registers and are therefore both 1023 bits long. The clock rate for the C/A code is 1.023 MHz, so each sequence lasts 1.0 ms. Figure 12.5 shows a generator diagram for the C/A code.

The C/A code for a particular satellite is created with an algorithm that includes the identification number of the GPS satellite, creating a unique code with a signal number that is the same as the GPS satellite number (*SV number*). The algorithm for generating a C/A code for SV number i is

$$C_{\mathrm{i}}(t) = G1(t) \oplus G2(t + n_{\mathrm{i}} T_{\mathrm{c}}) \tag{12.4}$$

where n_{i} is a unique value for each C/A code sequence and T_{c} is the C/A code chip period. The $\oplus$ symbol is the exclusive OR function.

There are 37 Gold code sequences available numbered 1–37, although sequences 33–37 are not used by GPS satellites; some are allocated to WAAS satellite transmissions and sequences 34 and 37 are the same. Low cross correlation of the sequences

is a requirement because the GPS receiver can pick up signals from as many as 12 satellites at the same time, so not all the possible sequences that can be generated are used.

As an example, in Figure 12.5 the output taps of the G2 shift register are set to three and eight. The phase selector is an exclusive OR adder with output $3 \oplus 8$, giving code number 31, which starts with the 10 bit sequence 1110010101 and has a code delay n_i of 861 bits. All the G2 code sequences are the same; they differ by a delay that is dependent on the G2 shift register tap settings. For the C/A code set, the delay n_i varies between 5 bits for code #1 to 950 bits for code #37 (GPS.gov 1995; Tsui 2000).

A correlator in the receiver looks for one of the sequences and must reject all other sequences that are present. Two C/A code sequences with zero cross correlation would achieve a rejection ratio of 1023, but the 37 available C/A code sequences will not all have zero cross correlation. The selected group of 37 are the sequences with the lowest levels of cross correlation among the available set of Gold code sequences. They also have low autocorrelation time sidelobes, another requirement of DSSS systems.

The C/A code sequence length of 1.000 ms gives range ambiguity of 300 km, since the code travels at a velocity of approximately 3×10^8 m/s and therefore has a length in space of 3×10^5 m. The entire C/A code sequence repeats in space every 300 km, leading to ambiguity of position only if the GPS receiver is in outer space. The ambiguity is easily resolved if the receiver knows roughly where it is; just knowing that the receiver is located close to the earth's surface is usually sufficient.

Figure 12.6 shows a simplified block diagram of a C/A code GPS receiver. The antenna is typically a circularly polarized patch antenna with a low noise amplifier (LNA) mounted on the underside of the antenna's ground plane. A conventional superhet receiver is used to generate an IF signal in the range 4–20 MHz with a bandwidth of about 2 MHz, which is sampled using I and Q sampling techniques and processed

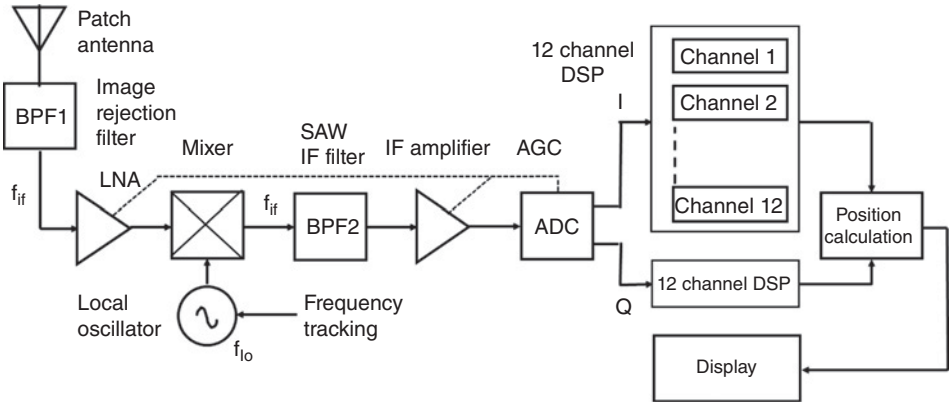


Figure 12.6 Simplified diagram of a single frequency C/A code GPS receiver. Signal frequency is L1 1575.42 MHz. BPF1 is a low loss image rejection and interference rejection filter. IF frequency is typically 4–20 MHz and there may be two down conversion stages. There are two analog to digital converters (ADC) in phase quadrature to preserve the phase of the IF signal. There are 12 parallel receiver channels in the I and the Q DSP receivers, which may be implemented as an ASIC. A separate microprocessor can be used for the position calculation and to drive the display, store maps, and other information. Frequency tracking is applied to the local oscillator to compensate for Doppler shift. Automatic gain control (AGC) can be used to help with weak satellite signals.

digitally. Some receivers use direct conversion to baseband. The digital portion of the receiver includes a C/A code generator, a correlator and a microprocessor, field programmable gate arrays (FPGAs), or application specific integrated circuits (ASICs) that makes the timing measurements and calculate the receiver's position. Most GPS receivers make use of a 12 channel IC chip set and some have an antenna mounted on top of the integrated circuit. Price for a single fully functional GPS integrated circuit varied between US$40 and US$80 in 2018 (Sparkfun 2018). GPS integrated circuits used in cellular telephones are purchased by cell phone manufacturers in huge quantities at much lower prices.

The analog to digital converter (ADC) used in GPS receivers has typically 1 bit to 3 bits. A 1 bit ADC has two possible outputs; a digital 1 if the signal is positive and a digital zero if the signal is negative. This adds quantization noise to the signal, but because the signal is about 20 dB below thermal noise the quantization noise is entirely masked by the thermal noise. There is a SNR penalty of 2 dB with 1 bit quantization (Grewal et al. 2013). Automatic gain control (AGC) is not required with 1 bit quantization, but is needed with a three bit ADC. I and Q ADCs are needed to preserve the phase of the IF signal.

12.4.2 The P-Code

The P-code for the *i*th satellite is generated in a similar way to the C/A code. The algorithm is

$$P_{\mathrm{i}}(t) = X_1(t) \oplus [X_2(t) + (i-1)T_{\mathrm{p}}] \tag{12.5}$$

where T_{p} is the chip period, the X_1 sequence contains 15 345 000 bits and repeats every 1.5 seconds, and the $X2$ sequence is 37 bits longer. The P-code repeats after 266.4 days, but is changed every 7 days for security reasons. The long length of the P-code sequence makes the distance measurements unambiguous. P-code sequences cannot be acquired easily because they do not repeat and are encrypted to form the Y code before transmission, features that prevent unauthorized users from operating high accuracy P-code GPS receivers. The C/A code provides information to authorized users on the starting time of the P-code; this is contained in the navigation message as an encrypted *handover word*. If the current feedback tap settings for the P-code generators are known, and the handover word is decrypted, the receiver can start the local X code generators close to the correct point in the P-code sequence. This allows rapid acquisition of the P-code, and is the origin of the name *coarse acquisition* for the C/A code.

12.5 Satellite Signal Acquisition

GPS signals are very weak, as in many spread spectrum systems, with noise power typically exceeding signal power by 19 dB in a receiver with an omnidirectional antenna. That makes it impossible to tune a conventional radio receiver to a GPS satellite. However, if an L-band antenna with a gain exceeding 30 dB is used to track a GPS satellite, the CNR will be approximately 10 dB and the BPSK C/A code can be demodulated and observed directly. In a conventional GPS C/A code receiver with an omnidirectional antenna the signal can only be observed and utilized after correlation between the received signal and an identical locally generated C/A code has been obtained. Given no

information about which satellites are visible (a cold start) a search must be conducted to find the exact frequency of the received signal and the start position of its C/A code sequence.

12.5.1 Searching for GPS Satellite Signals

GPS satellites have a high orbital velocity, 3.865 km/s, so there is significant Doppler shift in the received signals that exceeds the bandwidth of the receiver when in signal acquisition mode. A frequency search as well as a code search must be undertaken to obtain lock to a satellite's C/A code. The angle between the spacecraft velocity vector and a receiver on earth is 76.1° when a GPS satellite is at the horizon for a zenith pass, so the maximum velocity component toward a receiver is 928 m/s, giving a maximum Doppler shift in the L1 signal of $v_r/\lambda = 4.872$ kHz, ignoring the effect of earth rotation. Allowing the satellite to reach an elevation angle of 5° or 10° before it is used for a position measurement limits the Doppler shift that must be accommodated by the receiver to ±4 kHz. However, low cost GPS receivers do not have highly stable oscillators, which may drift in frequency as the receiver warms up, forcing a carrier frequency search across a wider frequency range. Rather than starting at the lowest frequency and working through to the highest frequency, a search begins with the most likely Doppler shift and receiver frequency offset, based on the last known values. The frequency step size is set by the bandwidth of the bandpass filter that follows the correlator. This is typically 500 Hz or 1 kHz. A narrow filter bandwidth improves the SNR of the signal but requires more steps in frequency in the acquisition process. Only when the receiver is set to the correct frequency is it possible to lock to the C/A codes of the GPS satellites.

A typical search and signal acquisition process follows these steps.

1. Select a receiver frequency.
2. Select a satellite (by SV number). Generate that satellite's C/A code.
3. Try all 1023 start positions of the code to attempt to correlate with the received signal.
4. Repeat step 3 several times. If correlation is obtained, repeat another three time to confirm.
5. If no correlation is obtained, try a different SV code. Repeat until all 24 operational SV codes have been tried.
6. If no correlation obtained, change to a different receiver frequency.
7. Repeat steps 2–6 until a satellite is acquired.
8. Once the first satellite is acquired, decode the navigation message to find other visible satellites and their Doppler shifts, then repeat the acquisition process, which will be much quicker given the additional information available.
9. Change to tracking mode in which Doppler shift and code rate are tracked through frequency and code lock loops.

A GPS receiver with a 1 kHz signal acquisition loop bandwidth must search up to eight receive frequencies to allow for the ±4 kHz Doppler shift of the received signal. The local oscillator in the GPS receiver consists of a numerically controlled oscillator (NCO) followed by a frequency multiplier. The NCO base frequency is the 10.23 MHz master oscillator in the receiver, which is locked to the GPS satellite clock. The receiver must also search each of 1023 C/A code positions in one chip steps, or 2046 code positions in half bit steps, for all of the 24 C/A codes. Hence the combination of Doppler frequency and code position potentially requires a search 8184 or 16 368 Doppler frequency and

code positions for each of the C/A codes. This can be a lengthy process from a cold start, which is why GPS receivers have 12 channels of digital signal processing (DSP) that can work in parallel to speed up the search process. If no signals are found from the first 12 C/A codes, some of the next 12 C/A codes must be present. Most GPS receivers can acquire four satellites within 30 seconds from a cold start.

12.5.2 Acquiring C/A Code Lock

The equation for a C/A code signal from the *i*th satellite is a BPSK modulated IF signal $s(t)$

$$s(t) = A_{\mathrm{i}} C_{\mathrm{i}}(t) D_{\mathrm{i}}(t) \sin[2\pi(f_{\mathrm{i}} + f_d)t - \varphi_i] \tag{12.6}$$

where

A_{i} is the amplitude of the received signal
$C_{\mathrm{i}}(t)$ is the Gold code sequence at 1.023 Mcps as a polar binary sequence
$D_{\mathrm{i}}(t)$ is the navigation message at 50 bps as a polar binary sequence
f_{i} is the nominal IF frequency of the received carrier in hertz
f_{d} is the Doppler shift of the received signal in hertz
φ_{i} is the phase angle of the received signal in radians

The spectrum of the signal $s(t)$ is 2.046 MHz wide between its first nulls and the carrier power level is typically 15–20 dB below the receiver noise floor. The first step in acquiring the signal is to remove the C/A code modulation. This is achieved by multiplying $s(t)$ by the C/A code sequence $C_{\mathrm{i}}(t)$. Provided we have synchronization between the received C/A code sequence and the locally generated C/A code sequence, multiplication of each of the chips in the sequence is either $+1 \times +1 = +1$ or $-1 \times -1 = +1$, which removes the chip modulation from the signal.

The resulting signal $s(t)$ then becomes $s'(t)$ where

$$s'(t) = A_{\mathrm{i}} D_{\mathrm{i}}(t) \sin[2\pi(f_{\mathrm{i}} + f_{\mathrm{d}})t - \varphi_{\mathrm{i}}] \tag{12.7}$$

The signal $s'(t)$ is at the IF frequency f_{i} Hz, offset by the Doppler frequency f_{d} Hz, but now has a bandwidth set by the navigation message modulation $D_i(t)$, which is a 50 Hz BPSK signal.

This is the correlation process that removes the C/A code modulation from the received signal leaving only the BPSK modulation of the navigation message, and allows the IF signal to be passed through a band pass filter (BPF) with a nominal 50 Hz bandwidth. The relative noise power passed by a 2.046 MHz BPF and a 100 Hz BPF (using null to null signal bandwidths) is at a ratio of 2 046 000/100 = 20 460 or 43.1 dB. The correlation process raises the SNR of the received signal from a typical −19 dB CNR to +24.1 dB SNR. However, we cannot readily acquire the signal in a 100 Hz bandwidth because the Doppler shift is an unknown at the beginning of the acquisition process. Typically, for acquisition purposes, the BPF after the correlator is set to 1000 Hz, so that eight frequency searches are needed to cover the ±4 kHz of possible Doppler shift, or 500 Hz if the approximate Doppler shift is known. In a 1000 Hz bandwidth, the SNR of the BPSK IF signal $s'(t)$ is 30.1 dB higher than the received signal $s(t)$, typically at 11.1 dB, which is sufficient to enable the code tracking loop to lock to the received C/A code. Once lock is achieved, the receiver switches to a tracking mode in which the bandwidth of the BPF

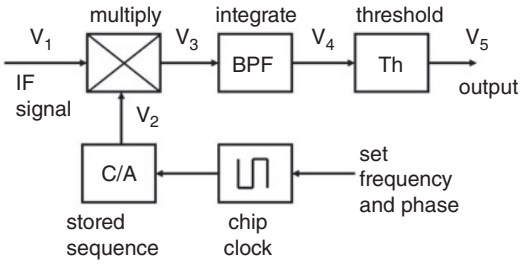


Figure 12.7 Single channel correlator for C/A code acquisition. The BPF has a bandwidth of 1 kHz during the acquisition process and is narrowed to 50 Hz once the signal is acquired. The C/A code sequence generator is stepped through all the GPS C/A codes until lock is established.

after the correlator is narrowed to 50 Hz and both the receiver local oscillator and code clock rate oscillator are driven by tracking loops.

12.5.3 C/A Code Correlator

Figure 12.7 is a simplified block diagram of a single channel correlator. The IF signal V1 is BPSK modulated by a 1.023 Mcps C/A code sequence and the signal is buried well below receiver noise. The signal V1 is multiplied by the output V2 of a C/A code generator, which produces the selected code sequence. The C/A code is 1023 chips long and repeats every millisecond. When the locally generated C/A code matches the received C/A code and is correctly synchronized, the output of the multiplier, V3, is a sine wave at the IF frequency with the C/A code removed. The signal still has BPSK modulation by the navigation message at 50 Hz. The threshold detector can be as simple as an envelope detector, which produces a constant DC output V5 once the correlator is locked to the received signal. During acquisition the locally generated C/A code is stepped in one chip increments through the entire 1023 chip sequence until the voltage V5 indicates that the C/A code has been detected, at which point the stepping of the code is stopped and the channel changes to a locked state.

The correlation process is illustrated in Figures 12.8 and 12.9 with a nominal IF frequency of 2.0 MHz. In Figure 12.8, the locally generated C/A code V2 exactly matches the BPSK modulated C/A code present in the IF signal V1. If the signal V1 were viewed on an oscilloscope it would appear to be white noise because the signal voltage V1 has an rms value one tenth of the rms noise voltage. The output of the multiplier V3 is the product of V1 and V2. Each time the IF BPSK signal V1 has a phase reversal, the locally generated C/A code signal V2 changes from −V to +V, or +V to −V, exactly in step with the IF signal modulation. The result is that the multiplier output is a sine wave at the IF frequency without C/A code modulation, for the 1.00 ms duration of the C/A code sequence. The process repeats each millisecond for 20 ms, or multiples of 20 ms, until the navigation message changes the polarity of the sine wave. The spectra of the V1 IF signal and the V3 multiplier output are illustrated in Figure 12.8. The IF signal has a null to null bandwidth of 2.046 MHz, corresponding to BPSK modulation by a 1.023 Mcps chip sequence. The V3 output of the multiplier has a null to null bandwidth of 100 Hz corresponding to the 50 Hz BPSK modulation of the navigation message. The frequency of the V3 signal may be shifted from the nominal IF center frequency by up to 4 kHz of Doppler shift caused by satellite motion; in Figure 12.8 there is a shift of just over 1 kHz, which keeps the narrow spectrum of the V3 signal outside the pass band of the 1 kHz acquisition filter. The correlator will not lock up under the conditions illustrated in Figure 12.8 until the local oscillator of the GPS receiver is stepped by 1 kHz to bring the V3 signal into the pass band of the acquisition filter.

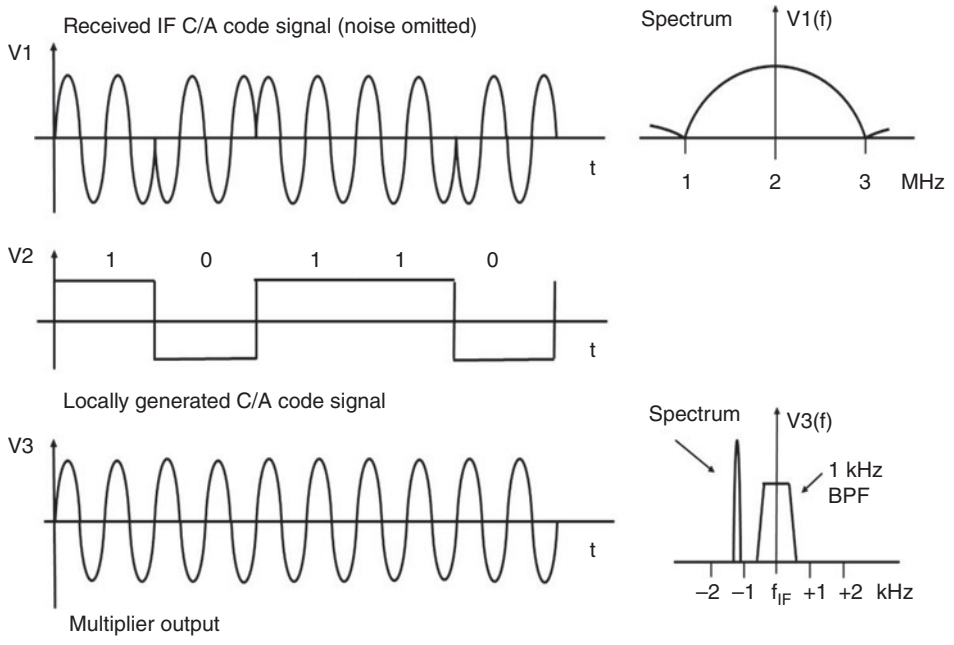


Figure 12.8 Illustration of the correlation process when the locally generated C/A code V2 has the correct timing. The 2 MHz wide C/A code BPSK signal V1 becomes a 50 Hz wide BPSK navigation message signal V3 at the output of the multiplier. However, the Doppler frequency offset is not correctly set, so the signal V3 lies outside the bandwidth of the 1 kHz acquisition band pass filter. A 1 kHz change in the Doppler frequency offset will bring the signal V3 into the pass band of the acquisition filter.

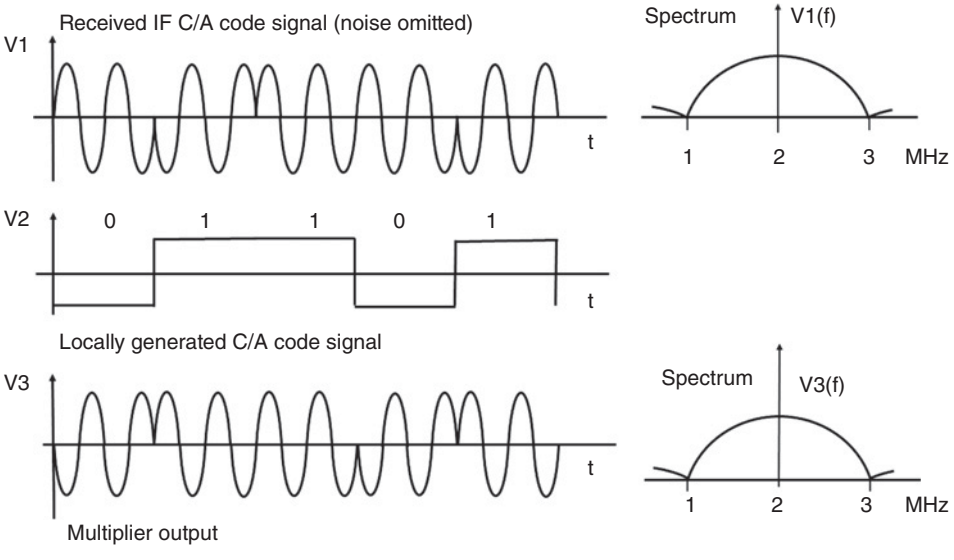


Figure 12.9 Illustration of the correlation process when the locally generated C/A code V2 has incorrect timing. The signal V3 at the output of the multiplier is a 2 MHz wide C/A code BPSK signal so there is no significant output from the 1 kHz BPF.

As noted earlier, it is very difficult to make a bandpass filter with a bandwidth of 1 kHz at a center frequency of 2 MHz, as this requires a Q factor of 2000, and the filter needs to be implemented digitally. Instead, a phase locked loop (PLL) is used, which has the characteristic of a narrow bandpass filter centered at a specific frequency.

Figure 12.9 illustrates the situation when the locally generated C/A code V2 is not synchronized to the received signal V1. The signal V2 may be the correct C/A code and only one chip away from the correct timing, or it could be the wrong C/A code, but in either case the output of the multiplier has BPSK modulation present at the 1.023 Mcps chip rate. The spectrum of the V3 signal is the same as the V1 signal and very little energy passes through the 1 kHz BPF. If the locally generated C/A code does not match any of the codes present in the received signal, the result will be the same – no output from the BPF. In most GPS receiver integrated circuits there are 12 parallel channels, each identical to the form illustrated in Figure 12.7, and each running a different C/A code sequence corresponding to the transmissions from all visible satellites. Some of the channels will have no output because there may be fewer than 12 GPS satellites in view, or there may be trees or buildings blocking the view of part of the sky. Satellites below 10° elevation angle may have signals that are too weak for reliable correlation.

12.5.4 Phase Locked Loop

The phase locked loop (also called a phase lock loop and abbreviated PLL) is a widely used circuit in communication and radio systems, as well as in the control of the speed of motors. In a GPS receiver the PLL can be used as the narrow BPF in Figure 10.7, centered on the IF frequency of the GPS C/A code receiver. The simplified circuit diagram for an analog PLL is illustrated in Figure 12.10. In a GPS receiver the PLL can be implemented using DSP routines.

The PLL will lock to an input signal V_1 when the input signal frequency is equal to the frequency of the voltage controlled oscillator (VCO) output V_2. Let V_1 be a simple sine wave given by

$$V_1(t) = V\cos\,\omega t \tag{12.8}$$

The output V_2 of the VCO is typically a square wave, but we need only consider the fundamental frequency. We will make the oscillator run at the same frequency as the input signal but with a phase difference φ and unity magnitude.

$$V_2(t) = \cos(\omega t + \varphi) \tag{12.9}$$

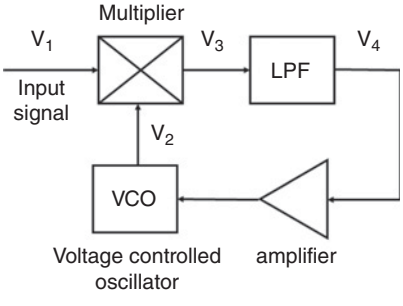


Figure 12.10 Simplified phase locked loop. LPF, low pass filter.

The multiplier has an output V_3 that is the product of V_1 and V_2.

$$V_3(t) = V\cos\omega t \times \cos(\omega t + \varphi) = \frac{1}{2}\mathrm{V}\cos(2\omega t + \varphi) - \frac{1}{2}\mathrm{V}\cos\varphi \tag{12.10}$$

The low pass filter (LPF) that follows the multiplier removes the double frequency portion of V_3 such that we have a DC signal V_4

$$V_4(t) = -\frac{1}{2}\mathrm{V}\cos\varphi \tag{12.11}$$

The amplifier adjusts the voltage level of the signal V_4 to match the input drive needed by the VCO. The loop will lock when the voltage V_4 is zero, which happens when φ is 90°. As the input frequency changes, V_4 will becomes positive or negative and drive the VCO to track the phase, and hence changes in frequency, of the input signal. When φ approaches 0° or 180° noise that accompanies the V_4 voltage will cause the VCO phase to go beyond 0° or 180°; this causes the loop to lose lock.

If we use a PLL as the narrow BPF in a GPS C/A code receiver, the input signal V_1 is the output of the correlator illustrated in Figure 12.8, which is a sine wave at the IF frequency of the GPS receiver with 50 Hz BPSK modulation by the navigation signal and a great deal of added noise in a 2 MHz bandwidth. The LPF in the PLL reduces the noise power in proportion to its bandwidth. If we set the LPF bandwidth to 500 Hz, the noise power at the LPF output is reduced by a factor of 2000, or 33 dB. (There are two noise sidebands present at the LPF output, which is why the reduction in noise power is not 36 dB.)

A PLL can successfully lock to an input signal if the SNR in the control loop is above 6 dB, but for reliable locking a SNR of 10 dB is desired. With the previously noted figures for CNR = −19 dB in a GPS C/A code receiver, the SNR of the V_4 signal should be +14 dB and the PLL will retain lock on the IF signal at its input. The loop bandwidth of 500 Hz is sufficient to allow the navigation signal to pass through the LPF and control the VCO. When the BPSK navigation signal reverses its phase, the PLL will track the change, so voltage V_4 is a demodulated form of the navigation signal.

All PLLs have two characteristics that define their operation: *capture range* and *lock range*. Capture range is the range of frequencies over which the PLL will lock to an input signal; it is typically a little larger than the LPF bandwidth either side of the free running frequency of the VCO. With a LPF bandwidth of 500 Hz the PLL cannot capture the input signal if it is much more than 500 Hz away from the VCO frequency, hence the need to step the receiver in 1 kHz steps in the acquisition phase to account for Doppler shift of the received signal. The lock range of a PLL is the range over which the input frequency can vary, once the loop is in lock, before unlocking occurs. Lock range is always larger than capture range, typically about twice the capture range. For more information on PLLs, the reader should consult reference texts on communication systems (e.g., Couch 2007; Haykin 2001). For more extensive treatment of PLLs, entire texts are devoted to the topic (Gardner 2005; Egan 1998).

12.5.5 Non-Coherent Delay Lock Loop

One of the simpler ways in which the acquisition process can be implemented is with *a non-coherent delay lock loop*. Figure 12.11 is a simplified diagram of a non-coherent delay lock loop in functional block form. The IF signal is digitized by two ADCs to generate I and Q channels and all signal processing is performed in a dedicated integrated

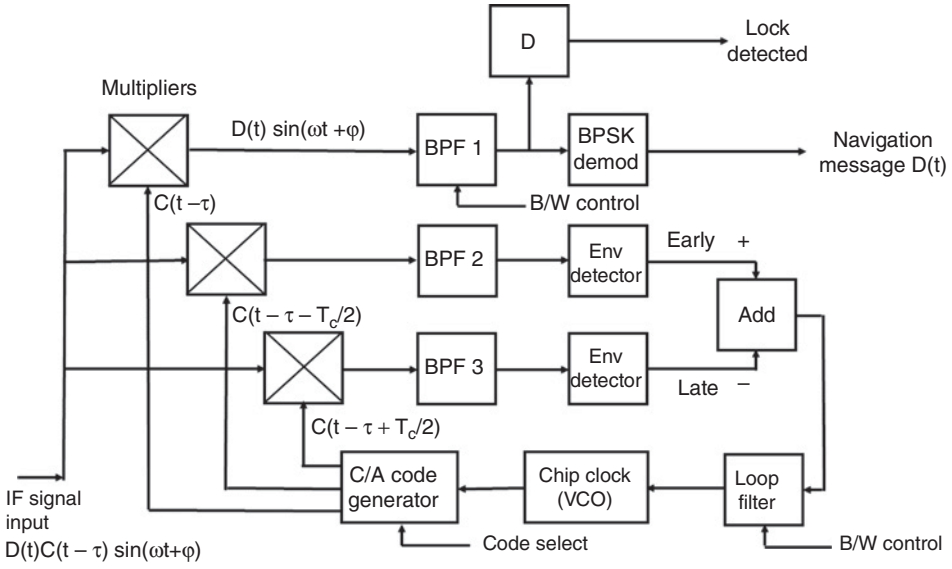


Figure 12.11 Non-coherent delay lock loop and navigation message recovery. The IF input signal has two BPSK modulations: the C/A code and the navigation message. When the loop is locked with the correct C/A code, multiplying the IF signal by the local generated C/A code removes the code modulation. The top channel is the punctual channel, the lower two channels run one half chip ahead (early) and one half chip behind (late) the punctual channel. BPF1 and BPF2 have bandwidths of 1000 Hz; their outputs are routed to envelope detectors that output the magnitude of the signal, but with opposite sign. Adding the early and late voltages indicates whether the C/A code is aligned correctly. The output of the adder drives the chip clock VCO to move the C/A code into alignment. The punctual channel drives a BPSK modulator to recover the navigation message. The punctual channel also has an envelope detector (D) that serves as a lock indicator. Once lock is achieved, the bandwidth of BPF1 can be narrowed to 50 Hz and the loop filter bandwidth can be reduced to a few hertz to ensure accurate tracking of the C/A code timing.

circuit. Most GPS ICs have 12 parallel channels that can track all the visible GPS satellites simultaneously.

The explanation of the signal processing techniques used in GPS receivers that follows is based on block diagrams that are implemented digitally. The function of the non-coherent delay lock loop illustrated in Figure 12.11 is to set the frequency and phase of the VCO that forms the local chip clock to align the received C/A code chip transitions precisely with the locally generated chip sequence transitions. The loop is called non-coherent because it uses the magnitude of the signal at the output of the C/A code correlator, obtained by squaring the I and Q signal voltages and then adding them. The result is a voltage that is always positive and proportional to the cross correlation of the received signal and the locally generated C/A code sequence. A coherent delay lock loop makes use of the amplitude and phase information in the correlator outputs by adding the I and Q voltages without squaring, and can therefore have an output that is both positive and negative.

The leading edge at the start of the C/A chip sequence generated in the receiver marks the time of arrival of the chip sequence from the satellite and is used to calculate pseudo range. Ideally, the alignment between the received sequence and the locally generated sequence should be within 10 ns, but noise in the tracking loop signals will cause the

timing to jitter. There are 1000 C/A code sequence received each second; by averaging the time of the start of each C/A code sequence over one second the required accuracy can be obtained.

The delay lock loop illustrated in Figure 12.11 has three paths: punctual, early (half chip ahead), and late (half chip behind). The punctual path corresponds to multiplication of the received signal at the IF frequency after correction for Doppler shift, by the correctly timed C/A code for the correct satellite. The early and late paths are used to steer the C/A code clock into phase with the received signal to maximize the output of the punctual channel. Doppler shift affects the C/A code clock much less than the IF carrier, with a maximum shift of 0.0026 Hz, but if left uncorrected will eventually cause the locally generated C/A code to lose synchronization with the received signal. A VCO is used to generate the code clock so that a voltage derived from the sum of the early and late channel outputs can be used to keep the VCO at the correct frequency and phase as the Doppler shift of the satellite changes over time. The maximum duration that any one GPS satellite remains in view is about three hours, so the Doppler shift of the received signal changes quite slowly, at less than 1 Hz per second (Grewal et al. 2013). A frequency lock loop is used to keep the receiver local oscillator at the correct frequency to compensate for Doppler shift.

The cross correlation of two rectangular pulses in time yields a result that is a triangle with its maximum when the pulses are coincident and zero when there is one pulse period difference. This is illustrated in Figure 12.12, with the outputs for the punctual and early and late channels in Figure 12.11 shown. Since the early and late channel timing is offset by half a chip, the cross correlation results in half the maximum value at the output of the envelope detector. The early and late outputs from the envelope detectors are added in opposite senses and used to steer the phase of the C/A code clock, as shown in Figure 12.12, with a positive voltage from the early channel and a negative voltage from the late channel. When the delay lock loop is correctly locked, as in Figure 12.12a, these two voltages will be equal and opposite, resulting in zero input to the chip clock. If the locally generated C/A code starts to slip in time and becomes late relative to the received signal, as in Figure 12.12b, the voltage from the late channel will increase, the

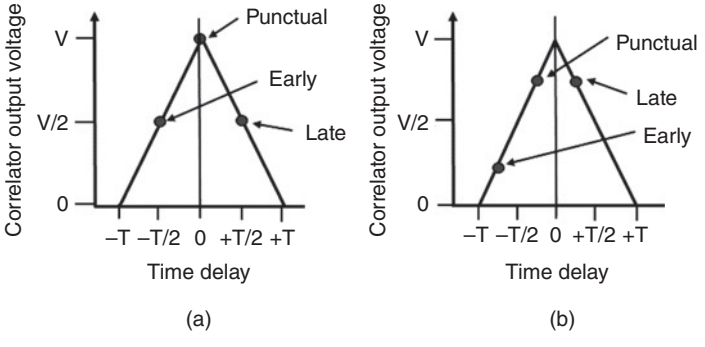


Figure 12.12 Delay lock loop correlator outputs corresponding to the punctual, early, and late channels in Figure 12.11. (a) Loop is correctly synchronized. Early and late voltages cancel and input to the VCO is zero. (b) Local C/A code is one quarter chip period delayed. Output from the late channel is much larger than the output from the early channel causing a negative voltage at the input to the VCO which will steer the loop back into synchronization. T, chip period.

voltage from the early channel will decrease, and the input to the VCO will move the local C/A code earlier in time to regain its correct synchronization.

The locally generated carrier that is used to demodulate the *C(t)* signal must be Doppler shifted to match the Doppler offset of the received signal, and modulated with the correct C/A code sequence, starting at the correct time. The correct Doppler shift, code sequence, and start time are all unknown when the receiver is first switched on. The signal is buried below the noise, so it is not possible to determine the correct parameters by direct analysis of the received signal. The receiver must therefore be designed to search all possible Doppler shifts, code sequences, and code start times until an output is obtained from the correlator indicating that a satellite signal has been found. Once one GPS satellite signal has been found, information contained in the navigation message can be used to steer the receiver to the parameters needed to acquire the other visible satellites. If the receiver is turned off and then turned on again, the microprocessor memory has the last known satellite configuration stored, and can derive expected signal parameters by allowing for the time for which the receiver was switched off.

The output of the C/A code correlator with Doppler corrected IF frequency for the satellite signal with code number i is

$$x(t) = A_{\mathrm{i}} R(\tau_{\mathrm{i}} - \tau) D_{\mathrm{i}}(t) \sin(2\pi f_{\mathrm{i}} + \varphi) + n(t) \tag{12.12}$$

where $R(\tau_{\mathrm{i}} - \tau)$ is the autocorrelation function of the wanted code number i, and $n(t)$ is the output from cross correlation with all other codes and white noise from the receiver. When the correlator loop is locked the punctual output has $R = 1$.

The time shift $(\tau_{\mathrm{i}} - \tau)$ to the correlation peak is the wanted measurement that provides the pseudo range to the satellite. The output of the correlator is a despread signal at baseband, which is modulated with the 50 bps navigation message. With the C/A code removed by the correlation process, it is straight forward to demodulate the navigation message $D_{\mathrm{i}}(t)$. The IF carrier can be recovered with a special type of PLL called a Costas loop. A Costas Loop compensates for the arbitrary phase of the received signal. The IF carrier signal is limited to remove any amplitude variations, or AGC is applied to the LNA and/or the IF amplifier, which sets $A_i = 1$. The receiver noise $n'(t)$ is restricted to the bandwidth of the LPF in the phase lock loop. Then the IF signal is $y(t)$ where

$$y(t) = D_{\mathrm{i}}(t) \sin(2\pi f_{\mathrm{i}} + \varphi) + n'(t) \tag{12.13}$$

The navigation message D(t) is recovered by multiplying the IF signal $y(t)$ by a reference carrier $\sin(2\,\pi\, f\, t)$, and low pass filtering to obtain the 50 bps signal. The reference carrier for the BPSK demodulator can be derived from the output of the Costas loop. The noise power in the PLL bandwidth is much smaller than the BPSK signal $D_{\mathrm{i}}(t)\ \sin\ (2\pi f_i, t)$, so navigation message is retrieved correctly. The SNR of the signal $y(t)$ is at least 23 dB, so there will be no bit errors. Even if a bit error occurs in the navigation message, it is removed when the next message is received 30 seconds later.

Figure 12.13 shows a Costas loop, which is often used as the demodulator for low speed BPSK signals such as the 50 bps GPS navigation message. The loop has an I channel and a Q channel driven by a VCO. The VCO frequency is set by the sum of the outputs from the I and Q channel detectors, which steers the VCO phase such that the I channel is in phase with the signal.

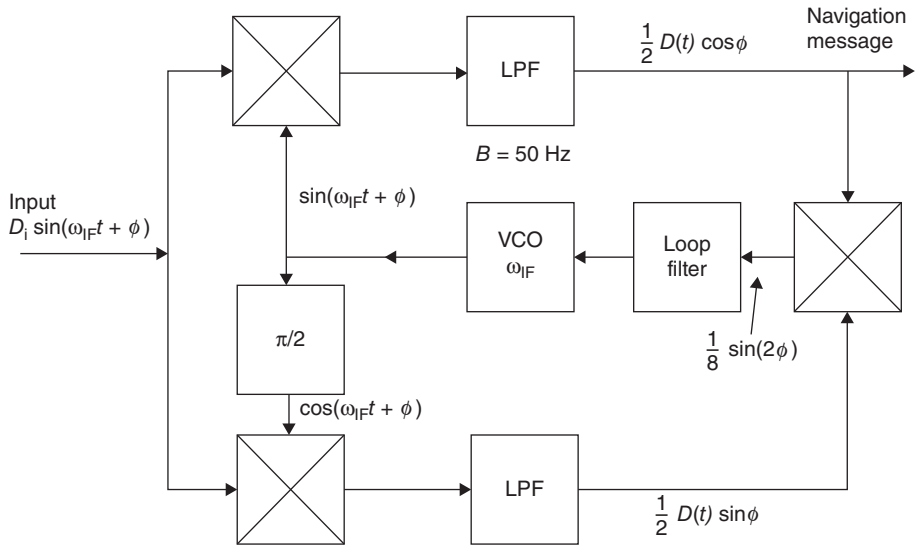


Figure 12.13 Costas loop used to demodulate the 50 Hz BPSK navigation message. LPF, low pass filter; VCO, voltage controlled oscillator.

12.6 GPS Signal Levels

Earlier in this chapter the C/A code received power level in a GPS receiver with an omnidirectional was given as −160 dBW. This figure is widely used in the literature of GPS, and derives from US government documents published around 1995 when the GPS system was declared fully operational with its first generation of satellites. A number of assumptions went into the calculation which are no longer relevant to most GPS receivers, and typical signal levels from Block II GPS satellites were found to be higher than −160 dBW.

12.6.1 Calculation of Receiver Signal Power

The assumptions made for early C/A code receivers included a linearly polarized antenna with vertical orientation and a 2 dB ionospheric and atmospheric loss. GPS satellites transmit RF signals with right hand circular polarization (RHCP). A circular polarized EM wave can be described as the sum of two equal magnitude linearly polarized waves in phase quadrature. If we describe the RHCP wave as two equal magnitude waves with vertical and horizontal polarization and have a receiving antenna that is vertically polarized, only the vertically polarized component of the RHCP wave will be received and there is a loss of 3 dB in signal strength. The early GPS link budget allocated a 3.5 dB loss for a vertical monopole antenna. A vertical monopole has a radiation pattern that is a maximum of 3 dB in the horizontal plane and falls to zero at the zenith. At an elevation angle of 60° the gain of the monopole has fallen by 3 dB relative to its maximum value (GPS signal specification 1995).

GPS satellites can appear anywhere in the visible hemisphere, but are more often at lower elevation angles. The area of a spherical cap is given by

$$A = 2\pi r^2(1 - \cos\theta) \tag{12.14}$$

where θ is the half angle of a cone with its axis vertical. For an elevation angle of 60°, $\theta = 30°$ and the area of the cap is 13.4% of the total area of the hemisphere. This means that GPS satellites can be found below an elevation angle of 60° for 86.6% of the time. The early GPS link budget was based on a GPS satellite at a low elevation angle and assumed 0 dB gain for the receiving antenna with the polarization loss canceling the antenna gain. The link budget also assumed a 2 dB loss as the L1 signal traversed the ionosphere and earth's atmosphere, which is much larger than the typical case.

Loss in the atmosphere for a 10° elevation angle is less than 0.16 dB at 1575 MHz in clear air. Rain causes little attenuation at 1575 GHz because the wavelength (0.19 m) is much larger than the largest rain drops, which do not usually exceed 6 mm. At low elevation angles the total attenuation in clear air plus heavy rain is unlikely to exceed 0.2 dB. The ionosphere is a region located between 100 and 1000 km above the earth's surface in which there are free electrons. The highest electron content is in the F layer, between 150 and 800 km. The ionosphere has two important effects on GPS signals: it causes phase changes along the path from the satellite to the GPS receiver, which causes errors in pseudo range calculations, and scintillations can be present that can cause the GPS receiver to lose lock on the signal. We will consider the phase shifting effects later and concentrate on scintillations that need to be taken into account in the link budget.

The ionosphere is largely unpredictable in its effect on radio signals. Sun spot activity and magnetic storms can cause huge fluctuations in electron content, and their occurrence is not well predicted. When the electron content of the ionosphere is high, electromagnetic waves can be scattered away from the path to the receiver resulting in a large drop in received power. L1 signal level drops in excess of 20 dB have been reported at equatorial latitudes at times of peak sunspot activity, but for mid-latitude locations, which includes all of the United States and Europe, scintillations rarely exceed ±2 dB (Ionospheric scintillation 2012). The inclusion of a 2 dB margin for ionospheric and atmospheric effects is adequate for all but a few hours per year in middle latitudes, but in the tropics and far north and south regions, greater outage time should be expected.

Block II satellites have an antenna beam that is slightly shaped to be 2.5 dB down at the edge of the earth coverage, relative to on axis gain. Block III satellites have beams that have 1.5 dB additional gain at the edge of earth coverage, compensating for the additional path loss at low elevation angles. From the GPS satellite altitude of 20 183 km, the earth subtends 30°. The expected gain for an antenna with a 30° beam is 15.2 dB, but the effect of shaping the beam will lower the on axis gain. The path length for a satellite with elevation angle of 10° is 24 700 km, with a path loss of 184.3 dB at 1575 MHz (L1 frequency). A zenith path has a length of 20 183 km, and a free space path loss of 182.5 dB, although satellites are much less likely to have elevation angles close to 90° than 10°. The start of life transmit power for Block II GPS satellites was 14.1 dBW with nominal on axis satellite antenna gain of 15.2 dB and 13.1 dB at elevation angles above 10°. End of life power was set at 13.7 dBW, giving a typical effective isotropically radiated power (EIRP) of 26.8 dBW for satellites seen at lower elevation angles. P-code transmit power for the L1 frequency is 3 dB lower than for C/A code signals, and 6 dB lower for L2 signals. With a linearly polarized antenna mounted vertically and 3.6 dB polarization loss, the gain of the antenna is set to −0.6 dB. Table 12.1 is based on the above parameters for an early Block II satellite, a linearly polarized receiving antenna, and a satellite beam with minimal shaping.

Later GPS satellites, and particularly Block IIR-M satellites launched after 2010 had modified antenna patterns that increased edge of beam signal by 1.5 dB and also increased L1 C/A code transmitter power by 1.5–2 dB. After 2000 most GPS receivers

Table 12.1 Signal power for early GPS satellites and receiver with linearly polarized antenna

Frequency	L1 1575.42 MHz	L2 1227.6 MHz	
Signal	C/A code	P-code	P-code
EIRP for satellite at 10° EL	26.8 dBW	23.8 dBW	19.7 dBW
Path loss for 10° elevation angle	184.2 dB	184.2 dB	182.1 dB
Atmosphere and ionosphere loss	2.0 dB	2.0 dB	2.0 dB
Receive antenna gain	−0.6 dB	−0.6 dB	−0.6 dB
Received power	−160.0 dBW	−163.0 dBW	−166.0 dBW

had an RHCP antenna that avoided the polarization mismatch loss of 3.6 dB that is included in the antenna gain in Table 12.1. The received signal power for the L1 C/A code increased substantially with these changes. Table 12.2 shows a link budget representative of a typical GPS handheld receiver and Block IIIR-M satellite at an elevation angle of 10°. With a RHCP patch antenna in the GPS receiver, a satellite at 30° elevation angle would deliver a received power 1.5–2 dB higher than indicated in Table 12.2.

The increase in receiver power and improvements in receiver design by the use of adaptive band pass filtering resulted in much improved SNR in the tracking loop of GPS receivers, with values of up to 30 dB under best case conditions. GPS receivers can be operated successfully inside houses where there are a sufficient number of windows. However, operation inside office blocks and tall buildings is less reliable because of the high attenuation through multiple floors and walls. Indoor navigation remains a challenging problem.

12.6.2 Receiver CNR and SNR

A handheld GPS receiver is often operated in an enclosed environment such as an automobile or inside a house. The antenna is surrounded by an environment with a typical physical temperature of 300° K that radiates noise into the GPS receiving antenna, resulting in a high antenna noise temperature. By contrast, an antenna mounted on a ground plane such as the roof of an automobile or the fuselage of an aircraft will pick up little radiation from its environment and should have an antenna temperature below 100 K. A bandpass filter is usually inserted between the antenna and the LNA to prevent the LNA being overloaded by interference and to act as an image rejection filter.

Table 12.2 Signal power for block IIR-M GPS satellite and L1 receiver with RHCP antenna

Frequency	L1 1575.42 MHz
Signal	C/A code
EIRP for satellite at 10° EL	29.8 dBW
Path loss for 10° elevation angle	184.2 dB
Atmosphere and ionosphere loss	1.0 dB
Receive antenna gain	0 dB
Received power	−155.4 dBW

The LNA is mounted immediately below the antenna, which is simple to do with a patch antenna that has a metallic ground plane, and DC power for the LNA is fed via the cable that connects to the GPS receiver, in exactly the same way as with direct broadcast satellite television installations. With additional loss from the BPF, a system noise temperatures of 500 K is often quoted as a typical value for a low cost C/A code receiver. With the antenna mounted on a ground plane a typical system noise temperature of 250 K is achievable.

The low CNR of the spread spectrum signal in a GPS receiver is converted to a usable SNR by correlation of the code sequences, which adds a despreading (processing) gain to the CNR. The theoretical processing gain of a DSSS signal is equal to the ratio of the chip rate to the bit rate in the spreading sequence. For the C/A code transmitted at 1.023 Mbps and a 1 ms correlation time, the theoretical processing gain is 1023, or 30.1 dB. The corresponding processing gain for the P-code is 40.1 dB. In this text we distinguish between SNR of the RF and baseband signals by using CNR for spread spectrum RF and IF signals in the GPS receiver, and SNR for baseband signals after correlation. Most of the literature on GPS refers to all signal to noise ratios as SNR, which can be confusing.

The GPS receiver can pick up signals from 9 to 13 satellites at the same time. The RF energy from the satellite spread spectrum transmissions adds to the noise in the receiver as an interference term, *I*. For simplicity, in the following analysis we will assume that there are 10 GPS satellites visible, that there are nine interfering satellites generating random signals (noise) out of which the receiver must extract the tenth signal, and that all the received signals are of equal strength. The signals from interfering satellites are treated as random noise because the Gold codes that they transmit have very low cross correlation with the code from the wanted satellite. Noise has zero cross correlation with the wanted signal, and the Gold codes used by GPS satellites are chosen because they closely approximate noise.

Nine interfering GPS satellites represents a worst case; in practice the signal strengths also vary depending on the elevation angle of the satellite and the antenna pattern at the receiver. The worst case is actually when a weak signal from a satellite at a low elevation angle must be extracted from stronger signals from satellites at higher elevation angles. GPS receivers automatically select the strongest signals for processing so that the worst case can be avoided, but if the sky is partially blocked by obstructions, a weak signal may have to be used. We will use the later figure of −155.4 dBW in Table 12.2 for the signal power received by a typical GPS receiver from a Block IIR-M GPS satellite.

The interference from nine C/A code spread spectrum signals of equal power is given by the sum of the received power (in watts) from each satellite. A power level of −155.4 dBW is 2.9×10^{-16} W, so the interfering power is *I* watts where

$$I = 9 \times 2.9 \times 10^{-16} = 2.62 \times 10^{15}\text{W or} - 145.8\ \text{dBW} \tag{12.15}$$

The thermal noise power, *N*, in a noise bandwidth of 2 MHz for a receiver noise temperature of 500 K is $k\ T_s\ B_n$ watts, where

$$N = kT_sB_n = 1.38 \times 10^{-14}\text{W or} - 138.6\ \text{dBW} \tag{12.16}$$

The noise and interference powers must be added in watts, not in decibels

$$N + I = (1.38 + 0.26) \times 10^{-14} = 1.64 \times 10^{-14}\text{W or} - 137.9\ \text{dBW} \tag{12.17}$$

Table 12.3 CNR, noise, and interference budget for L1 and L2 carriers

Carrier	L1	L1	L2
Code	C/A code	P-code	P-code
System noise temp T_s 500 K	27.0 dBK	27.0 dBK	27.0 dBK
Noise bandwidth B_n	63.0 dBHz	73.0 dBHz	73.0 dBHz
Thermal noise power N	−138.6 dBW	−128.6 dBW	−128.6 dBW
Interference power I (nine satellites)	−145.8 dBW	−148.8 dBW	−152.9 dBW
Noise plus interference $N + I$	−140.7 dBW	−128.7 dBW	−128.7 dBW
Receiver power P_r	−155.4 dB	−155.4 dB	−161.4 dB
$C/(N + I)$	−14.7 dB	−26.7 dB	−32.7 dB
Processing gain G_{proc}	30.1 dB	40.1 dB	40.1 dB
SNR in acquisition mode	15.4 dB	13.4 dB	8.1 dB
SNR in 50 Hz bandwidth	28.4 dB	26.4 dB	15.1 dB

Hence the worst case CNR for one C/A code signal in this scenario is

$$\frac{C}{N+I} = -155.4 + 137.9 = -17.5 \text{ dB} \tag{12.18}$$

Similar analysis yields the C/(N + I) values for the two P-code signals. CNR and SNR values are given for the typical scenario analyzed here in Table 12.3. CNR and SNR values are 4.7 dB higher than the original published figures for early GPS satellites and receivers.

Note that thermal noise is the major factor in setting C/(N + I), since in the worst case of interference caused by nine visible satellites, all received at maximum power, the interference power level is 7.2 dB below the thermal noise power. The CNR at the receiver is 0.7 dB lower when the interference from the nine visible satellites is included. A more realistic scenario would have four satellites at the maximum receive power level and the remainder at a lower level, since GPS satellites orbit in constellations of four, with one constellation always visible to improve the accuracy of position location measurements. Thus we would expect less than 0.7 dB degradation in the CNR due to interference by other satellites' CDMA signals for almost all of the time.

12.7 GPS Navigation Message

A key feature of the GPS C/A code is the navigation message. The navigation message contains a large amount of information that is needed by GPS receivers to acquire GPS satellite signals and calculate the position of the receiver. The navigation message is sent at 50 bps by BPSK modulation of the C/A code. Effectively, 20 C/A code sequences form one navigation message bit. The phase of the 20 sequences is inverted between the 1 and 0 bits of the message by modulo two addition of the navigation message data with the C/A code sequence. The navigation signal is extracted by a 50 bps BPSK demodulator that follows the C/A code correlator. The narrow bandwidth of the navigation message

Table 12.4 GPS Navigation message – subframe details

Subframe #1	TLM	HOW	Health of satellite; clock correction data
Subframe #2	TLM	HOW	Ephemeris data for this satellite
Subframe #3	TLM	HOW	Ephemeris data for this satellite
Subframe #4	TLM	HOW	Almanac; ephemeris data for satellites with PRN# ≥ 25; Ionospheric correction data
Subframe #5	TLM	HOW	Almanac; ephemeris data for satellites with PRN# 1–24; satellite health for all satellites; estimate of user range error

ensures a high SNR at the demodulator input and correspondingly low probability of bit errors in the navigation message.

The complete navigation message is 1500 bits, sent as a 30 second frame with five subframes. However, some information is contained in a sequence of frames and the complete data set requires 12.5 minutes for transmission. The most important elements of the message are repeated in subframes 1, 2, and 3 every 30 seconds. The subframes contain the satellite's clock time data, orbital ephemeris for the satellite and its neighbors, and various correction factors. Some important details of the subframes are given in Table 12.4 (Grewal et al. 2013).

Subframes #1, #2, and #3 repeat all data every 30 seconds.

Subframe #4 and #5 repeat every 30 seconds, but transmission of the full data set requires 25 subframes over a period of 12.5 minutes.

TLM is the telemetry word with an 8 bit flag marking the start of the subframe.

HOW is the handover word allowing authorized users to synchronize to the P-code from the C/A code in the GPS receiver.

The calculation of position in a GPS receiver requires very accurate knowledge of the location of the satellite at the time that the measurements of pseudo ranges are made. If the pseudo range is measured to an accuracy of 2.4 m, we must know the satellite position to an even greater accuracy, and that requires very accurate calculation of the GPS satellite orbits.

By comparison, the orbit of a communication satellite does not need to be known to the same level of accuracy. The GPS system uses modified WGS 84 data to define the earth's radius, Kepler's constant, and the earth's rotational rate. The WGS 84 data set also includes a very detailed description of the earth's gravitational field, which is essential for precise location of the satellites in their orbits. All of these parameters and corrections are stored in every GPS receiver and used in calculating its location.

12.8 GPS C/A Code Standard Positioning System Accuracy

Position location accuracy is defined as the probability that the measured location of a GPS receiver is within a specified distance of its true location. For example, typical accuracy for a GPS receiver using the GPS C/A code is 5 m, defined as a 2DRMS error, which means that 95% of measurements will be within 5 m of the receiver's true location. The term DRMS means the RMS error of the measured position relative to the true position of the receiver. If the measurement errors are Gaussian distributed, as if often the case, 68% of the measured position results will be within a distance of 1DRMS from the true

location and 95% of the results will be within 2DRMS of the true location. For 99.8% of measurements the accuracy will be within 3DRMS. Accuracy in GPS measurements is usually defined in terms of 2DRMS, in the horizontal or vertical plane. Accuracy is always better in the horizontal plane than the vertical plane because GPS satellites can surround the receiver in the horizontal plane, but can only be above the receiver in the vertical plane.

12.8.1 Estimating Location Accuracy

The process of estimating GPS measurement accuracy requires several steps. The basic figure is the error in a single range measurement, called a URE – user range error. Four range measurements are required to calculate position, and it is assumed that the errors add in a random fashion. The RMS range error is calculated by taking the square root of the sum of four range errors squared.

$$RMS\ error = \sqrt{(URE_1)^2 + (URE_2)^2 + (URE_3)^2 + (URE_4)^2} \tag{12.19}$$

where the subscript corresponds to the range from each of four satellites. If the range error is the same for each of the four satellites the RMS error will be two URE.

The major sources of error in a GPS receiver location measurement are: satellite clock error, ephemeris errors, ionospheric delay, tropospheric delay, receiver noise, and multipath. When a value has been assigned to each of these parameters, the URE is calculated in the same fashion as in Eq. (12.19) by taking the square root of the sum of each error contribution squared. Estimates of position error for C/A code receivers are included in the GPS standard positioning system performance standard ICD-GPS-200C issued by the US Department of Defense (ICD-GPS-200C). This is a 185 page document that describes the parameters and operation of the GPS standard positioning system in considerable detail. There have been four issues since the GPS system was declared operational in 1995, the latest being in 2001, with periodic updates through 2018.

The estimated 2DRMS accuracy of GPS SPS measurements has improved steadily over the years, from an average value of 19 m in 1995 to 3 m in 2001 for a single RF measurement without wide area augmentation. Where two RF measurements are available the accuracy is consistently better than 2 m. The accuracy that can be achieved with a GPS C/A code receiver can be found by using a range error budget. Typical values of URE are given in Table 12.5. All values are in meters.

The estimated 1DRMS position error values are twice range error, at 19 m in 1995 and 5.3 m in 2001. These values are close to the expected average error that would be observed under the operating conditions used in deriving the URE values in Table 12.4. The 2DRMS position error that should not be exceeded more than 5% of the time was estimated to be 38 m in 1995 and 10.6 m in 2001. Under best case conditions, which apply across the 48 contiguous states of the United States for much of the time, measurements show an average position error of 3 m. With WAAS augmentation, the average position error has been measured as 1.87 m (WAAS accuracy 2017).

The above analysis provides an overview of the accuracy that can be achieved under good conditions with a good quality C/A code GPS receiver. However, there are some factors that can cause accuracy to be significantly degraded. The most likely cause of larger errors is the ionosphere. As noted earlier in the chapter, the ionosphere is

Table 12.5 Estimated range error for single frequency C/A code measurements

Error contribution	1995 value (m)	2001 value (m)
Satellite clock error	3.5	1.43
Ephemeris errors	4.3	0.57
Ionospheric delay	6.4	1.3
Tropospheric delay	2.0	0.25
Receiver noise	2.4	1.4
Multipath	3.0	1.0
RMS range error	9.51	2.66

extremely variable and somewhat unpredictable. Electrons in the ionosphere causes a delay in the C/A code as the L1 signals pass through the ionosphere, corresponding to a reduction in the velocity of EM waves. Similarly, the refractive index of air is greater than unity and causes EM waves to slow down. The effect of the troposphere under clear air conditions can be calculated accurately, but clouds and rain add an uncertainty in the calculation of delay. The density of electrons varies greatly, with higher concentrations below latitudes 30°N and 30°S, and much higher concentrations during periods of sunspot activity and magnetic storms. Sunspot activity varies over an 11 year cycle; in an active year at low latitudes the 1.3 m range error attributed to the ionosphere can increase to 14 m, resulting in a URE of 14.2 m and a 1DRMS position error of 28 m. In one recorded instance in 2000, a year of peak solar activity, a monitoring station in SE Australia at 20°S latitude recorded a horizontal position error of 25 m and a vertical error of 60 m (GPS SPS performance standard 2001).

The Navigation Message contains estimates of ionospheric and tropospheric delay in subframe #4. The estimates are based on algorithms and electron density measurements that have improved with time, leading to the fivefold reduction in ionospheric range error. The navigation message also contains estimates of satellite clock error and ephemeris error, allowing C/A code receivers to adjust pseudo range measurements to compensate for the errors. Errors due to receiver noise have been reduced in the 2001 results in Table 12.4 because the signal strength is higher, as indicated in Table 12.3. Range error caused by receiver noise is proportional to the square root of SNR (Skolnik 1981). Multipath error is reduced in the 2001 analysis because the antenna of the C/A code receiver in 1995 was assumed to be a vertical monopole with maximum gain in the horizontal plane. Since multipath reflections occur primarily at low elevation angles, the use of patch antennas in later receivers with lower gain at low elevation angles reduced the error contribution from multipath. However, if a GPS receiver is operated among tall buildings or indoors, multipath error can increase substantially.

The range error introduced by the ionosphere and the troposphere can be partially removed by receiving identical signals at two different carrier frequencies. This technique is used by high precision P-code receivers. The P-code signal is transmitted on the L1 carrier at 1575.42 MHz, in phase quadrature with the C/A code signal. The P-code is also transmitted on the L2 carrier at 1227.60 MHz. Algorithms are used in the P-code receiver to calculate the net delay of the signal caused by the ionosphere and the atmosphere, and to then remove the errors from the calculated ranges. Some Block II and all

block III GPS satellites transmit a second C/A code at the L5 frequency of 1176.45 MHz, which provides improved accuracy with C/A code receivers. There is also a L2C signal at 1227.6 MHz on later satellites that serves the same purpose. Accuracy similar to WAAS augmented service is achieved with dual frequencies.

Receiver position is calculated in (x, y, z) coordinates, and the errors in x, y, and z depend on the elevation angle of satellites, the satellite geometry, and the other parameters in the error budget. The calculated position will have different levels of error in the x, y, and z directions. To account for these differences several *dilution of precision* factors (DOP) are defined. A DOP factor multiplies the basic position measurement error to give a larger error caused by the particular DOP effect.

12.8.2 Dilution of Precision – HDOP, VDOP, and GDOP

Horizontal DOP is one of the most important DOP factors for most GPS users. A typical value of HDOP is 1.5, and it is often the smallest of the DOPs. There are many DOP factors in GPS. The more important ones are horizontal dilution of precision, HDOP, vertical dilution of precision, VDOP, and geometric dilution of precision, GDOP. Other DOPs include position dilution of precision, PDOP, and time dilution of precision, TDOP. In general, VDOP and GDOP are most likely to degrade the accuracy of GPS position measurements. VDOP accounts for loss of accuracy in the vertical direction caused by the angles at which the satellites being used for the position measurement are seen in the sky. If the satellites are all close to the horizon, the angles between the satellites and the receiver are all similar and VDOP can be large. In the worst possible case, if all the satellites were at the horizon, it would be impossible to make an accurate measurement in the vertical direction. A change in range to at least one satellite must occur when the receiver is moved, otherwise the receiver cannot detect that change. If all the satellites are at the horizon, no range change occurs for small vertical movements of the receiver and consequently vertical accuracy is very poor. Similarly, if all the satellites were clustered directly overhead, HDOP would be large.

VDOP is important in aircraft position measurements, where height above the ground is a critical factor, especially when landing. C/A code receivers suffer from significant VDOP and cannot provide sufficient vertical accuracy for automated landing of aircraft. C/A code GPS receivers cannot guarantee sufficient vertical accuracy unless operated in a DGPS mode. ATC relies on aircraft altimeters, rather than GPS vertical measurements.

The GPS satellites are configured in orbit to minimize the probability that a DOP can become large, by arranging the orbits to provide clusters of four satellites with suitable spacings in the sky. However, if the receiver's view of the sky is restricted, for example by buildings, the geometry for the position calculation may not be ideal and GDOP can become large. This causes all the other DOP values to increase. Aircraft, and ships at sea always have a clear view of the sky, but automobiles often do not. C/A code receivers may revert to two dimensional measurements (x and y) using three satellites when the sky is obstructed.

12.8.3 Wide Area Augmentation System

In the *wide area augmentation system* (WAAS) developed by the FAA for aircraft flying in North America, 28 WAAS reference stations continuously monitor signals from all visible satellites in the GPS system. The stations use the P-code transmissions to make

accurate differential measurements of the pseudo range to each visible satellite. The actual position of the WAAS stations is known very accurately from prior survey data, so each WAAS station can calculate the error in the pseudo range to each visible satellite. The WAAS stations send their data to two central stations with uplinks to two GEO satellites. The central station validates the data, combines all the information to create a wide area map of pseudo range errors for each satellite, and sends a sequence of pseudo range correction data to all GPS users via the GEO satellite. The central station also determines whether any of the data is in error, and sends a warning signal called an *integrity message* to instruct aircraft not to use the GPS system, or a particular satellite, because the data are not reliable. This is an essential part of the FAA strategy for using GPS as the primary means of aircraft navigation. If the aircraft is relying on GPS information alone to determine its position, that information must have a very high reliability.

In 2018, WAAS data was transmitted by transponders on two GEO satellites, Intelsat Galaxy 15 and Telesat Anik F1-R. The transponders transmit signals at the L1 frequency using two PRN codes selected from numbers 33 to 37; C/A code GPS receivers accept the GEO signal as it is identical to a GPS satellite L1 transmission. A conventional GPS receiver with suitable software can extract the pseudo range error values from the WAAS satellite navigation message and obtain markedly improved accuracy in its position determination. Thus no hardware changes are needed to convert a GPS receiver to use WAAS data. The GEO satellite can also be used to augment GPS satellites for position measurements, since it radiates the same signal format. The calculation of pseudo range error from the P-code sequence, rather than (x, y, z) position data error from the C/A code, significantly increases the accuracy of the WAAS DGPS system. This is an essential part of the FAA strategy for using GPS as the primary means of aircraft navigation. If the aircraft is relying on GPS information alone to determine its position, that information must have a very high reliability.

The WAAS Navigation Systems Verification and Monitoring Branch analyzes the accuracy of the GPS SPS performance in quarterly GPS performance analysis (PAN) reports. These reports contain analysis performed on data collected at 28 WAAS reference stations (WAAS accuracy 2017). In an example report for the period 1 October through 31 December 2016, a magnetic storm occurred on 25 October 2016. The 2DRMS (95% of time) accuracy achieved in the three month period was 1.89 m in the horizontal plane and 3.88 m in the vertical plane. On 25 October 2016, when the magnetic storm occurred, the worst station position error had increased by less than 1 m. This suggests that the 2 dB allowance for atmospheric and ionospheric disturbances is much larger than the average value experienced in the United States.

12.9 Differential GPS

The accuracy of GPS measurements can be increased considerably by using *differential GPS* (DGPS) techniques. There are several forms of DGPS, all of which are intended to increase the accuracy of a basic GPS position measurement. A second, fixed, GPS receiver at a *reference station* is always required in a DGPS system. In the simplest forms of DGPS, a second GPS receiver at a known position continuously calculates its position using the GPS C/A code. The calculated (x, y, z) location is compared to the known location of the station and the differences in x, y, and z are sent by a radio telemetry link

to the first GPS receiver. However, this technique works well only if the two stations are close together and use the same four satellites for the position calculation.

In a more sophisticated form of DGPS, the monitoring station at a known location measures the error in pseudo range to each satellite that is visible at its location, and telemeters the error values to users in that area. This allows other GPS users to select which satellites they want to observe, and extends the area over which the DGPS system can operate. The errors in a C/A code measurement are reduced to well below 1 m using this approach.

The most accurate forms of DGPS use the relative phase of the many signals in the GPS transmissions to increase the accuracy of the timing measurements. Suppose that you count the number of cycles of the 1575.42 MHz L1 carrier wave between a satellite and a GPS receiver, and that the GPS satellites are stationary for the length of time it takes to make the count at two separate locations. The wavelength of the L1 carrier is 0.19043 m, so movement of the receiver by 0.01 m directly away from the satellite would change the phase angle of the received wave by 18.9°. If the total number of cycles between the satellite and the receiver is known, and fractional cycles are measured with a phase resolution of 20°, the true distance to the satellite can be found to 0.01 m accuracy. In principle, measurements that compare the phase angle of the received L1 carriers from several GPS satellites could therefore be used to detect receiver movements at the centimeter level. This is called *differential phase* or *kinematic* DGPS.

The obvious difficulty is that we cannot count the number of cycles of the L1 carrier between the satellite and the receiver. However, we can make phase measurements and time of arrival comparisons for various GPS signals at two different locations and resolve motion between the two locations. If one of the receivers is a fixed reference station, it is then possible to locate the second GPS receiver very accurately with respect to that fixed location.

This technique is valuable in land surveying, for example, where a reference station can be set up at a known location, such as the corner of a plot of land, and the position of the plot boundary relative to that point can be measured. The same technique can be used to find the position of an aircraft relative to an airport runway so that a precision approach path can be established.

The difficulty with DGPS phase comparison measurements is that the L1 carrier has cycles that repeat every 0.19043 m, and one cycle is identical to the next. This creates *range ambiguity,* which must be resolved by reference to the wavelengths of other signals.

The 10.23 MHz P-code transmission of the L1 carrier has a P-code chip length in space of 29.326 m, which is 154 cycles of the L1 carrier. The ambiguity of the carrier waveform can be resolved within the 29.326 m length of a P-code chip by comparison of the time of arrival of a particular cycle of the L1 carrier with the time since the start of the P-code chip. Similar ambiguity resolution for the 29 m P-code chips is possible using the length of the C/A code chip and the C/A code sequence. The length of a C/A code chip at 1.023 MHz is 293.255 m, and the length of a C/A code sequence is 293.255 km. When ambiguity resolution is applied using all of these waveforms, very small movements of the receiver can be detected and ambiguity out to 293 km can be removed. Aircraft flight paths have been tracked to an accuracy of 2 cm over distances of tens of kilometers using phase comparison DGPS techniques.

This explanation of kinematic DGPS is oversimplified, because the satellites are moving and measurements over a considerable time are required to resolve ambiguity to the

centimeter level. The P-code can be used for real time differential measurements without knowledge of P-code itself, because only a comparison of the time of arrival of the code bits is required.

12.9.1 Local Area Augmentation Systems (LAAS)

Local area augmentation has been applied to the control of moving vehicles and aircraft. One example is in agriculture, where self-driving tractors can maneuver with centimeter accuracy across fields of hundreds of acres. Tractors can be made to drive between rows of crops to plant and harvest crops, and to deliver fertilizer and pesticides with much greater efficiency than with a human operator. Automatic landing (autoland) of aircraft has also been demonstrated with cargo aircraft. Local area augmentation system (LAAS) DGPS systems have been demonstrated to achieve better than 1 m accuracy in three dimensions as an aircraft lands on a runway, with update rates sufficiently fast to control a large commercial aircraft. DGPS position data can be coupled to the aircraft autopilot so that blind landings can be made automatically in zero visibility conditions. Several demonstrations of autoland using DGPS were made in the late 1990s using Boeing 737 and 757 aircraft and operational use of such systems began in 2018 (Aviation week 2018).

Aircraft used by overnight delivery companies will likely be fitted with GPS blind landing systems first, since cargo aircraft are subject to fewer operating restrictions than passenger aircraft and overnight delivery is subject to delays when airports are closed by low visibility weather. Typically, a GPS based autoland system fitted to a large aircraft can achieve more consistent landings than a skilled pilot, so autoland may eventually become as common for landings as autopilot use is for en route operation. Weather may eventually be less of a factor in causing delays to passenger aircraft arrivals and departures. Despite the many successful demonstrations, the FAA has not approved GPS autoland for general operations. This may be because of the difficulty of guaranteeing that the system has not been spoofed, as discussed in the next section.

12.10 Denial of Service: Jamming and Spoofing

GPS signals are very weak, in the −153 to −160 dBW range for C/A code receivers using the SPS L1 frequency. A transmitter located near the GPS receiver operating at the L1 frequency and modulated with white noise can easily create an interfering signal at −116 dBW in the GPS receiver. Despite the high processing gain of the correlators, around 30 dB for a C/A receiver, the white noise will generate a random sequence of bits in the receiver that is much larger than the C/A code signals from the GPS satellites and thus prevent the receiver from making position measurements. This is called *jamming*; it is a well-known technique for denying service in any radio or radar system, and has been widely employed by governments and military operations to disrupt radio communications and create false targets in radars. In the radar and military communications world, denial of service is known as countermeasures, and attempts to defeat countermeasures are counter-countermeasures. Countermeasures against radar detection of targets are almost as old as radar itself. The two techniques have progressed hand in hand for the past 80 years and were effectively used during the D-Day invasion of France in 1944. Spoofing transmitters on the English coast were successfully used to create false ship

targets on German radars covering the Straits of Dover, to persuade the German High Command to divert defensive forces to that area and lessen defense of the invasion area of Normandy (Schofield 2008, p. 65).

Jamming and Spoofing are of particular concern to military users of GPS, as GPS provides navigation guidance for military aircraft, ships, missiles, UAVs, and drones (GPS Jamming 2012). Unmanned systems such as missiles, UAVs and drones are particularly vulnerable because there is no operator present who can observe changes in the GPS receiver output that indicate jamming or spoofing is present. The L2 signals of GPS are more difficult to spoof than the L1 civil signals, but a determined adversary can be successful (Black Sea spoofing incident 2017). Demonstrations of spoofing raised the issue to public notice in 2013, followed by increasing research to find ways to defeat spoofers (GPS Spoofing 2016; Ship navigation 2017). Cellular phones include GPS C/A code receivers that send location information back to cell towers and the computers that provide E911 location service. A spoofing device placed close to a cell phone can make the cell phone system think the user is in a different location, diverting attention from the location of an illegal activity.

12.10.1 Jamming Example

Suppose we set up a white noise jammer with an omnidirectional antenna having 0 dB gain in the horizontal plane and 100 mW transmitter at a distance of 1 km from a GPS receiver. Path loss at 1575.42 GHz for a distance of 1 km is 96.4 dB. Assuming that the GPS receiver has an antenna with gain 0 dB in the direction of the jammer, the received power in the GPS receiver is

$$P_r = -10 \text{ dBW} + 0 \text{ dB} + 0 \text{ dB} - 96.4 \text{ dB} = -104.4 \text{ dBW} \tag{12.20}$$

This is a received power level 48 dB greater than the maximum received power in a GPS C/A code receiver and will jam the receiver. The distance between the GPS receiver and the jamming transmitter can be increased to 5 km and the receiver is still rendered ineffective.

A relatively low power jammer with an omnidirectional antenna can jam all GPS receivers within 5 km. This is a useful technique for denying a hostile opponent use of GPS L1 signals in a military environment. Mounting the jammer on a drone can make it even more effective.

In the civil world, using any form of jammer is illegal; nevertheless, advertisements for GPS jammers can be found on the internet. The Federal Communications Commission (FCC) has prosecuted people in the United States for using GPS jammers. In one case, a trucker whose vehicle was equipped with a GPS receiver so that his employer could monitor the location of the truck used a GPS jammer to prevent the employer from knowing where he was. In the process, he jammed all the GPS receivers at the Newark International Airport in New Jersey. The FCC fined the driver US$31 875 and he was fired by his employer (GPS jamming 2013).

12.10.2 Spoofing

A more subtle form of denial of service called *spoofing* uses a transmitter that mimics the signals transmitted by a GPS satellite to cause the GPS receiver to indicate a false position. This is particularly dangerous to aircraft when landing or ships operating in

fog. Aircraft can crash land away from the intended runway and ships can founder on rocks if the location data is spoofed. Typically, a spoofing system tries to mimic the signal being received from one satellite that is being used by the target GPS receiver. A replica C/A signal is transmitted at a low level such that the spoofing C/A code has the correct timing in the GPS receiver. The power level is then increased until the receiver's correlator is synchronized to the spoofing signal instead of the GPS satellite. The timing of the C/A code is then altered slowly, causing the GPS receiver to generate an incorrect pseudo range, which will result in an incorrect position output. This is easy to do with a stationary or slowly moving target, but more difficult with a fast moving aircraft.

Software defined radios can be programmed to behave like GPS satellites transmitting L1 signals. GPS simulators are widely available and all the information needed to create false C/A code signals is publicly available. GPS receivers with AGC are vulnerable to spoofing by false signals. As the received power from the spoofer increases, the receiver RF and IF gain of the receiver is reduced and the real GPS signals become weaker, allowing the false signals to take over the correlators.

GPS receivers are used to coordinate timing at electrical power generating stations to synchronize the power grid and also at cell phone towers to synchronize transmissions. There is concern that these receivers may be vulnerable to jamming and spoofing. Susceptibility to jamming and spoofing can be reduced in a number of ways. The received power level of each C/A code signal can be monitored. If any signal increases unexpectedly, or goes above −150 dBW at the receiver input, some form of interference should be suspected and an alert issued.

Sudden jumps in the reported position of the GPS receiver are an indication of spoofing. Antennas mounted on a ground plane have low gain in the horizontal plane and can be surrounded by a surface that further reduces horizontal gain. With a phased array antenna, the direction of a jamming signal can be found and a null in the antenna pattern steered to that direction. Mounting a GPS antenna on the wing of an aircraft provides blocking of the antenna pattern below the horizontal and diminishes the effect of a ground based jammer.

The FAA's original NextGen proposals in 2000 included the decommissioning of all NDB and VOR beacons, all ILS approaches to airports, and all secondary radar installations. Primary radars are shared with military air defense and need to be retained. By 2015 the possibility of jamming and spoofing of the GPS receivers on aircraft led to revised thinking, and a network of VOR stations and ILS approaches will be retained in the United States indefinitely as a backup to GPS navigation. Aircraft will always be within 100 mi of an ILS approach and 70 mi of a VOR beacon when flying in the lower 48 states.

The FAA has published GPS approaches to over 5000 airports in the United States. A GPS approach is similar to an ILS approach, essentially creating a path through the sky that leads to the runway threshold. However, unlike an ILS approach, the signal strength of a GPS receiver remains constant and is vulnerable to jamming or spoofing throughout the approach. Because the ILS system has powerful transmitters located on the airport, the signal gets steadily stronger as the aircraft approaches the runway threshold, the most dangerous part of the approach procedure, and is therefore much less vulnerable to jamming or spoofing than a GPS based approach. This may be one reason that autoland systems, which require a GPS receiver on the airport have not progressed as quickly as expected.

12.11 ADS-B and Air Traffic Control

Commercial and military ATC systems have used radar to manage air traffic since the 1960s. Prior to the use of radar for ATC, commercial aircraft flew on defined airways and pilots would report their position as they passed over specific landmarks.

Two passenger aircraft collided over the US Grand Canyon in 1955 causing 128 fatalities, proving that a better method of controlling air traffic was needed (Grand Canyon mid-air 1955). The US Congress established the Federal Aviation Authority (FAA) with a directive to develop an ATC system using radar that would guarantee aircraft would not collide in mid-air. ATC radar systems have two different types of radar. Primary radar, used mainly in air defense installations but also by ATC, uses RF energy reflected from an aircraft's surface to detect and locate aircraft. Large rotating antennas and high power transmitters are needed to detect aircraft at long ranges, as the reflected power from the target decreases as the fourth power of the range to the target. When a returned signal is detected the azimuth angle of the target and its range are measured, but no altitude information is obtained. Stealth aircraft are designed to reflect minimal RF energy from radars over a wide range of frequencies to make them difficult to detect at all but very short range.

Secondary radar employs a transponder on the aircraft that sends out a reply when interrogated by a secondary radar transmitter. An aircraft transponder is similar to an onboard processing satellite transponder in that it consists of a receiver and signal processing unit that transmits a coded message in response to a received signal. The basic transponder is known as mode *C* and sends out aircraft identity and altitude information. Secondary radar has two advantages over primary radar: it uses a communication link between the aircraft and ATC, so signals decrease in strength in proportion to range squared, rather than range to the fourth power as in primary radar, and the coded response includes aircraft altitude and identity. Combined with angle and range measurements from the radar, the three-dimensional location of the aircraft is determined.

The information about each aircraft in a given area is displayed on a TV type screen as a map showing the identity, altitude, speed, and direction of each aircraft. The task of the air traffic controller is to direct the aircraft so they never come close. A backup system on passenger aircraft called TCAS (traffic collision avoidance system) uses the transponder transmissions from nearby aircraft to determine whether they are getting too close and to warn the pilots to take evasive action if necessary. Almost all the world's ATC systems use secondary radar and have aircraft equipped with TCAS. The effectiveness of the system is illustrated by the very rare occurrence of mid-air collisions between commercial aircraft. As noted earlier in this chapter, there were no fatal injuries to passengers traveling by commercial aircraft in the United States between 2009 and 2017, with 2.5 million air travelers every day.

Primary and secondary radars are expensive to build and operate, and the coded messages used in secondary radar are 21 μs long, so occupy 6.3 km (4 mi) in space. Aircraft that are closer than 3 km (2 mi) in range and angle have overlapping transponder replies, so minimum separation is 2 mi for close-in target and 5 mi for long range (en route) aircraft. In 2000, the US FAA determined that GPS could locate aircraft much more effectively than radar for ATC purposes and provide more accurate navigation than existing ground based navaids. An intention to abandon all ground based radio and radar systems by 2020 was declared, replaced with GPS based location and navigation. A program

called NexGen (short for Next Generation ATC) was established, with transition to GPS based transmissions from mode S transponders.

A mode S transponder is similar to a mode *C* transponder but has an extended reply. The aircraft ID and altitude (read from the aircraft altimeter) are transmitted followed by information on the aircraft's location in latitude and longitude, its GPS reported altitude, and speed, derived from a dedicated GPS C/A code receiver on the aircraft. With WAAS capability, the average accuracy of reported aircraft position is ±2 m, providing ATC with knowledge of each aircraft's precise location and speed. The aircraft altitude is still read from the onboard altimeter, which works on barometric pressure and is required to be accurate to ±23 m (75 ft), because of the VDOP uncertainty in GPS altitude. The GPS reported altitude provides a check that the aircraft's altimeter is working correctly.

The mode S transponder carried by aircraft is part of a larger FAA program known as ADS-B (2015). There are two parts to ADS-B: ADS-B IN and ADS-B OUT. ADS-B OUT is the GPS receiver and mode S transponder that reports the aircraft location to ATC via hundreds of ground based stations, and also to all nearby aircraft, twice every second. One of the problems with secondary radar systems is that when a large number of aircraft are present, for example, in the airspace around New York, multiple overlapping replies can be transmitted from the aircraft mode *C* transponders causing confusion at the radar receiver (known as garbling). A mode S transponder automatically transmits twice every second with much lower risk of overlapping transmissions and much faster updates than the 6 or 10 second scan rate of ATC radars. ADS-B can also be used to track aircraft and vehicles on the ground at airports, where they are too close to the radar antenna for secondary radar to work. This is particularly important when mist and fog obscure parts of the airport from visual observation by ATC ground controllers in a control tower. Runway incursions by aircraft and trucks is a major concern at busy airports. Figure 12.14 shows an overview of the ADS-B system.

ADS-B IN is a service to aircraft that transmits information from the ground based stations. There are two types of information; Traffic information service, broadcast (TIS-B) and flight information service, broadcast (FIS-B). TIS-B data includes the location of all aircraft in the area, derived from ADS-B and radar observations. FIS-B data includes weather radar maps showing rain, warnings of hazardous weather conditions, information about nearby airports, and any notices to airmen (notams). These services are especially valuable to private pilots flying general aviation aircraft under visual flight rules where weather and other aircraft present possible hazards. The TIS-B and FIS-B information can be displayed on a tablet computer in the cockpit using a Bluetooth link to the ADS-B receiver. Foreflight® and Wing-X® are two popular subscription services that offers this capability (Foreflight 2018; Wing-X 2018). ADS-B OUT equipment has been mandated for all aircraft flying in controlled airspace from 1 January 2020. Only general aviation aircraft flying under visual flight rules well away from cities will be permitted to fly without ADS-B capability (AOPA NexGen 2018).

12.12 GPS Modernization

A number of upgrades have been made to the GPS satellites to improve the accuracy and availability of the civil C/A signals. The largest uncertainty in C/A code location is the delay of the C/A code as the signal passes through the ionosphere. The algorithms used with single frequency operation can lead to pseudo range errors as high as 14 m

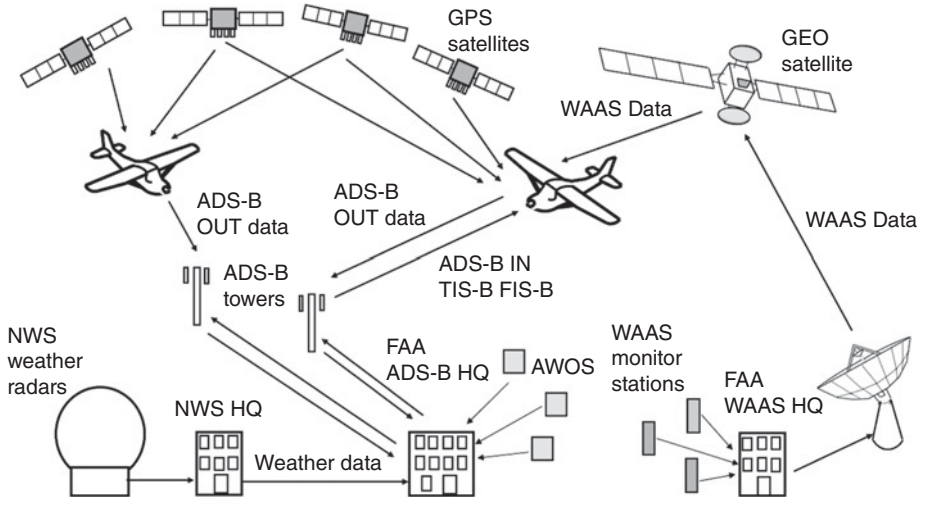


Figure 12.14 Overview of the ADS-B system for air traffic control. The system is based on GPS WAAS enabled receivers with location accuracies of 2 m on all aircraft flying in controlled airspace. The aircraft report their location and speed via an ADS-B OUT transmitter to a network of ADS-B towers across the United States that connect to a data center. Weather data is collected by the National Weather Service (NWS) from a network of weather radars, and local weather data is obtained from automatic weather observation stations (AWOSs) at thousands of airports. The weather radar maps and AWOS data, together with other information useful to pilots is transmitted by the ADS-B towers as a flight information signal (FIS-B) to aircraft that have an ADS-B IN receiver. Traffic information is transmitted as a TIS-B signal. The aircraft on the right of Figure 12.14 is equipped with ADS-B IN and OUT units and can fly under ATC control. The aircraft on the left does not have an ADS-B IN receiver and cannot receive the FIS-B and TIS-B data.

at tropical latitudes when sunspots are active. The ionospheric delay is approximately proportional to the inverse of the RF frequency squared. The P-code system has always used two frequencies, L1 and L2 that give different pseudo range values, allowing for accurate calculation of ionospheric delay. Two frequency operation with the C/A code is possible with either of two new transmissions from later Block II and Block III GPS satellites. L2C is a second C/A code transmission at the L2 frequency, in quadrature with the P-code L2 signal. L5 is a new frequency, 1176.45 MHz, with a modified C/A code format. Both of the new signals reduce ionospheric delay uncertainty to less than 2 m. The L2C signal was available from 19 GPS satellites in March 2018, with all satellites transmitting L2C by 2022 (GPS Modernization 2018; McDonald 2002).

The L5 transmission has two signals in phase quadrature; the I channel carries the navigation message and the Q channel has no modulation. The chip rate is increased to 10.23 Mcps with different but synchronized codes on both I and Q components and a 1 ms sequence. The unmodulated Q channel signal is used for acquisition and *C* code tracking, reducing the risk of loss of lock. Resistance to multipath and jamming is improved and accuracy is increased by the 10.23 Mcps chip rate to a level comparable to the P-code. L5 signals will be available from all Block III GPS satellites by the late 2020s. Using the C/A code L1, L2C, and L5 signals in a three channel receiver allows the accuracy of a civil code receiver to equal or exceed that of the encrypted P-code system (IEEE Spectrum 2018).

A new C code signal, L1C, using multiplexed binary offset carrier (MBOC) modulation will be transmitted by Block III GPS satellites, enabling international cooperation while protecting US national security interests. The design will improve mobile GPS reception in cities and other challenging environments (GPS Modernization 2018). For full details of the updated GPS system see (Interface specification IS-GPS-705B 2010).

12.13 Summary

GNNS systems have revolutionized navigation and become a consumer product with many applications. A system that was originally conceived for targeting nuclear weapons has become a worldwide commodity. Eventually GPS receivers will be present in all aircraft, ships, and automobiles so that navigation will become a simple task, and no one with a GPS receiver (and the ability to read a map) will ever be lost. GPS receivers are now embedded in all cellular telephones so that the user will know his or her location, and an emergency call will always contain location information. GPS may eventually replace all existing aircraft and ship radio navigation systems and become the sole means of navigation, permitting aircraft to select their own routes between airports and to be landed automatically in zero visibility conditions under the control of a local area DGPS system.

The principles of GPS position location are simple, but its execution with the required accuracy is complex. Position location with GPS is based on trilateration; measurement of the distances from an unknown point to three known points (GPS satellites) yields a unique solution for the position of the unknown point. The apparent distance between each satellite and the receiver is determined by measuring the time of flight of a signal transmitted by the satellite, using real time clocks on the satellite and at the receiver. Electromagnetic waves travel at approximately 3×10^8 m/s, or 300 m per μs. To achieve position location accuracy of a few meters, the time of flight of GPS signals must be determined to one hundredth of a microsecond or better accuracy.

GPS satellites have three atomic clocks with extreme accuracy, but most commercial GPS receivers use low cost quartz crystal controlled clocks that are not as accurate. The signals from a fourth GPS satellite are used to calculate the offset error in the clock of the GPS receiver, which allows the receiver to track time with an accuracy better than one hundredth of a microsecond. Thus four GPS satellites are needed to make an accurate three-dimensional position measurement. The distances between the GPS satellites and the receiver are known as pseudo ranges, because they may be in error by thousands of kilometers. The receiver clock error data provided by the fourth pseudo range measurement allows correction of pseudo ranges to true ranges. The satellites all transmit navigation messages that contain the ephemeris for each GPS satellite; these data are used by the GPS receiver to calculate satellite orbital positions.

GPS satellites use DSSS techniques, transmitting two PRN sequences known as C/A and P-codes at two frequencies in L-band. Each satellite is assigned a unique set of codes, which are used by the receiver to make time of flight measurements and also to identify the satellites. Commercial GPS position measurement receivers use the course acquisition (C/A) code transmitted at a chip rate of 1.023 Mbps by GPS satellites on the L1 carrier at a frequency of 1575.42 MHz. The C/A codes are never changed. Authorized users, primarily military, use the P-code, which has a chip rate 10.23 Mbps to provide

greater positioning accuracy. The P-code is encrypted and changes once a week. P-code signals are transmitted on the L1 carrier, in phase quadrature with the C/A code signal, and also on the L2 carrier at a frequency 1227.60 MHz. Receivers that use both L1 and new L5 transmissions are able to remove some timing errors introduced by atmospheric and ionospheric delays to achieve consistent accuracy of 2 m.

DGPS makes use of a fixed GPS receiver in a known location to calculate errors in the GPS position measurement, or errors in the pseudo ranges to all visible satellites, which are then sent over a radio link to other GPS receivers. The errors do not vary greatly over a wide area, allowing C/A code GPS receivers to achieve far greater accuracy than is possible with a single receiver. DGPS is the basis of the US FAA's WAAS that improves the accuracy of C/A code receivers to 2 m.

Exercises

12.1 Find the exact altitude of a GPS satellite that has an orbital period equal to precisely one half of a sidereal day. Use a value of mean earth radius $r_e = 6378.14$ km and a sidereal day length of 23 hours 56 minutes 4.1 seconds.

12.2 Find the maximum Doppler shift of the L1 signal frequency for a GPS satellite at an altitude of 20 200 km when the satellite has an elevation angle of 20°. Hint: Maximum Doppler shift occurs when the observer is in the plane of the satellite orbit. Find the velocity of the satellite and the component of velocity toward the observer.

12.3 An observer at the geographical north pole has a GPS receiver. At an instant in time, four GPS satellites all have the same range from the observer, and the GPS receiver records a measured delay time for the C/A signal of 0.170 975 28 s for each satellite. The four satellites' coordinates are calculated to be (0, −13 280.5, 23 002.5), (0, 13 280.5, 23 002.5), (−13 280.5, 0, 23 002.5), (13 280.5, 0, 23 002.5), where all distances are in km. Assuming an earth radius of 6378.0 km at the north pole, so that the observer's coordinates are (0, 0, 6378), determine the clock offset error in the GPS receiver. (Use Eqs. (12.1) and (12.3), and take the velocity of light in free space to be 2.99792458×10^8 m/s.)

12.4 Accurate position location using GPS requires precise knowledge of the speed of light. In most applications, we use a velocity of light of 3.0×10^8 m/s. Solve problem #3 above and then recalculate the clock offset using $c = 3 \times 10$ m/s instead of the more precise value given in problem #3. What is the error in the clock offset? What is the difference in the ranges to the satellites when the approximate value for $c = 3 \times 10^8$ m/s is used? Discuss the corresponding position error due to the approximation. Why is it essential to use the exact value of the velocity of EM waves?

12.5 A C/A code GPS receiver is located at the geographic south pole, coordinates $(0,0, z_p)$. Four GPS satellites are used to determine the radius of the earth at the south pole. At the instant of time that the measurement is made, the satellites have coordinates

#1: (0, −13 280.500, −23 002.500)	#2: (0, 13 280.500, −23 002.500)	#3: (13 280.500, 0, −23 002.500)	#4: (0, 0, −26 561.000)

The corresponding measured delay times for the C/A code sequences from the satellites are

#1 : 0.12102731 s	#2 : 0.12102731 s	#3 : 0.12102731 s	#4 : 0.11738995 s

Find the clock offset in the GPS receiver, and determine the radius of the earth at the south pole. Use a value for the velocity of light in free space $c = 2.99792458 \times 10^8$ m/s, and work your solution to a precision of 1 m. You will need to solve two simultaneous non-linear equations from the set in Eq. (12.3) in which the unknowns are the clock offset and the value of z_p. Start with an estimated value $z_p = 6378$ km, and then solve the two simultaneous equations. This will give two unequal values for the clock offset. Use iteration of the value of z_p to find the correct values for clock offset and earth radius at the south pole. An equation solver can also be used.

References

ADS-B (2015). https://www.faa.gov/nextgen/programs/adsb (accessed 6 June 2018).

AOPA NexGen (2018). https://www.aopa.org/advocacy/advocacy-briefs/air-traffic-services-brief-automatic-dependent-surveillance-broadcast-ads-b (accessed 6 June 2018).

Aviation Week (2018). http://aviationweek.com/commercial-aviation/faa-targets-2018-gps-based-autoland-capability (accessed 6 June 2018).

Beidou (2018). https://www.bipm.org/en/CGPM/db/15/2 (accessed 25 April 2018).

BIPM (1975). https://www.bipm.org/en/CGPM/db/15/2 (accessed 25 April 2018).

Black sea spoofing incident (2017). http://gpsworld.com/spoofing-in-the-black-sea-what-really-happened (accessed 11 May 2018).

Clarke, B. (1996). *Aviation Application of GPS*. NY: McGraw-Hill.

Couch, L.W. (2007). *Digital and Analog Communication Systems*, 7e. Upper Saddle River, NJ: Pearson Education Inc.

Egan, W.F. (1998). *Phase-Lock Basics*. NY: Wiley.

FAA NexGen (2018). https://www.faa.gov/nextgen (accessed 25 April 2018).

Foreflight (2018). https://www.foreflight.com (accessed 13 May 2018).

Galileo (2017). https://www.esa.int/Our_Activities/Navigation/Galileo/What_is_Galileo (accessed 15 May 2018).

Gardner, F.M. (2005). *Phaselock Techniques*, 3e. Holboken, NJ: Wiley.

GLONASS (2017). https://en.wikipedia.org/wiki/GLONASS (accessed 15 May 2018).

GPS Control Segment (2017). https://www.gps.gov/systems/gps/control (accessed 15 May 2018).

GPS cost (2011). http://www.gpsalliance.org/docs/Government_GPS_Investments.pdf (accessed 24 April 2018).

GPS.gov (1995). Global positioning system standard positioning service signal specification. https://www.gps.gov/technical/ps/1995-SPS-signal-specification.pdf (Accessed 23 May 2018).
GPS.gov (2018). https://www.gps.gov/systems/gps/control (accessed 24 April 2018).
GPS Jamming (2012). https://www.gps.gov/governance/advisory/meetings/2014-06/scott.pdf (accessed 11 May 2018).
GPS Jamming (2013). http://www.nj.com/news/index.ssf/2013/08/man_fined_32000_for_blocking_newark_airport_tracking_system.html (accessed 11 May 2018).
GPS Modernization (2018). https://www.gps.gov/systems/gps/modernization/civilsignals (accessed 11 May 2018).
GPS Satellites (2018). www.losangeles.af.mil/About-Us/Fact-Sheets/Article/343724/gps-iif (accessed 24 April 2018).
GPS signal specification (1995). https://www.gps.gov/technical/ps/1995-SPS-signal-specification.pdf (accessed 3 May 2018).
GPS Spoofing (2016). Protecting GPS from Spoofers Is Critical to the Future of Navigation, *IEEE Spectrum*. http://spectrum.ieee.org/telecom/security/protecting-gps-from-spoofers-is-critical-to-the-future-of-navigation (accessed 12 May 2018).
GPS Standard positioning system performance standard (2001). https://www.gps.gov/technical/ps/2001-SPS-performance-standard.pdf (accessed 10 May 2018).
Grand Canyon Midair (1955). https://en.wikipedia.org/wiki/1956_Grand_Canyon_mid-air_collision (accessed 12 May 2018).
Grewal, M.S., Andrews, P.A., and Bartone, C.G. (2013). *Global Navigation Satellite Systems, Inertial Navigation, and Integration*. Hoboken, NJ: Wiley.
Haykin, S.S. (2001). *Digital Communications*, 4e. Hoboken, NJ: Wiley.
IEEE Spectrum (2018). Superaccurate GPS Chips Coming to Smartphones in 2018. *IEEE Spectrum: Technology, Engineering, and Science News. September 2017*.
IIHS (2018). http://www.iihs.org/iihs/topics/t/general-statistics/fatalityfacts/overview-of-fatality-facts (accessed 25 April 2018).
Interface specification IS-GPS-705B (2011). https://www.navcen.uscg.gov/pdf/gps/IS-GPS-705B.pdf (accessed 14 May 2018).
Ionospheric scintillation (2012). http://gpsworld.com/gnss-systemsignal-processinginnovation-ionospheric-scintillations-12809 (accessed May 8, 2018).
McDonald, K.D. (2002). The modernization of GPS: plans, new capabilities and the future relationship to Galileo. *Journal of Global Positioning Systems* 1 (1): 1–17.
Parkinson, B.W. and Spilker, J.J. (1996). *The Global Positioning System – Theory and Applications,* Vols I and II, AIAA Progress in Astronautics and Aeronautics, vol. 164. Washington, DC: AIAA Inc.
SAI (2015). 2015 Satellite Industry Association Report https://brycetech.com/downloads/SIA_SSIR_2015.pdf (accessed 25 April 2018).
Schofield, B.B., Vice Admiral (2008). Operation Neptune, Barnsley, Pen and Sword Military. Available in Kindle edition at https://www.amazon.com/Operation-Neptune-B-B-Schofield/dp/1844156621 (accessed 6 June 2018).
Ship navigation (2017). https://cacm.acm.org/magazines/2017/9/220436-why-gps-spoofing-is-a-threat-to-companies-countries/fulltext (accessed 12 May 2018).
Skolnik, M.I. (1981). *Introduction to Radar Systems*, 2e, 402. New York, NY: McGraw Hill.
Sparkfun (2018). https://www.sparkfun.com/pages/GPS_Guide (accessed 10 May 2018).

Strang, G. and Borre, K. (1997). *Linear Algebra, Geodesy, and GPS*. Wellesly, MA: Cambridge Press.

Tsui, J.B.-Y. (2000). *Fundamentals of Global Positioning System Receivers: A Software Approach*, 3e, 80–81. Hoboken, NJ: Wiley.

WAAS accuracy (2017). www.nstb.tc.faa.gov/reports/PAN96_0117.pdf (accessed May 9 2018).

Wing-X (2018). http://hiltonsoftware.co (accessed 13 May 2018).

Glossary

$\boldsymbol{\alpha}$ Roll off factor of a SRRC filter
$\boldsymbol{\delta(t)}$ Unit impulse
$\boldsymbol{\gamma}$ Central angle in orbital calculations
$\boldsymbol{\gamma_R}$ Specific attenuation
$\boldsymbol{\eta_A}$ Aperture efficiency
$\boldsymbol{\eta_c}$ Coupling coefficient
$\boldsymbol{\theta_1, \theta_2}$ Beamwidths of antenna in orthogonal directions
$\boldsymbol{\theta_{3dB}}$ 3 dB beamwidth of an antenna
$\boldsymbol{\lambda}$ Wavelength
$\boldsymbol{\mu}$ Kepler's constant (3.986004418×10^5 km^3/s^2)
μW Microwatts
$\boldsymbol{\sigma}$ Gaussian distributed noise voltage
$\boldsymbol{\tau}$ Receiver clock error (offset or bias)
$\boldsymbol{\varphi}$ Phase angle
$\boldsymbol{\omega}$ Frequency in radians per second
$\boldsymbol{\Delta N}$ Increase in noise power
$\boldsymbol{\Delta N_{rain}}$ Increase in noise power with rain in the slant path
3IM point The third order intermodulation point
8-PSK Eight phase PSK
8-VSB Eight level vestigial amplitude modulation
16-APSK 16 level amplitude-phase shift keying
a Radius of geostationary orbit (42,164.17 km)
$\boldsymbol{A}$ Area
$\boldsymbol{A}$ Attenuation
(a, b) *Semimajor, semiminor axes of an ellipse*
ACK Acknowledge
ACM Adaptive coding and modulation
ADC Analog to digital converter
ADDR Address
ADF Automatic direction finder
ADPCM Adaptive differential PCM
ADS-B Automatic dependent surveillance - broadcast
ADS-B in Signals received by aircraft from ADS-B ground transmitters
ADS-B out Signals transmitted by aircraft to ADS-B ground receivers
$\boldsymbol{A_e}$ Effective aperture of an antenna

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

$A(f_n)$ Attenuation at frequency n
AGC Automatic gain control
AH Ampere hours
AKM Apogee kick motor
AM Amplitude modulation
AOCS Attitude and Orbit Control System
AOPA Aircraft Owners and Pilots Association
AOR Atlantic Ocean Region
ANSI American National Standards Institute
A_r Receiving antenna aperture area
ARQ Automatic Repeat Request
ASCII American standard code for information interchange
ASIC Application specific integrated circuit
AT&T American telephone and telegraph company
ATC Air traffic control
ATM Asynchronous transfer mode
ATSC Advanced Television Standards Committee
AWGN Additive white Gaussian noise
Az Azimuth
Az-El Antenna mounting with a horizontal elevation axis and a vertical azimuth axis
B Bandwidth in hertz
BCH Bose-Chaudhuri Hocquenghem code
BER Bit error rate
BIPM Bureau International des Poids et Mesures (International Bureau on Weights and Measures)
B_n Noise bandwidth in hertz
BOL Beginning of life
BPF Band pass filter
BPSK Binary phase shift keying
BSS Broadcast satellite service
BTC Block turbo code
C Carrier power
C-band (4–8 GHz)
CDF Cumulative distribution function
C/(N+I) Ratio of carrier power to noise plus interference power
C/A Code Course acquisition code (GPS)
C/I Carrier to interference ratio
CATC Cable TV
C_b Total number of bits in a TDMA frame
C_{ca} Carrier power with clear sky conditions
CCDev Commercial Crew Development
CCM Constant coding and modulation
CD Compact disc
CDI Course deviation indicator
CDMA Code division multiple access
$C_i(t)$ Gold code sequence for C/A code (GPS)
CKS Checksum
CKSM Checksum

CLPC Code book excited LPC
$(CNR)_{dn}$ Downlink carrier to noise ratio
$(CNR)_{eff}$ Effective CNR (includes implementation margin)
$(CNR)_{IM}$ Intermodulation carrier to noise ratio
$(CNR)_{o}$ Overall carrier to noise ratio
$(CNR)_{up}$ Uplink carrier to noise ratio
CNTL Control
COFDM Coded orthogonal frequency division multiplexing
CONUS Contiguous United States (lower 48 states)
CP Circular polarization
CPA Co-polar attenuation
CPFSK Continuous phase FSK
C_{rain} Carrier power with rain in the slant path
CRBS Carrier recovery and bit synchronization
CRC Cyclic redundancy check
CRC-8 8-bit cyclic redundancy error check
CRS2 Commercial resupply service 2
CSC Common signaling channel
CTC Convolutional turbo code
d Distance to a satellite from an earth station
D Diameter of an antenna aperture
DA Demand assignment
DAB Digital audio broadcasting
DAC Digital to analog converter
DAMA Demand assignment multiple access
dB Decibel
dBm Decibels greater than one milliwatt
DBSS Direct broadcasting satellite service
DBS-TV Direct broadcast satellite TV
dBW Decibels greater than one watt
DCT Discrete cosine transform
DDM Digital demodulator
DEMUX Demultiplexer
DGPS Differential GPS
$D_i(t)$ Navigation message (GPS)
D_m Median drop diameter
DME Distance measuring equipment
DNTX Do not transmit code
DoD United States Department of Defense
DOP Dilution of precision
DSL Digital subscriber line
DS-n Hierarchy of digital data rates
DSNG Digital satellite news gathering
DSP Digital signal processing
DSSS Direct sequence spread spectrum
DVB-C Digital video broadcasting standard – Cable TV
DVB-RCS Digital video broadcast system with a return link
DVB-S, DVB-S2 Digital video broadcasting standards (ETSI) satellite

DVB-T Digital video broadcasting standard – terrestrial broadcast
DVD Digital video disc
e Eccentricity of a satellite orbit
E_{av} Average energy
E_b Energy per bit
E_b/N_o Ratio of energy per bit to noise power spectral density
ECEF Earth centered earth fixed (coordinate system)
EELV Evolved expendable launch vehicle
EGNOS European geostationary navigation overlay service
EHF Extra high frequency
EIRP Effective isotropically radiated power
El Elevation
ELT Emergency locator transmitter
ELV Expendable launch vehicle
ENST Ecole Nationale Superieure des Telecommunications de Bretagne
EOC Edge of coverage
EODL End of design life
EOL End of life
EOML End of maneuvering life
erfc(x) Complementary error function, probability a noise voltage exceeds a value of x volts
E_s Energy per symbol
ESA European Space Agency
E_s/N_o Ratio of energy per symbol to noise power spectral density
ETSI European Telecommunications Standards Institute
EU European Union
f Frequency in hertz
FA Fixed access
FAA US Federal Aviation Administration
FCC US Federal Communications Commission
f_d Doppler frequency
FDMA Frequency division multiple access
FEC Forward error correction
FEP Front-end processor
FET Field effect transistor
FFSK Fast frequency shift keying
FH-SS Frequency hopping spread spectrum
FIR Finite impulse response (digital filter)
FIS-B Flight information signal
FM Frequency modulation
FPGA Field programmable gate array
FSS Fixed satellite system
FTTH Fiber to the home
G Antenna gain, amplifier gain
G Universal gravitational constant ($6.672 \times 10^{-11} \mathrm{Nm}^2/\mathrm{kg}^2$)
$G(\theta)$ Gain as a function of angle
G1, G2 Shift registers used to generate C/A code (GPS)
GaAsFET Gallium arsenide field effect transistor
GAGAN Indian GPS aided GEO augmented navigation

$G_D(P)$ Diversity gain
GDOP Geometric dilution of precision
GEO Geostationary earth orbit
GES Gateway earth station
G_{IF} IF amplifier gain
G_l Gain less than unity (a loss)
G_m Mixer gain
GMSK Gaussian minimum shift keying
GMT Greenwich mean time
GNSS Global navigation satellite system
GPS Global positioning system
G_r Receiving antenna gain
G_{RF} LNA gain
GSM Global system mobile (worldwide cell phone standard)
GSFC Goddard Space Flight Center
GSO Geostationary earth orbit
G_t transmitting antenna gain
G/T Antenna gain to system noise temperature ratio
GTO Geostationary transfer orbit
h Altitude of a satellite above earth
h_r Rain height
h_s Earth station altitude
H Channel capacity
H Horizontal polarization
HBE Hub baseband equipment
HD High definition (digital television)
HDL Data transmission protocol
HDOP Dilution of horizontal precision
HDTV High definition TV
HEO Highly elliptic orbit
HEX Hexadecimal
HF High frequency
hPa Hectopascal
HPA High power amplifier
HPF High pass filter
HUMINT Human intelligence
I In-phase
I_D Diversity improvement
IBS Intelsat business service
ICBM Intercontinental ballistic missile
IEEE Institute of Electrical and Electronic Engineers
IET Institute of Engineering Technology
IF Intermediate frequency
IGO Intergovernmental organization
IIR Infinite impulse response (digital filter)
ILS Instrument landing system
IM Intermodulation
ImpM Implementation margin

Inmarsat International Maritime Satellite Organization
Intelsat International Telecommunications Satellite Organization
ICAO International Civil Aviation Organization
IOR Indian Ocean Region
IOT Internet of things
ISI Intersymbol interference
ISL Intersatellite links
ISO International Systems Organization
ISARA Integrated solar array reflectarray
ISS International Space Station
ITU International Telecommunications Union
JPEG Joint pictures expert group
JPL Jet Propulsion Lab
JTC Joint Technical Committee (ETSI)
k Boltzmann's constant (1.39×10^{-23} J/K or -228.6 dBW/K/Hz)
k Number of parity bits in a codeword
K-band (18–27 GHz)
Ka-band (27–40 GHz)
Ku-band (12–18 GHz)
kW Kilowatts
kWH Kilowatt hours
L-band (1–2 GHz)
L1, L2, L5 CDMA signals (GPS)
L_a Attenuation in the atmosphere
LAN Local area networks
LDPC Low density parity check code
LED Light emitting diode
L_{eff} Effective pathlength
LEO Low earth orbit
LET Linear energy transfer
LHCP Left hand circularly polarization
LMDS Local multipoint distribution service
LNA Low noise amplifier
LNB Low noise block converter
LNC Low noise converter
LO Local oscillator
L_p Path loss
LP Linear polarization
LPA Low power amplifier
LPC Linear predictive encoding
LPF Low pass filter
L_{ra} Losses in a receiving antenna
L_{ta} Losses in a transmitting antenna
LTE Long term evolution (cell phones)
M Number of levels in an ADC
M Number of TDMA frames transmitted per second
M Ratio of chip rate to bit rate in DSSS-CDMA
M2M Machine to machine

M_e Mass of the earth (5.98×10^{24} kg)
m-APSK Multi-level m-phase PSK
MAD Mutually Assured Destruction
MBOC Multiplexed binary offset carrier
MCDDD Multi-carrier demodulation, demultiplexing, and decoding
MCS Master control station (GPS)
MF-TDMA Multi-frequency time division multiple access
MIRV Multiple independently targeted vehicles (nuclear warheads)
mm wave Frequencies above 110 GHz
ModCon Modulation and coding combination
MOS Mean opinion score
MPEG-2, MPEG-4 Video compression techniques
m-PSK PSK modulation with m phase states
MSAS Japanese multi-functional satellite augmentation system
MSK Minimum shift keying
MTBF Mean time between failures
MUX Multiplexer
MW Megawatts
mW Milliwatts
N Noise power, number of bits in an ADC or DAC.
n Number of bits in a code word
n(t) Noise voltage waveform
NAK Not acknowledge or negative acknowledge
NASA US National Aeronautics and Space Administration
NCO Numerically controlled oscillator
NDB Non-directional beacon
***N*(*D*)** Drop size distribution
NexGen Next generation air traffic control system
NF Noise figure
NGSO Non-geostationary orbit
NIST US National Institute of Science and Technology
N_o Single sided noise power spectral density
NOC Network operations center
NPSD Noise power spectral density
NRZ Non-return to zero
N-S North-South
NTSC National Television Standards Committee
NWS National Weather Service
N_{xp} Noise power at transponder input
OBP Onboard processing
OC-n Hierarchy of optical fiber data rates
OFDM Orthogonal frequency division multiplexing
OMT Orthogonal mode transducer
OQPSK Offset QPSK
OSI Open Systems Interconnect
P Power
P Pascal (unit of pressure)
p Probability of a single-bit error

$P(\theta)$ Power as a function of angle
$P(v)$ Probability distribution function of v
PA Preassigned access
PAM Pulse amplitude modulation
PCM Pulse code modulation
P-code PRN code of precise positioning system (GPS)
pdf Probability distribution function
P_e Probability of error
PLL Phase locked loop
P_n Noise power
POR Pacific Ocean Region
PPS Precise positioning service (GPS)
PR Pseudorange
P_r Received power
PRBS Pseudo random sequence used for scrambling DBS-TV signals
PRN Pseudo random sequence
PSK Phase shift keying
PSTN Public switched telephone network
P_t Transmitted power
pW Picowatts
Q Quadrature
$Q(z)$ Q function, probability that a noise voltage exceeds a value of z volts
QAM Quadrature amplitude modulation
QEF Quasi-error free
QPSK Quadrature phase shift keying (four phase)
r_e Mean earth radius (6378.137 km)
R Distance or range
R Rainfall rate
$R(\tau_i - \tau)$ Autocorrelation function
rad Radian
rad(Si) is a unit of radiation energy
RAIM Receiver autonomous integrity monitoring
RB Radiocommunication Bureau (ITU)
R_d Message data rate
RF Radio frequency
RHCP Right hand circularly polarization
RHI Range height indicator
RMS Root mean square
ROM Read only memory
RRB Radio Regulations Board (ITU) Radiocommunication Bureau (RB) of the ITU
R-S Reed-Solomon code
R_{tc} Data rate for the packets
S-band (2–4 GHz)
SA Selective availability
SAR Specific absorption rate
SBAC Satellite broadcasting and communications association
SBAS Satellite based augmentation system
SCPC Single channel per carrier

SCPC-FDMA-DAMA Single channel per carrier frequency division multiple access demand access multiple access
SD Standard definition (digital television)
SDARS Satellite digital audio radio service
SDI Strategic Defense Initiative (Star Wars)
SDTV Standard definition TV
SES Société Européenne de Satellites (European Society for Satellites)
SHF Super high frequency
SIA Satellite Industry Association
SIGINT Signals intelligence
SISO Soft input soft output
SMDC U.S. Army Space and Missile Defense Command
SMS Short messaging system
$\mathbf{(SNR)_{in}}$ Signal to noise ratio at the input to a device
$\mathbf{(SNR)_{out}}$ Signal to noise ratio at the output of a device
$\mathbf{(SNR)_{PCM}}$ Signal to noise ratio in a PCM link
$\mathbf{(SNR)_{q}}$ Signal to noise ratio for quantization noise
$\mathbf{(SNR)_{t}}$ Signal to noise ratio for thermal noise
SOC Satellite operations center
SOF Start of frame sequence
SOI Silicon-on-insulator
SONET Synchronous optical network
SOS Silicon-on-sapphire
SPS Standard positioning service (GPS)
SRRC Square root raised cosine
SSPA Solid state high power amplifier
SSTL Surrey Satellite Technology Ltd
SV Space vehicle (GPS)
$\mathbf{T_{antenna}}$ Antenna noise temperature
$\boldsymbol{T_b}$ Bit period
$\boldsymbol{T_c}$ C/A code chip period (GPS)
TCAS Traffic collision avoidance system
TCP/IP Transmission control protocol/Internet protocol
TEC Total electron content
$\boldsymbol{T_d}$ Burst transmission duration of traffic bits in TDMA
TDD Time division duplexing
TDM Time division multiplexing
TDMA Time division multiple access
TDRS Tracking and Data Relay Satellite
$\boldsymbol{t_{frame}}$ Frame time
$\boldsymbol{t_g}$ Guard time
$\boldsymbol{T_{IF}}$ IF amplifier noise temperature
$\boldsymbol{T_{in}}$ Input noise temperature
TIROS Television infrared observation satellite
TIS-B Traffic information service
$\boldsymbol{T_m}$ Mixer noise temperature
$\boldsymbol{T_n}$ Noise temperature in kelvins
$\boldsymbol{T_{no}}$ Output noise temperature

T_p Physical temperature in Kelvin degrees
t_{pre} Preamble time
T_{RF} LNA or RF amplifier noise temperature
T_s Symbol period
T_s System noise temperature
T_{sca} System noise temperature in clear sky conditions
T_{sky} Sky noise temperature
TTC&M Telemetry, tracking, command, and monitoring
TV Television
TWTA Travelling wave tube amplifier
T_{xp} Noise temperature at transponder input
ULA United Launch Alliance
UPC Uplink power control
URE User range error
USAT Unwanted satellite
UTC Universal coordinated time
UW Unique word
V Vertical polarization
$v(t)$ Signal voltage waveform
V-band (40–75 GHz)
VCO Voltage controlled oscillator
VDOP Dilution of vertical precision
VHF Very high frequency
VLEO Very low earth orbit
VOIP Voice over Internet protocol
VOR VHF omni-range navigation system
VSAT Very small aperture terminal
VBP VSAT baseband processor
VSAT/WLL Very small aperture terminal/wireless local loop
Vu Digital signal voltage
W Watts
W band (75–110 GHz)
WAAS Wide area augmentation system
WRC World Radio Conference
SAT Wanted satellite
X-band (8–12 GHz)
X_1 P-code PRN sequence of 15,345,000 bits
X_2 P-code PRN sequence of 15,345,037 bits
X.25 Data transmission protocol
XPD Cross-polarization discrimination
XPI Cross-polarization isolation
x – y mount Antenna mounting with two orthogonal horizontal axes
z Zulu time (GMT)

Appendix A

Decibels in Communications Engineering

Most readers of this book will be familiar with the practice of expressing power ratios in decibels, abbreviated dB. The dB ratio A of two power levels P_1 and P_2 is given by

$$A = 10 \log_{10}\left(\frac{P_1}{P_2}\right) \tag{A.1}$$

provided that P_1 and P_2 are expressed in the same units. Although the decibel is formally defined only for a power ratio, P_1 and P_2 and A can also be expressed in terms of many combinations of voltage, current, resistance, electric field strength, magnetic field strength, and so on. It is common practice in communications engineering to use decibels and the mathematical properties of the logarithm to transform multiplicative equations to additive equations, to manipulate the additive equations into particularly convenient forms, and to define new logarithmic units with dB in their names for some of the quantities that appear. When first presented, this practice is confusing to many people, and we hope to clarify it here.

Consider the simple voltage divider circuit with resistors R_S and R_L shown in Figure A.1. The rms voltages across the source and the load resistance are V_S and V_L, respectively; the rms power supplied by the source is P_S W and the rms power delivered to the load is P_L W. From elementary circuit theory, these quantities are related by

$$P_S = \frac{{V_S}^2}{R_S + R_L} \tag{A.2}$$

$$P_L = \frac{{V_L}^2}{R_L} \tag{A.3}$$

$$V_L = \frac{V_S R_L}{R_S + R_L} \tag{A.4}$$

$$P_L = \frac{P_S R_L}{R_S + R_L} \tag{A.5}$$

letting

$$R_S + R_L = R_T \tag{A.6}$$

then

$$P_L = \frac{P_S R_L}{R_T} \tag{A.7}$$

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

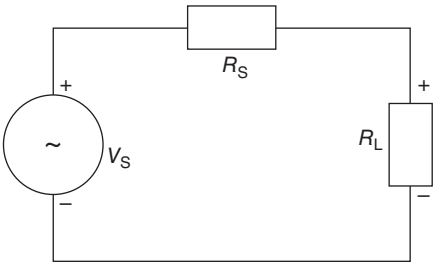


Figure A.1 Illustration of a voltage divider circuit. V_S is a voltage source, R_S is the source resistance, and R_L is a load resistance.

Eq. (A.7) is a multiplicative equation, and P_S and P_L must have the same units.

Expressed another way, whatever units we substitute in for P_S, P_L will have the same units. The reader should keep this in mind for what comes next. Solving for the ratio and expressing the result in decibels, we have

$$10\ \log_{10}\left(\frac{P_L}{P_S}\right) = 10\ \log_{10}\left(\frac{R_L}{R_T}\right) \tag{A.8}$$

Note that R_T is always equal to or greater than R_L and P_L is always less than or equal to P_S. This makes both sides of Eq. A.8 and several of the following equations have negative decibel values on both sides of the equation. A negative decibel value simply indicates that the value is less than unity. Note also that the log of zero is negative infinity.

Invoking the properties of logarithms we may rewrite Eq. A.8 as

$$10\ \log_{10} P_L - 10\ \log_{10} P_S = 10\ \log_{10}\left(\frac{R_L}{R_T}\right) \tag{A.9}$$

Without affecting the correctness of this equation or of its predecessor we may divide P_S and P_L by 1 W. Expressed in the form of Eq. (A.9), the result is

$$10\ \log_{10}\left(\frac{P_L}{1\ \mathrm{W}}\right) - 10\ \log_{10}\left(\frac{P_S}{1\ \mathrm{W}}\right) = 10\ \log_{10}\left(\frac{R_L}{R_T}\right) \tag{A.10}$$

The first term above is the decibel ratio of P_L to 1 W. This is defined as P_L expressed in units of decibels greater than 1 W, or P_L in dBW. If we represent this quantity as P_L (dBW) then

$$P_L\ (\mathrm{dBW}) = 10\ \log_{10}\left(\frac{P_L}{1\ W}\right) \tag{A.11}$$

Likewise the source power in dBW is P_S (dBW) where

$$P_S\ (\mathrm{dBW}) = 10\ \log_{10}\left(\frac{P_S}{1\ W}\right) \tag{A.12}$$

Substituting P_L and P_S into Eq. A.10 yields

$$P_L\ (\mathrm{dBW}) - P_S\ (\mathrm{dBW}) = 10\ \log_{10}\left(\frac{R_L}{R_T}\right) \tag{A.13}$$

If we had expressed the powers in Eq. A.9 in milliwatts and then divided both of them by 1 mW, the only effect would have been to express P_S and P_L in decibels greater than

1 mW or P_L and P_S in dBm.

$$P_L \text{ (dBm)} = 10 \log_{10}\left(\frac{P_L}{1 \text{ mW}}\right) \tag{A.14}$$

$$P_S \text{ (dBm)} = 10 \log_{10}\left(\frac{P_S}{1 \text{ mW}}\right) \tag{A.15}$$

and Eq. A.9 becomes

$$P_L \text{ (dBm)} - P_S \text{ (dBm)} = 10 \log_{10}\left(\frac{R_L}{R_T}\right) \tag{A.16}$$

Eqs. A.13 and A.16 differ only in the logarithmic power units that appear on their left hand sides. These equations are true so long as P_S and P_L are both expressed in the same logarithmic units. This happens because units that cancel by division in multiplicative equations like Eq. A.7 cancel by addition or subtraction in additive decibel equations like Eqs. A.13 and A.16. We can use this to write a general form for both these equations as

$$P_L - P_S = 10 \log_{10}\left(\frac{R_L}{R_T}\right) \text{dB} \tag{A.17}$$

The dB unit after the equation means that the quantities involved must be expressed in a consistent set of logarithmic units. Eq. A.17 is typical of many of the equations used in this text that contain a mixture of decibel and non-decibel units. We can change the equation to one involving only decibel quantities if we divide both the resistances by 1 Ω (Ω) and transform the ratio on the right hand side to a difference

$$P_L - P_S = 10 \log_{10}\left(\frac{R_L}{1\ \Omega}\right) - 10 \log_{10}\left(\frac{R_S}{1\ \Omega}\right) \tag{A.18}$$

Now let us define a new unit for our own use called the dBΩ for decibels above one ohm. This is a dubious but expedient use of the term decibel! Thus

$$R_L \text{ (in dB}\Omega\text{)} = 10 \log_{10}\left(\frac{R_L}{1\ \Omega}\right) \tag{A.19}$$

We can also express R_T in dBΩ by taking its log. Thus in these new units Eq. A.16 becomes

$$P_L \text{ (dBm)} - P_S \text{ (dBm)} = R_L \text{ (dB}\Omega\text{)} - R_T \text{ (dB}\Omega\text{)} \tag{A.20}$$

Likewise we could have expressed the resistance in kilohms (kΩ) and then divided all of the resistance terms by 1 kΩ, inventing a new unit that we will call the dBkΩ. The result could have been either

$$P_L \text{ (dBm)} - P_S \text{ (dBm)} = R_L \text{ (dBk}\Omega\text{)} - R_T \text{ (dBk}\Omega\text{)} \tag{A.21}$$

or

$$P_L \text{ (dBW)} - P_S \text{ (dBW)} = R_L \text{ (dBk}\Omega\text{)} - R_T \text{ (dBk}\Omega\text{)} \tag{A.22}$$

The units in these two equations cancel by addition and subtraction rather than by multiplication and division. Hence so long as both powers are in the one common decibel unit and so long as both resistances are in another common decibel unit, we may write a general form of these equations

$$P_L - P_S = R_L - R_T \tag{A.23}$$

A common alternative is to rearrange Eqs. (A.21) and (A.22) so that input quantities and output quantities are on opposite sides of the equation. For example, if the usual problem is to find the load power, then Eq. A.2 l would most usefully be expressed as

$$P_L \text{ (dBW)} = P_S \text{ (dBW)} + R_L \text{ (dB}\Omega\text{)} - R_T \text{ (dB}\Omega\text{)} \tag{A.24}$$

Readers seeing expressions like Eq. A.24 for the first time object to its apparent addition of dBW and dBΩ, since they have learned by sad experience that quantities with different units should not be added. But logarithmic units are different: quantities with different logarithmic or decibel units are added; the test of correctness is whether or not the units cancel by addition or subtraction. Adding and subtracting in dB units is the equivalent of multiplying and dividing in conventional arithmetic, something we frequently do. For example, to calculate gas mileage in a car, we divide miles driven by gallons used, giving miles per gallon. Thus in Eq. A.24 the dBΩ units cancel in the subtraction of R_T in dBΩ from R_L in dBΩ and the dBWs cancel because they appear on both sides of the equal sign.

The approach we have followed in this text is to use decibel units where the practice is common (principally for power) and to call the reader's attention to other common dB units at the point of first introduction. Besides the dBW and the dBm, a common power unit is the dBp, which is power expressed in dB above 1 picowatt. We must emphasize again that decibel units of resistance are irregular and that we have introduced them here only as a teaching tool.

Other dB units used frequently in this text are dBK, for decibels greater than 1 K, for noise temperature, and dBHz, for decibels greater than 1 Hz for bandwidth. Combined with Boltzmann's constant, k, the expression $k\,T\,B$ is noise power, usually calculated as

$$N = k + 10\,\log_{10}\,T + 10\,\log_{10}\,B \text{ dBW} \tag{A.25}$$

where k has a value −228.6 dBW/K/Hz.

One error that students frequently make as a result of first meeting decibels in the electronics lab is to write amplifier voltage gain as

$$G = 10\,\log_{10}\,\frac{{V_{out}}^2}{{V_{in}}^2} = 20\,\log_{10}\,\frac{V_{out}}{V_{in}} \text{ dB} \tag{A.26}$$

and then use 20 $\log_{10}$ (…) for power calculations. Eq. A.26 is correct only if the input and output impedances are identical, because a decibel is, by definition, a power ratio. However, voltage gain will undoubtedly continue to be defined by Eq. A.26, strictly a misuse of dB. The critical point to remember is that throughout the world of communications, all decibels calculations require 10 $\log_{10}$ (…), never 20 $\log_{10}$ (…) unless the quantities are squared. Also be careful not to multiply two decibel value together. That is equivalent to raising the first value to the power of the second value.

Another possible dB unit is the dB$. Everyone is looking for a 3 dB increase in their salary!

Appendix B

Antennas

B.1 Introduction

This appendix attempts to introduce the reader who is unfamiliar with radio antennas to a topic that is the subject of many textbooks and semester-long graduate courses. It is therefore a brief overview of a complex subject and omits most of the details related to the many and varied forms that antennas can take. The first issue to be emphasized is that the plural of a radio antenna is *antennas*, not antennae. Beetles and other insects have antennae, but radios have antennas. The IEEE standard definitions of terms in electrical and electronic engineering is the reference work that defines how technical terms are to be used in these fields (IEEE 2013). A receiving antenna must collect radio frequency (RF) energy, and a transmitting antenna must launch RF energy into space. From this perspective, an antenna is a kind of coupling device between the transmission line circuit and free space.

The function of a transmitting antenna is to convert radio signals in a waveguide or coaxial cable to radio waves in free space. A receiving antenna must convert incident radio waves to signals in a waveguide or coaxial cable. Antennas are often a critical element of a radio communication link because the gain of the transmitting and receiving antennas are two parameters in the link equation. Higher antenna gain means stronger signals and greater communication capacity in the link; a 6 dB increase in the gain of the transmitting antenna or the receiving antenna can double the capacity of a radio link, measured in bits per second.

All antennas have two defining parameters: *gain* and *beamwidth*. We will define these terms first for a transmitting antenna and then explain how they relate to a receiving antenna. Gain describes the ability of a transmitting antenna to increase the electromagnetic energy directed to a given point, relative to the energy received at that point from an *isotropic* antenna with a gain of one, driven by the same transmitter power. An isotropic antenna radiates equal power in all directions and is used as a reference, but does not exist in practice. Real antennas always radiate more in one direction than other directions.

An *omnidirectional* antenna radiates radio waves equally in all directions, has a gain of one like an isotropic antenna, but does not exist in practice. It is a useful reference when discussing antennas, and can be approximated by some real antennas that need to transmit or receive in all directions. One example is a GPS antenna, which must be able to receive signals from at least four satellites that are located anywhere in the visible sky. The GPS antenna needs to be omnidirectional only in the half space that exists above

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

the earth's surface, and should ideally have a gain of 3 dB (a factor of two). However, in some directions gain will be less than 3 dB, so for link calculations, the gain is often set to one (0 dB). Omnidirectional antennas have dimensions that are usually a fraction of a wavelength, so are physically small.

Gain and beamwidth are inversely related; the higher the gain of an antenna, the narrower its beamwidth. As the gain of a transmitting antenna is increased, more energy is directed toward one point, but some energy is still radiated in all directions. The *antenna pattern* or *polar diagram,* also called a *radiation pattern,* describes the way in which energy is distributed by the antenna into three-dimensional space. An antenna pattern is typically a cut through the three-dimensional radiation pattern in a particular plane. These concepts are illustrated in Figure B.1a for a typical antenna with a gain of 33 dB, plotted on Cartesian coordinates as gain in dB versus angle away from the direction of maximum gain, called the *antenna axis* or *boresight*. The antenna pattern consists of a *main lobe* with maximum gain along the antenna axis, and a series of *sidelobes* extending away from the main lobe in all directions. The peaks of the sidelobes are separated by *nulls* where the radiated power falls to a minimum, but never to zero, which corresponds to negative infinity in decibels. High gain transmitting antennas used in satellite communication links are often required to have a pattern of sidelobes that must lie below a specified *envelope* to minimize interference with adjacent satellites, as discussed in Chapter 8. Antenna designers can control the sidelobe levels of an antenna to some degree, so the first and second sidelobe peaks of an antenna are often quoted to describe the way in which sidelobes decrease with angle away from the antenna axis. Figure B.1b shows the same antenna pattern plotted on polar coordinates. The polar diagram gives a better illustration of the way the lobes are distributed in space, but is not widely used for high gain antennas because it is more difficulty to make measurements in polar coordinates that with Cartesian coordinates.

Figure B.1c, illustrates the 3 dB beamwidth of the antenna, which is the angular distance between the half power points of the antenna pattern, where the gain of the

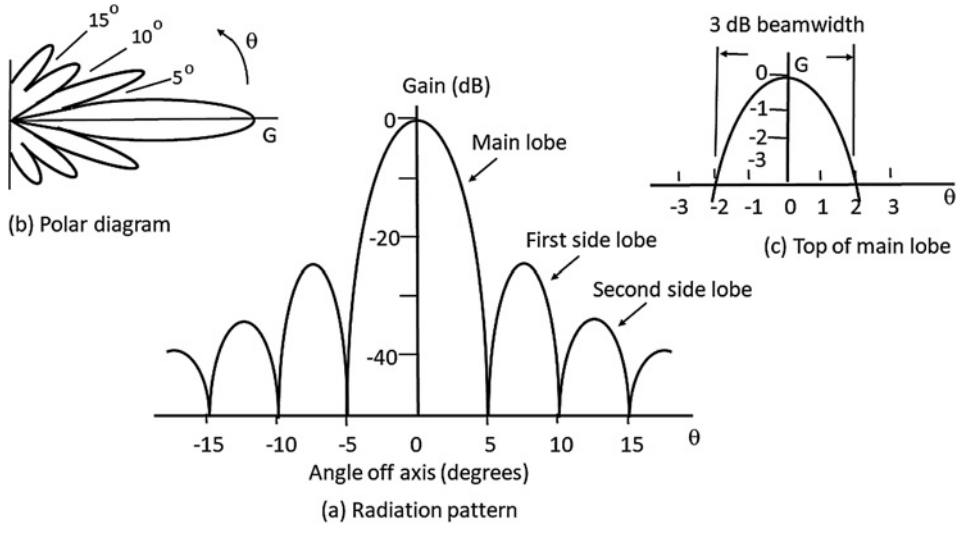


Figure B.1 Radiation pattern for an antenna with a gain of 33 dB. (a) Pattern plotted on Cartesian coordinates. (b) Polar diagram. (c) Top of the main lobe showing the 3 dB beamwidth.

antenna has decreased by 3 dB from its maximum value. In discussing antenna patterns, the maximum gain of the antenna is referenced as 0 dB, and the pattern is plotted in (negative) decibels below the maximum.

The term antenna gain is used rather loosely in general discussion of communication systems, and usually means the maximum gain of an antenna. However, gain is always a function of angle, so the term is also used to describe the gain of the antenna at some angle other than the direction of maximum gain. Antenna experts make a distinction between gain and *directivity*, the difference being the inclusion of internal losses in the antenna, but in this discussion we will assume they are the same.

A basic property of almost all antennas is *reciprocity* which means that the gain, beamwidth, and antenna pattern of a given antenna are identical when transmitting and receiving. It is convenient to describe antennas in terms of their performance when transmitting, and then to assume that those same parameters apply when the antenna is used for the reception of radio signals. Some phased array antennas are non-reciprocal because of the inclusion of phase shifters that have different phase shifts when transmitting and receiving.

High gain antennas are needed in satellite communication links that transfer information at high bit rates so that the capacity of the link is maximized. Antenna gain is directly related to the physical dimensions (in terms of wavelengths) of an antenna. To achieve a high gain the antenna *aperture* must have an area of many square wavelengths. An example is direct to home satellite television, known as *direct broadcast satellite television* (DBS-TV), described in Chapter 10. A DBS-TV receiving antenna must collect the signals transmitted by a direct broadcast satellite and extract a block of TV channels transmitted by one of the satellite's transponders. The bandwidth of the RF signal is typically in the range 20–40 MHz, which sets the noise power in the receiver. The received signal must be greater than the noise power by 2–15 dB depending on the modulation and forward error correction methods employed, which requires the receiving antenna to have a high gain. The discussion in Chapter 10 shows that the gain of the DBS-TV receiving antenna must usually be at least 33 dB to satisfy the CNR requirement in the receiver.

B.2 Gain and Beamwidth

A fundamental relationship in antenna theory is between antenna aperture area and gain. Proving the relationship is difficult, but observation indicates its validity (Stutzman and Thiele 2013). An antenna with an aperture of A square meters operating at a wavelength of λ meters has a gain given by

$$G = \eta_A 4\,\pi A/\lambda^2 \tag{B.1}$$

where η_A is the *aperture efficiency* of the antenna. Note that A and λ must have the same units, meters in this example. The antenna gain in Eq. B.1 is a linear value, not in decibels.

If the aperture is circular with a diameter D m, as is often the case, the area is given by

$$A = \pi r^2 = \pi D^2/4 \text{ m}^2 \tag{B.2}$$

Substituting in Eq. B.1 gives

$$G = \eta_A\,(\pi\, D/\lambda)^2 \tag{B.3}$$

Reflector antennas like those used for DBS-TV reception have aperture efficiencies of 70%, so to achieve a gain of 33 dB, a ratio of 2000, requires an aperture area of A square meters where

$$A = \frac{G\,\lambda^2}{\eta_A \times 4\pi} = 227\,\lambda^2\ \mathrm{m}^2 \tag{B.4}$$

If the DBS-TV system operates in the Ku-band at a frequency of 12 GHz where the wavelength is 0.025 m, the antenna must have an aperture area of 0.142 square meters. With a circular aperture area, the diameter of the antenna is found by rearranging Eq. B.3

$$D = \sqrt{\frac{G}{\eta_A}} \times \frac{\lambda}{\pi} \tag{B.5}$$

Hence the antenna must have a diameter of at least 0.425 m, or 1.43 ft (≈17 in.). A diameter of 0.46 m (18 in.) is common for DBS-TV receiving antennas.

The beamwidth of all antennas is inversely related to the gain of the antenna. For antennas with an aperture, such as *waveguide horns* and reflector antennas, there is a direct relationship between the aperture dimension in a particular plane and the beamwidth in that plane. The relationship can be written as

$$\text{3 dB beamwidth} = B \times \lambda/D \text{ degrees} \tag{B.6}$$

where D is the width of the aperture in the plane under consideration and B is a factor that can vary between 58 and 100 depending on the distribution of electric field in the aperture. D and λ must have the same units. A value of $B = 75$ is often used to estimate the beamwidth of a reflector antenna.

The distribution of electric field in an antenna aperture is called *illumination*, and leads to *illumination efficiency*. Uniform illumination, a constant amplitude and phase across the aperture, gives the highest illumination efficiency of 100%, but cannot be achieved in a reflector antenna because *spillover* of energy from the feed occurs at the edge of the reflector. The combination of many factors that include illumination efficiency, spillover, phase error in the aperture, and blockage by a feed in a reflector antenna sum together to create the aperture efficiency of the antenna, η_A.

B.3 Polarization

Electromagnetic (EM or radio) waves are polarized and consist of an electric field (E) and a magnetic field (H) at right angles to one another, oscillating at the frequency of the EM wave and traveling with the speed of light. Polarization is defined by the direction of the electric field component of the wave. A linearly polarized EM wave has an E field in a specific direction, for example vertical. The wave is then said to be vertically polarized, and an antenna that receives this wave must have its electric field vertical. Wire antennas such as the whip antennas seen on automobiles are vertically polarized. The orientation of a linearly polarized wave can be vertical, horizontal, or any angle in between. A circularly polarized (CP) EM wave has an electric field that rotates about the axis of travel, making one complete revolution in one wavelength distance. Circularly polarized waves are either right hand (RHCP) or left hand (LHCP) depending on the direction of rotation of the electric field, and are often thought of as two linearly polarized waves at right angles to each other and with a 90° phase difference. Circular

polarization is often used in satellite communication systems when the polarization of a received signal may vary. For example, a low earth orbit satellite with a vertically polarized transmitting antenna will produce a signal at an earth station that has a varying angle of polarization as the satellite travels across the sky. Tracking the polarization of a received EM wave is difficult, so circular polarization is a convenient solution, although making an omnidirectional circularly polarized antenna is challenging.

B.4 Low Gain, Medium Gain, and High Gain Antennas

Antennas can be broadly classified by their gain and application. Small antennas have gain less than 20 dB and broad antenna patterns. They are used where a broad beam is required, as elements in phased array antennas, and also as the *feed* for reflector antennas. Medium gain antennas employ a reflector and a feed to achieve gains in the 20–40 dB range, as typified by the DBS-TV example above. These antennas are manufactured in very large quantities for direct broadcast TV applications, and have become low cost devices. A reflector antenna has a beam that points in a specific direction – along the antenna's axis – and the entire structure must be mechanically rotated to move the direction of the beam. This presents some challenges in a LEO satellite system where the satellites cross the sky in a few minutes and two antennas are needed to maintain continuous communication. The beam of a *phased array antenna* can be steered electronically to follow the track of a LEO satellite, and can be repositioned in a fraction of a second to a new satellite. The challenge for LEO satellite systems is to manufacture phased array antennas at an acceptable price for large scale use. Large antennas with high gain from 40 to 65 dB are employed at gateway earth stations. These are invariably reflector antennas, although large phased arrays are used in some radar applications. Achieving gain greater than 65 dB is difficult and expensive, requiring a very large antenna structure, so this represents an upper limit on gain. An antenna with a gain of 65 dB has a very narrow beam, and must be rotated mechanically to follow any motion of a satellite. The following sections discuss the different antenna types that can be used to meet the gain requirements in each of these categories.

B.5 Small Antennas

Small antennas are needed where a broad beam is required, as in the GPS example above. This class includes *monopoles, dipoles,* and *patch antennas,* with gain of a few dB and near omnidirectional radiation patterns, as illustrated in Figure B.2. Y*agi antennas, helix antennas,* and *waveguide horns* have gain in the range 5–20 dB and are illustrated in Figure B.3. Monopole antennas consist of a wire that is usually about one quarter of a wavelength long, mounted above a ground plane, and connected to a coaxial cable. These antennas are widely used on automobiles for reception of radio broadcasts and for two-way radio communications in the vhf and uhf bands. Vertical whip antennas on cars are vertically polarized, while both TV and FM broadcast signals are horizontally polarized, although some FM radio transmissions are circularly polarized specifically to accommodate the vertical antenna on many automobiles. In the middle of the FM band, at 100 MHz, a quarter wave monopole is 0.83 m in length (2.5 ft). The vertical monopole has a broad beam in the horizontal plane, making is ideal for reception of terrestrial radio broadcasts. The ground plane is formed by the metal structure of the

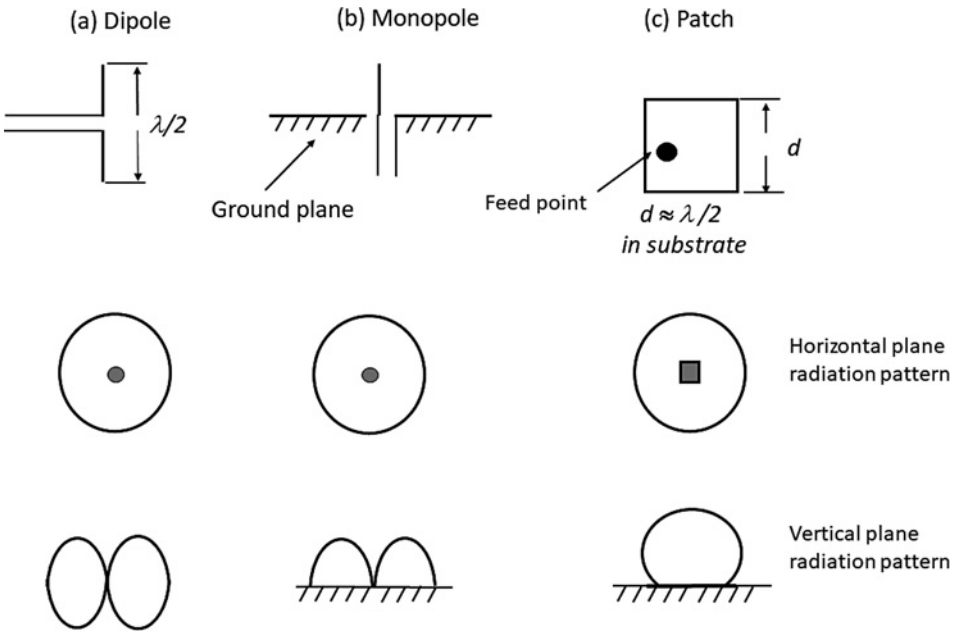


Figure B.2 Small antennas and their radiation patterns. (a) A vertical half wave dipole in free space. (b) A quarter wave monopole over a ground plane. (c) A patch antenna on a horizontal ground plane.

automobile. Where fiberglass is used for a car or aircraft body, wires or wire mesh must be embedded in the surface where a monopole antennas is mounted to provide a ground plane.

A dipole is basically two monopoles pointed in opposite directions and does not need a ground plane. All antennas have a *radiation resistance,* which defines the ratio of current to voltage when the antenna is used for transmission. By reciprocity, the antenna has

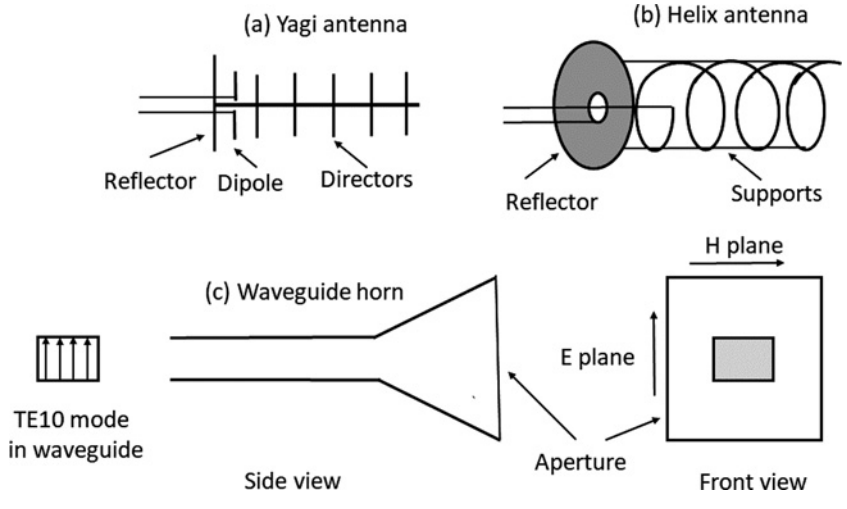


Figure B.3 (a) A Yagi antenna with seven elements. (b) A three turn helix antenna. (c) A rectangular waveguide horn.

the same resistance when receiving. A dipole antenna with length equal to one half wavelength has a radiation resistance of 73 Ω at its resonant frequency, and can be connected directly to a 75 Ω coaxial cable. (As far is the transmitter is concerned, the antenna looks like a 73 Ω resistor. Resistors are often used as dummy antennas for transmitter testing.) Any other dipole length, or monopole length other than a quarter wavelength, has reactive impedance so such antennas are not used as widely as the quarter wave monopole and half wave dipole. The antenna pattern for a dipole is similar to a monopole; both do not transmit or receive in the direction of the wire, so neither is useful for a LEO satellite that flies overhead if the antenna is mounted vertically. Small satellites often use *circular polarization* because of the effect of Faraday rotation in the ionosphere at vhf and uhf frequencies. The quadrafilar helix is a popular circularly polarized antenna for these applications.

Satellites frequently have several monopole antennas at very high frequency (VHF) and ultra high frequency (UHF) to serve telecommand and telemetry links during the launch phase. At least one of the antennas should be able to receive and transmit regardless of the orientation of the satellite. Thin wire monopoles and dipoles have a narrow bandwidth, typically only 1% or 2% of the radio frequency, but can be made *fat* by using rods rather than wires to increase their bandwidth.

Yagi and helix antennas typically have gains in the 5–15 dB range. A Yagi antenna consists of a dipole connected to a coaxial cable and a series of *directors* that are wires or rods cut to slightly less than one half wavelength spaced ahead of the dipole, as shown in Figure B.2. There is a reflector consisting of a wire mesh or sheet metal plate, or a director rod, behind the dipole to reduce radiation in a backward direction. Polarization is in the direction of the dipole and director wires. Yagi antennas can be built with two sets of rods and dipoles at right angles to make a *crossed Yagi*. A network between the two dipoles shifts the phase of one dipole by 90° with respect to the other dipole to make a circularly polarized antenna. The dipoles used in Yagi antenna are often *folded* into a narrow loop. A folded dipole has an impedance of 300 Ω, but in a long Yagi with many director elements, the impedance can be close to 75 Ω.

As so often happens in life, those who deserve credit do not receive it. The antenna known worldwide as the Yagi antenna was invented in 1926 by Professor Shintaro Uda of Tohoku Imperial University, Japan, and his research student Hidetsugu Yagi. Professor Uda wrote eleven papers on the design of the antenna, all published in Japan, in Japanese. Mr. Yagi was more fluent in English than Professor Uda, and published an English language paper describing their work in the IRE Journal in 1928 (Yagi 1928). He also applied for a patent on the antenna without including Professor Uda's name. The patent was later transferred to the UK Marconi company. The antenna became popular for VHF and UHF radars, which were widely employed in WWII. Attempts were made by the IEEE Antennas and Propagation Society to rename the antenna as Yagi-Uda, but have not been successful. Hundreds of millions of Yagi-Uda antennas have been installed on rooftops for terrestrial television reception, but very few of their owners know that credit for the antenna should really go to Professor Uda.

A helix antenna consists of a wire wound on a former, or freestanding, which has a helical shape as seen in Figure B.3. A single turn of the helix has a length equal to

one wavelength at the required RF frequency of operation and the antenna sends and receives circular polarization.

Patch antennas use printed circuit techniques to deposit a layer of copper on a dielectric with a continuous backing plate that acts as a ground plane. A single patch radiates and receives normal to the backing plane, and is therefore useful as an omnidirectional antenna for GPS and satellite radio applications. The size and shape of the patch, usually about a quarter wavelength on a side, and the location of the feed point determine the operating frequency and impedance. The pattern is very broad, making a patch antenna well suited to receiving GPS signals. A low noise amplifier can be built onto the same printed circuit board as the antenna, keeping transmission line loses to a minimum. Adding more patch elements and an electronic phasing network creates a phased array antenna with a beam that can be steered electronically. The successful development of LEO satellite systems for internet access requires a low cost electronically steered phased array antenna that can both transmit and receive, and track LEO satellites as they cross the sky. Patch antennas can be designed to transmit linear or circular polarization.

Waveguide horns are widely used as antennas and as the feed system for a reflector antenna. The function of the feed is to collect EM waves from the reflector surface in receive mode, and to transmit rays toward the reflector when transmitting. The distribution of energy from the feed when it arrives in the reflector's aperture creates the illumination referred to earlier.

A waveguide is a hollow rectangular or circular tube with a conducting inner surface that carries an electromagnetic field in a specific configuration. In a rectangular waveguide, the most common field configuration, called a *waveguide mode*, is known as TE10. The TE10 waveguide mode has the electric field directed across the narrow dimension of the waveguide, as illustrated in Figure B.3. Typical rectangular waveguide dimensions are 0.4 λ in the narrow dimension and 0.9 λ in the wide dimension, internally. These dimensions allow only the TE10 mode to propagate in the waveguide. Larger dimensions, in excess of 0.5 λ in the narrow dimension and 1.0 λ in the wide dimension can support other waveguide modes. The propagation velocity of EM waves in a waveguide is different for different modes, which creates problems with communication signals if there is more than one mode present. Circular waveguides carry a TE11 circular mode, and are used when two polarizations need to be carried by the same waveguide. At microwave frequencies, currents in the walls of a waveguide do not penetrate very far into the surface, a phenomenon known as *skin effect*. Waveguides can be made of plastic or carbon fiber with a thin layer of copper or silver plating.

A waveguide horn is created by flaring the sides of the waveguide to create an aperture, as illustrated in Figure B.3. The broad dimension of a rectangular waveguide, the narrow dimension, or both, can be flared to make a rectangular or square aperture. Gain is set by the aperture area according to Eq. B.1, with aperture efficiency η_A around 80%. In a rectangular waveguide horn, the beamwidths are determined by the aperture dimension in the two orthogonal planes. The electric field is across the narrow dimension of the waveguide, defined as the E plane, and defines the polarization of the horn. The wide dimension is defined as the H plane. The factor B in Eq. B.5 is theoretically 58 in the E plane and 83 in the H plane. However, as the aperture of the horn is made larger, the wavefront becomes curved, gain falls, and beamwidths increase. Loss of gain restricts waveguide horn dimensions to about three wavelengths. With a square horn three wavelengths on a side, aperture area is nine square wavelengths and Eq. B.1 gives an antenna gain of 90, or 19.5 dB, assuming an aperture efficiency of 80%. Beamwidths are expected

to be 20° in the E plane and 28° in the H plane. Waveguide horns are often made rectangular with a larger width in the H plane to equalize the beamwidths.

B.6 Reflector Antennas

When antenna gain greater than 23 dB is required, a reflector antenna is the lowest cost and most widely used configuration. The principle of a reflector antenna is most easily explained using rays, which are the paths of EM waves, or light, which is an EM wave with a much shorter wavelength than microwaves. Figure B.4 illustrates a parabolic reflector antenna with a feed at its focus in receiving mode. Rays from a distant source are parallel when they arrive at the reflector surface and are reflected to the focus of the antenna where the feed collects the EM waves into a waveguide. By reciprocity, a transmitting feed radiates rays toward the reflector that become parallel rays after reflection. A car headlight works on this principle. Almost all reflector antennas are based on parabolic reflectors, since this is the only geometric shape that has a single focal point at which to place a feed. Circular waveguide supporting the TE11 circular waveguide mode can be flared to give a circular aperture horn with similar performance to a rectangular horn, and is often used with circular polarization and where two orthogonal polarizations are to be received. Circular waveguide can be transitioned to square or rectangular waveguide over a distance of a few wavelengths.

There are a number of feeds for reflector antennas based on circular waveguide. The scalar feed, illustrated in Figure B.4, has a flange at the aperture of the waveguide with a series of grooves that control the radiation pattern to make it circularly symmetric. The corrugated horn, also illustrated in Figure B.4, has internal grooves that serve the

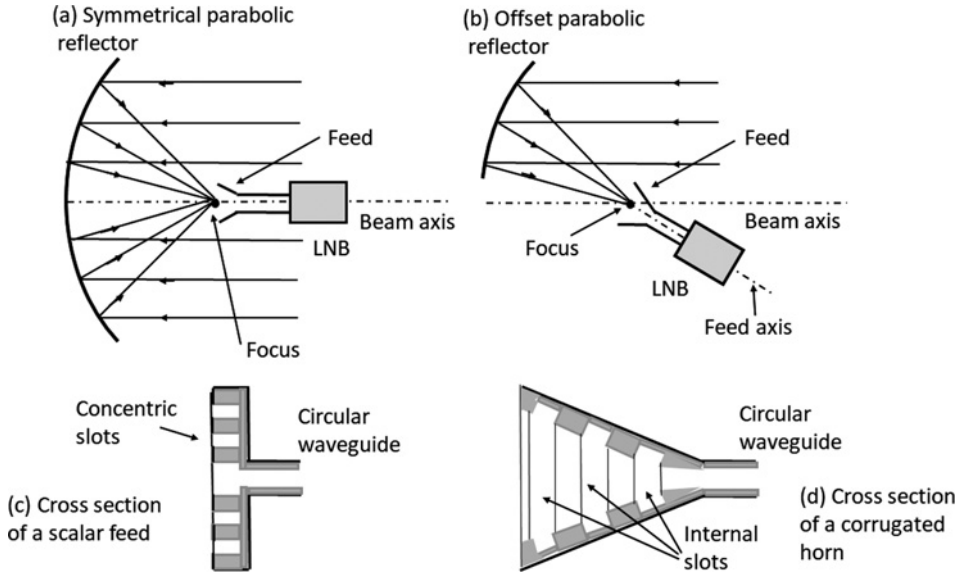


Figure B.4 (a) Symmetrical parabolic reflector antenna. The feed and LNB block part of the aperture. (b) Offset parabolic antenna. The feed and LNB lie below the aperture, avoiding blockage. (c) A scalar feed. (d) A conical corrugated waveguide horn. Both feeds (c) and (d) have circularly symmetric patterns and are often used with dual polarizations.

same purpose. Both types are widely used as the feed element in reflector antennas for satellite communication earth stations, and are manufactured in large quantities using die casting techniques. Metalized plastic castings are also widely used for waveguide components.

A plane across the outer edge of the reflector is called the antenna *aperture*, a term derived from earlier work on lenses used to focus light, as in a telescope. Lenses are not widely used in microwave antennas because they are typically made of plastic materials that absorb energy, which generates noise when receiving and heat when transmitting. The beam formed by a paraboloidal reflector antenna generates a *plane wave* in the aperture because the path length from the focal point to the aperture is equal for all rays. In general, reflector antennas are always designed to produce a plane wave in the antenna aperture. Departure from a plane wave in the antenna aperture, for example by creating a curved wave front, reduces the gain of the antenna, widens its beamwidth, and increases the height of the sidelobes, all of which are undesirable effects. The parabolic reflector with a circular aperture and a feed at its focus is widely used for point to point microwave links. The main disadvantage of this configuration is that the feed system and its supports blocks some of the energy in the aperture, which reduces the aperture efficiency and lowers the antenna gain. The receiving electronics must be mounted behind the feed increasing its weight and blocking the aperture, or a long lossy transmission line is needed to place the receiver behind the reflector. Any spillover at the reflector edges "sees" the earth, which is hot compared with the sky. This means that more noise is introduced into receiving antenna systems pointed down at earth, as is the case for communication satellites. Offset reflector designs overcome this problem, and this is the configuration of choice for DBS-TV antennas used for satellite television reception. The offset reflector design is illustrated in Figure B.4, and a photograph of two early DBS-TV antennas is shown in Figure B.5. The antenna on the left in Figure B.5 has a single feed and can receive TV signals from one satellite only; the antenna on the right has

Figure B.5 Early design of DBS-TV receiving antenna with a circular aperture and a single feed (left) and a dual feed (right). Photo credit: Tim Pratt.

a dual feed and can receive signals from two satellites. The feed, and the LNB located immediately behind the feed lie below the aperture of the antenna and do not block the incoming waves.

Reflector antennas that have gains of 45 dB and greater are needed for gateway earth stations and mobile uplinks used for news gathering and sporting events. Many of these antennas have dual reflector designs that follow the configuration of reflector telescopes built by early astronomers, notably Cassegrain and Gregory. The Cassegrain antenna, illustrated in Figure B.7a is the most widely used dual reflector design. It places the feed system at the vertex of a paraboloid *main reflector*, allowing bulky transmitting equipment to be located behind the main reflector. A *subreflector* with a hyperboloid shape reflects energy radiated by the feed onto the main reflector and thus to the aperture where it forms a plane wave front. A hyperboloid has two foci; when used as a subreflector the feed is at one focus of the hyperboloid and the focus of the main reflector is at the other hyperboloid focus, as illustrated in Figure B.6a. An alternative dual reflector configuration is the Gregorian antenna, illustrated in Figure B.6b. The Gregorian configuration is less widely used than the Cassegrain as it places the subreflector farther away from the feed and therefore requires a heavier support structure. Both configurations require subreflectors that are at least ten wavelengths in diameter to avoid diffraction losses, and because the subreflector blocks the aperture and causes a reduction in aperture efficiency, the main reflector needs to be fifty wavelengths or larger in diameter to keep blockage loss to an acceptable level. Dual reflector configurations tend to be used when antenna gain greater than 40 dB is required. Both dual reflector designs can be offset, as illustrated for an offset Cassegrain antenna in Figure B.6c, to remove the subreflector and feed from the aperture. This configuration is often used for mobile earth stations used for news gathering (satellite trucks) as it can be folded down for transportation, and also for folding antennas against the body of a satellite during launch.

The radiation pattern of a reflector antenna is determined by the electric field distribution in the antenna aperture, which is created by the antenna feed. The field has two parameters, amplitude and phase. Generally, the phase needs to be uniform across the aperture to maximize gain, but the amplitude needs to be reduced, or *tapered,* at the

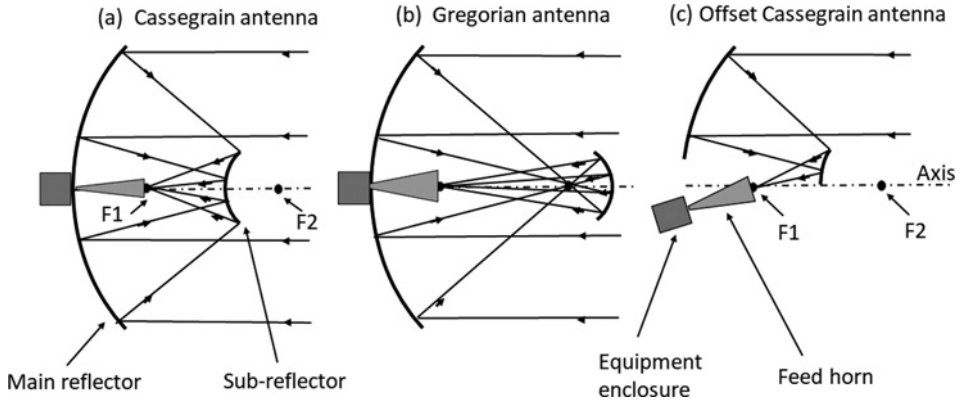


Figure B.6 (a) Symmetrical Cassegrain antenna. The main reflector shape is a paraboloid and the subreflector is a hyperboloid. (b) Symmetrical Gregorian antenna. The main reflector shape is a paraboloid and the subreflector is a hyperboloid set beyond the focus of the paraboloid. (c) Offset Cassegrain reflector. Both feed and subreflector lie below the aperture to avoid blockage.

edge of the aperture to avoid energy from the feed missing the reflector. This is called *spillover* and is undesirable because it reduces the antenna efficiency and sends power behind the main reflector when the antenna is transmitting. When the antenna is receiving, spillover allows interference and thermal noise to enter the feed, both undesirable effects. Near uniform amplitude of the aperture electric field gives the highest antenna gain, so feed systems for reflector antennas are designed to achieve uniform phase in the aperture, nearly uniform amplitude across the aperture, and steep cutoff at the edge of the reflector. The feed system also attempts to produce a circularly symmetric field distribution in the aperture, which is why scalar feeds and corrugated horns are popular. Near uniform illumination of the reflector produces high sidelobes, with the first sidelobe approaching −17 dB with a circular aperture, whereas tapering the illumination reduces the sidelobes. This is an important consideration in earth station antennas that transmit to GEO satellites because the satellites are often spaced two degrees apart in the geostationary orbit. High sidelobes can cause interference to an adjacent satellite, so the radiation pattern of these transmitting antennas is regulated by the ITU. (See Chapter 4 for details.)

The field in the antenna aperture has two parameters, amplitude and phase. In a single reflector configuration, the feed controls both parameters. In a dual reflector configuration there are two reflectors that can be used to control amplitude and phase. By changing the shape of the subreflector and main reflector, uniform phase can be maintained in the aperture by keeping the length of ray paths from the feed to the aperture constant. The shape of the two reflectors can then be adjusted to give near uniform field in the aperture with the desired sharp cut off at the main reflector periphery. This technique of shaping the reflectors in a Cassegrain antenna achieves approximately 1 dB greater gain than with a conventional paraboloid – hyperboloid combination (Galindo 1964; Williams 1965).

Large Cassegrain antennas with gains in excess of 60 dB are used at gateway earth stations to provide high EIRP on the uplink to geostationary satellites and high CNR on downlinks. The beamwidths of such large antennas may be as small as 0.1°, requiring the antenna to be mechanically steered on two axes to follow movement of a GEO satellite and to allow repointing of the antenna to a different satellite. The cost of these large antennas can exceed 1 million dollars, so they are employed only where there is sufficient traffic to justify their high cost. A useful approximate relationship between 3 dB beamwidth and gain is

$$G = 30\,000/(\theta_1 \times \theta_2) \tag{B.7}$$

where G is a linear gain, not in decibels, and θ_1 and θ_2 are the 3 dB beamwidths of the antenna in degrees in two orthogonal planes. If the antenna has symmetrical beamwidths, the product of $(\theta_1 \times \theta_2)$ becomes the square of the 3 dB beamwidth. An antennas with a gain of 50 dB has $G = 100\,000$ and Eq. B.7 gives a 3 dB beamwidth of 0.56°. This beamwidth is sufficiently wide to allow fixed pointing of the antenna toward a geostationary satellite. A mechanical mechanism allows repointing of the antenna to a different satellite, typically by adjustment of the length of the legs that support the reflector. This illustrates a very important property of all antennas: large antennas have narrow beamwidths and small antennas have wide beamwidths. You can choose the gain or the beamwidth, but not both. For example, a satellite antenna designed to produce coverage of the earth from GEO orbit must have a beamwidth of 17°, and will therefore have a gain of 20 dB according to Eq. B.7.

Large antennas, especially those that are fully steerable are expensive. As gain is increased, an antenna must grow in three dimensions, and an approximate relationship between cost and aperture diameter for large steerable antennas has been proposed based on a study in 2008 of the actual cost of a number of large antennas (D'Addirio, 2008).

$$\text{Cost} = \text{US\$260 000} \times (D/12\text{ m})^{2.7} \tag{B.8}$$

where D is the diameter of the antenna aperture in meters. As an example, suppose we know that a large steerable antenna for a gateway earth station has a diameter of 10 m (33 ft). According to Eq. B.8, the cost of the antenna will be US$159 000. To increase the gain by 3 dB requires a doubling of the aperture area, which corresponds to a $\sqrt{2}$ increase in aperture diameter to 14.1 m. According to Eq. B.8 the cost will increase by a factor $1.41^{2.7} = 2.53$, to US$402 000. More sophisticated cost models have been proposed that take account of the frequency at which the antenna operates because higher frequencies require more rigid reflectors to maintain surface tolerance. An upper limit of reflector diameter for C-band fully steerable antennas appears to be about 28 m (85 ft). Eq. B.8 predicts a cost for this antenna of US$2.56M in 2008 dollars. Eq. B.8 applies specifically to steerable antennas with diameters exceeding 5 m. Fixed pointing and repositionable antennas cost much less than fully steerable antennas.

Examples of dual reflector antennas can be seen on the satellites illustrated in Figure 3.3 in Chapter 3, Figure 10.4 in Chapter 10, and Figure 11.2 in Chapter 11. The WDBJ satellite truck in Figure 10.13b has a front fed antenna that folds down for transport. Photographs of earth station antennas are included in Chapters 10 in Figures 10.1 and 10.13a. Illustrations of many large reflector antennas can be found by searching the internet for *satellite earth station antennas*.

B.7 Antenna Theory

Antenna theory explains how the distribution of electric field in the aperture of an antenna leads to the familiar antenna pattern of main lobe and sidelobes, and allows the gain of an antenna to be calculated with accuracy. The earliest analysis was by Airy in 1835 to explain the phenomenon of light and dark rings produced when sunlight was allowed to fall on a pinhole in a sheet and projected onto a dark background. The central bright spot and the bright rings observed by Airy are the projection of the radiation pattern produced by the pinhole aperture onto a plane (Airy 1835). Exactly the same mathematical analysis can be applied to the antenna pattern of a large Cassegrain antenna. In the following section the analysis is confined to a narrow linear aperture, with a summary of results for circular apertures.

Figure B.7 shows a linear aperture with length D lying along the x axis of Cartesian coordinates and width dy, where dy is less than one wavelength. (Think of this antenna as a narrow slot.) The axis of the antenna is along the z direction. We can divide the length of the aperture into small segments dx, again less than one wavelength. At point P, a distance R from the center of the aperture, radiation from each element in the aperture has a different path length unless P lies on the axis of the antenna.

The difference dr in path length to the point P relative to a path from the center of the aperture at $x = 0$ is given by

$$dr = x \sin \varphi \tag{B.9}$$

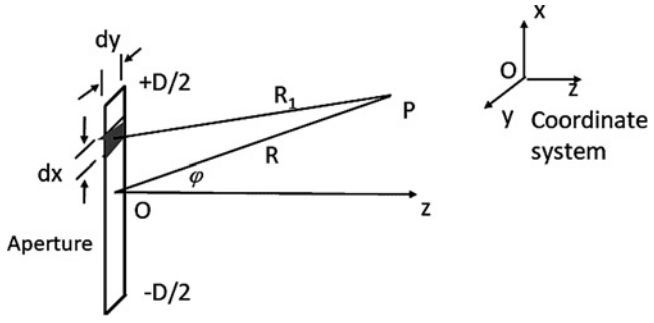


Figure B.7 A linear radiating aperture. The field at point P is the summation of contributions from elements *dx x dy* in the aperture.

where φ is the angle between the antenna axis and a line from the center of the aperture to the point P.

The electric field at point P from an element of field dF in the antenna aperture is given by summation of all the elemental field contributions to the total electric field $E(\varphi)$ at point P. The field contribution dE from the element of electric field dF in the aperture at a distance x from the center of the aperture is

$$dE = E(x) \exp(-jkx \sin \varphi) \tag{B.10}$$

where $E(x)$ is the electric field distribution along the linear aperture and $\exp(-j\,k\,x \sin \varphi)$ is the phase shift to the point P from the specific element dF relative to the phase for an element positioned at $x = 0$. The total field at point P is given by the summation of all field contributions in Eq. B.10 along the length of the aperture. As the length of field segments dx in the aperture becomes small, the summation becomes an integral giving the field at point P as

$$E(\varphi) = \int_{-D/2}^{D/2} E(x) \exp(-jkx \sin \varphi)\, dx \tag{B.11}$$

with $k = 2\pi/\lambda$.

Eq. B.11 is valid only in the *far field* region of the antenna. In the classical definition, the far field begins at a distance $2\,D^2/\lambda$ from the antenna aperture. This is a distance along the axis of the antenna where the phase difference between a ray from the center of the antenna and a ray from the end or edge of the antenna differ by 45°. The radiation pattern of the antenna is assumed to remain constant in the far field, but changes as the distance between the observation point and the aperture is reduced below $2\,D^2/\lambda$. Very close to the aperture, at distances less than $D^2/10\lambda$, the field distribution in space is similar to the field in the aperture. At distances between D^2/λ and $2\,D^2/\lambda$ the main difference observed in the radiation pattern is filling of the nulls between sidelobes. With a large antenna the far field distances can be many kilometers, which poses difficulties when trying to measure the radiation pattern. For example, a 25 m C-band antenna has a far field distance of 25 km at a frequency of 6 GHz. Beacon transmissions from satellites can be used for pattern measurements with large antennas.

If the field in the aperture is uniform and with constant phase, $E(x) = 1$, and the limits of integration are extended to infinity since there is no field beyond the ends of the aperture, then Eq. B.11 becomes

$$E(\varphi) = \int \exp(-jkx \sin \varphi)\, dx \tag{B.12}$$

This is identical to the form of the Fourier transform of a rectangular pulse so we can use the results of Fourier transform theory to find the shape of the radiation pattern of a linear aperture. For the specific case of $E(x) = 1$, known as *uniform illumination,* we know that the result is $\sin x/x$, a sinc function, with a main lobe and sidelobes. The 3 dB beamwidth is 51 λ/D degrees, and first sidelobe level is −13.2 dB. The above results can be extended to a square or rectangular aperture by applying Eq. B.11 to the field distributions in the x and y directions, giving two radiation patterns in orthogonal planes.

A similar approach for an antenna with a circular aperture requires the integration of field contributions from annular rings in the aperture. With uniform illumination, a circular aperture produces a beam with 3 dB beamwidth of 58 λ/D degrees and a first sidelobe level of −17.6 dB. Known results from Fourier transform analysis of waveforms was applied to evaluate Eq. B.12 for different linear aperture distributions. A half cosine raised to the power n was frequently used as an approximation to the field across a reflector antenna aperture when a horn was used as the feed device. The illumination function is

$$E(x) = \cos^n\left(\frac{\pi x}{d}\right) \text{ with } |\, x \,| \leq d/2 \tag{B.13}$$

When $n = 0$ we have uniform illumination of the aperture.

Microwave reflector antennas were developed in the United States at the Radiation Lab during World War II as part of the effort to create microwave radars. The work of the Radiation Lab on antennas was published in 1949 as volume 12 of a series of books describing the work of the Radiation Lab, and has been republished several times (Silver 1983). This was the first text that covered the design, construction, and performance of reflector antennas. The availability of main frame digital computers in the late 1960s made evaluation of the radiation pattern of any reflector antenna feasible by numerical evaluation of Eq. B.11 using the measured radiation pattern of the feed, with much better accuracy than the previous Fourier transform technique. Sophisticated computer

Table B.1 Properties of linear and circular apertures with half cosine illumination function (Pratt et al. 1986)

Power of n	Illumination efficiency	3 dB beamwidth (degrees)	First sidelobe level (dB)
Linear aperture			
0	100%	51 λ/d	−13.2
1	81%	69 λ/d	−23
2	67%	83 λ/d	−32
Circular aperture			
0	100%	58 λ/d	−17.6
1	75%	73 λ/d	−24.6

programs are now available that can analyze the performance of many different antenna types, for example (Antenna Toolbox® by Mathworks®).

B.8 Multiple Beam Antennas

Computer simulation is essential in the design of multiple beam and shaped beam antennas. Shaped beam antennas are used on GEO satellites for direct broadcast television to direct additional power to areas with higher rates of heavy rainfall to reduce the occurrence of outages. (See Chapter 10 for a description of this technique.) Multiple spot beams over the footprint of a satellite provide a large increase in satellite capacity over a shaped beam. ViaSat I launched in 2011 has 72 spot beams and 18 fold frequency reuse providing a capacity of 120 Gbps (ViaSat 2017). Previous GEO satellites with single beam footprints had capacities of less than 10 Gbps. LEO satellites for communications and internet access also use multiple spot beams, with the Iridium satellites being an early example (Fossa et al. 1998).

Multiple beams from a single antenna aperture can be achieved in two ways. A reflector antenna can be equipped with many feed horns located close to its focus. Provided the reflector has shallow curvature, achieved by using a large *f/D ratio,* multiple beams that point a few degrees away from the axis of the antenna can be created with multiple feeds. (*f/D* ratio is the ratio of the focal length *f* of a reflector to its aperture diameter *D*.) Offset fed paraboloidal reflector antennas have been developed for DBS-TV reception that incorporate multiple feeds to allow reception of signals from several closely spaced GEO satellites and are a familiar sight on the roofs of many houses. See Chapter 10 for further details. A second approach is to use a phased array, discussed in the following section, with or without a reflector. A phased array can generate overlapping beams whereas multiple feeds generate side by side beams.

Figure B.8 illustrates an offset Casegrain antenna with a nine horn feed of the type used on a large GEO satellite. The antenna uses reflectors with low curvature to allow horns to be placed well off the feed axis and the subreflector folds in toward the main reflector for launch. This antenna produces nine beams that cover the contiguous 48 states of the United States using four frequencies and two polarizations to minimize interference between adjacent beams. The circles in Figure B.8b represent the 3 dB beamwidth of each of the multiple beams. Beams with the same frequency and polarization are set well apart to avoid excessive interference. Figure B.8 represent a much simplified version of the antennas on high capacity Ka-band satellites, which have a very large number of beams. For example, ViaSat I launched in 2011 generated 72 beams to cover the continental United States, Alaska, and Hawaii, using a phased array feed and separate antennas for transmit and receive. ViaSat 1 and its spot beam footprint are illustrated in Figure 11.2 in Chapter 11 (ViaSat 2017).

B.9 Phased Arrays

An active phased array antenna consists of multiple small elements such as monopoles, dipoles, or patches arranged in a line or a plane to create an antenna aperture. When transmitting, the elements are each driven by a separate power source via a phase shifter. By varying the phase of the signal at each element, the antenna beam can be moved to

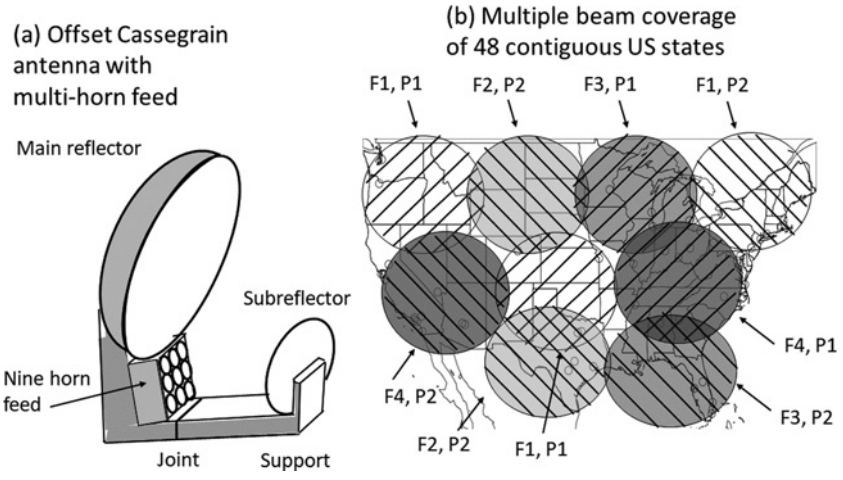


Figure B.8 Illustration of an offset Cassegrain antenna for a large GEO satellite with a nine horn feed producing nine beams to cover the 48 contiguous states of the United States. (a) Offset Cassegrain antenna with nine feed horns and long focal length. The subreflector support hinges at the joint to fold against the main reflector for launch. (b) Nine beams covering 48 states. F is frequency, P is polarization. Overlapping beams use opposite polarization and adjacent beams have different frequencies to minimize interference. Beams with different frequency and polarization are sometimes referred to by color, probably from being plotted in that fashion.

point in different directions. Reciprocity tells us that the antenna will receive signals from the same direction as the transmitted beam provided the phase shifters are reciprocal. Figure B.9 illustrates the principle for a short linear array.

We will analyze a phased array operating in transmitting mode; by reciprocity, the same antenna performance will be obtained when the phased array is used to receive

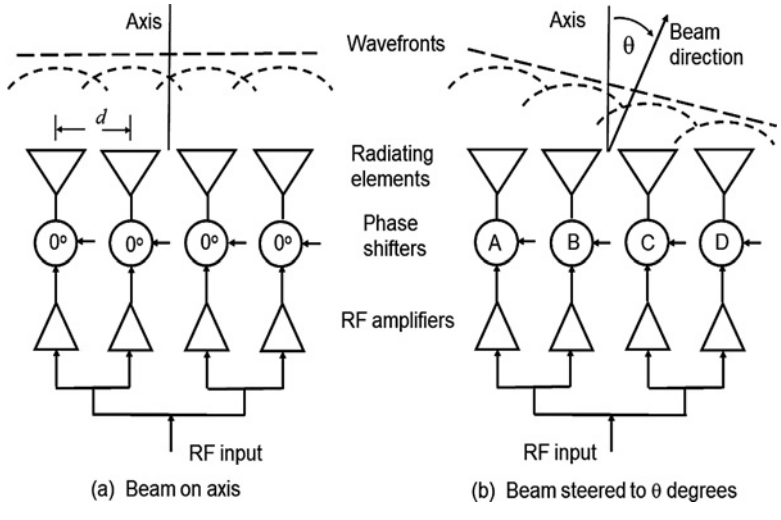


Figure B.9 Four element phased array. Each element radiates a spherical wavefront that sum to form a beam. (a) Phase shifters are all set to 0° to produce a beam along the antenna axis. (b) Phase shifters have a progressively increasing delay along the array to produce an off axis beam.

signals. Figure B.9 shows a four element phased array with an RF amplifier to drive each element. The elements are less than half a wavelength across and are spaced a distance d apart. The transmitters are supplied by a common frequency source and a phase shifter is inserted between each element and its RF amplifier. Each element in the array acts as a *Huygen's source* radiating a spherically expanding wavefront. The elemental wavefronts add together to form a plane wave in the far field. In Figure B.9a, the phase shifters are all set to zero so the transmitted wavefront is parallel to the face of the array and normal to the antenna axis. In Figure B.9b, the phase shifters introduce a progressive phase shift of Φ radians per element across the array, so that element A radiates its wavefront ahead of elements B, C, and D, and element B radiates its wavefront ahead of elements C and D. The result is to create a plane wave in the far field that is angled away from the antenna axis by an angle θ where Φ is the incremental phase shift between elements

$$\sin\theta = \Phi \times \lambda/(2\pi d) \tag{B.14}$$

As an example, let's set the incremental phase shift between elements to $45° = \pi/4$ rad and set $d = 0.6\ \lambda$. Then

$$\sin\theta = (\pi/4 \times \lambda) \,/\, (2\pi \times 0.6\ \lambda) = 0.208$$

giving $\theta = 12.0°$. The phase shifts for the four elements are then A = 0°, B = 45°, C = 90°, D = 135°. If we increase the elemental phase shift to 90° ($\pi/2$), the beam angle becomes $\sin^{-1}$ 0.416 = 24.6° and the phase shifter settings are A, B, C, D = 0°, 90°, 180°, and 270°. Thus the beam radiated by the phased array can be scanned to any angle away from the antenna axis by setting the phase shifters to the appropriate values. There are two limitations on the maximum scan angle that can be achieved with a phased array. If the elements are spaced too far apart, spurious main lobes called *grating lobes* appear in the radiation pattern. Grating lobes appear when the scan angle θ exceeds a value given by

$$\sin\theta = \lambda/d - 1 \tag{B.15}$$

With an element spacing of 0.55 wavelengths, grating lobes appear at a scan angle of 55° away from the antenna axis, so maximum scan angle needs to be restricted to ±50°. With an element spacing of one half wavelength, the scan angle can be increased to near 90° before grating lobes arise. A second problem is more easily understood if the array is in receiving mode. When viewed from an angle θ away from the antenna axis, the array length D is foreshortened to $D\ cos\ \theta$ and therefore intercepts less of the incident wavefront. (See Chapter 9 for a discussion of this effect.)

The main disadvantage of phased arrays is the large number of active RF devices that are required to make up an array with a gain comparable to that of a much simpler reflector antenna. In the example above for a Ku-band DBS-TV receiving antenna the gain required was 33 dB, which was satisfied with an offset paraboloidal reflector with an aperture diameter of 0.425 m, and an aperture area of 0.142 m^2 at a wavelength of 0.025 m. A square phased array with an area of 0.142 m^2 has sides of length 0.377 m. If elements are spaced a half wavelength apart, their separation is 0.0125 m and there are 30 elements on a side for a total of 900 elements. Each element requires a transmit amplifier and a phase shifter as well as a diplexer and low noise receiver when a single array is used for both transmit and receive. Alternatively, separate transmitting and receiving arrays can be used. The same number of elements is required in a Ka-band antenna, but the dimensions are smaller. For example, a square 30 GHz transmit array

with a gain of 33 dB has dimensions 0.3 m × 0.3 m, and a 20 GHz receive array has dimensions 0.45 m × 0.45 m. Placed side by side, the combined transmit-receive antenna has dimensions 0.75 m × 0.3 m (approximately 30 in. by 12 in.). Each transmitting element is required to radiate just over one milliwatt to achieve a transmit power of one watt, which is likely to be the maximum permitted transmit power based on radiation safety and interference considerations. The challenge is to be able to manufacture the phased array at a cost that is not prohibitive. In 2018 there are no active phased array antennas available for DBS-TV home installations, but they may well become available for use with LEO satellite constellations for internet access if the cost can be reduced sufficiently.

Almost all active phased array antennas have been built for military applications where low cost is not the primary objective. They are known as active phased array radars (APARs) and have been employed by the US army, navy, and air force for the detection of hostile aircraft and missiles. Large VHF and UHF active phased array radars for the detection of ICBMs at long ranges have been built by the US Department of Defense at a cost exceeding $100M per installation (Pave Paws 2017). The US Navy Aegis radar systems on guided missile cruisers employs four S-band phased arrays with over 4000 elements in each array. These arrays contain thousands of active elements and consequently are costly to construct. Both of these phased array antennas have the ability to track multiple targets simultaneously (actually sequentially by rapidly switching the beam direction electronically) which justifies their very high cost in critical military applications. In 1996 it was estimated that a Pave Paws antenna cost US$123M (Pave Paws cost 1996).

B.10 Phase Shifters

The phased arrays illustrated in Figure B.9 require an active phase shifter behind each radiating element. Phase shifts in 45° increments are often used in phased arrays by switching different lengths of transmission line in or out of the phase shifter. Lines with lengths of λ/8, λ/4, and λ/2 provide phase shifts of 45°, 90° … 315°, but must somehow be incorporated into an array where the elements are spaced half a wavelength apart. A three bit word is sufficient to select the required phase shift, and either a single computer able to address 900 elements is needed or an individual processor must be located behind each element. Some of the techniques used to produce large flat screen TV displays may be applicable, since they require millions of LEDs that can be addressed individually.

References

Airy, G.B. (1835). On the diffraction of an object-glass with circular aperture. In: *Transactions of the Cambridge Philosophical Society*, vol. 5, 283–291.

Antenna Toolbox by Mathworks® (2019). https://www.mathworks.com/products/antenna.html (accessed 21 June 2018).

D'Addario, L.D. (2008). Jet Propulsion Laboratory SKA TDP Antennas Meeting, 13 March 2008. http://www.http://webcache.googleusercontent.com/search?q=cache:3eUuVh0hP1EJ:www.astro.cornell.edu/SKATDP/files/Meet_2008Mar_AWG/AWG_Mar2008_Talk13_d%27Addario.ppt+&cd=1&hl=en&ct=clnk&gl=us (accessed 30 December 2017).

Fossa, C.E., Raines, R.A., Gunsch, G.H., and Temple, M.A. (1998). An overview of the IRIDIUM low Earth orbit (LEO) satellite system. Proceedings of the IEEE 1998 National Aerospace and Electronics Conference, NAECON 1998. pp 152–159. *doi*:*10.1109/NAECON.1998.710110* (accessed 30 December 2017).
Galindo, V. (1964). Design of Dual-reflector Antennas with Arbitrary Phase and Amplitude Distributions, IEEE Trans on Antennas and Propagation, AP-12, pp. 403–408.
IEEE (2013). https://ieeexplore.ieee.org/document/6758443 (accessed 21 June 2013).
Pave Paws (2017). https://en.wikipedia.org/wiki/PAVE_PAWS (accessed 30 December 2017).
Pave Paws cost (1996). https://www.forecastinternational.com/archive/disp_old_pdf.cfm?ARC_ID=269 (accessed 21 June 2018).
Pratt, T. and Bostian, C.W. (1986). *Satellite Communications*. New York, NY: Wiley.
Silver, S. (ed.) (1983). *Microwave Antenna Theory and Design*. London, UK: IET, Reprint of Volume 12 of the Radiation Lab series published in 1949.
Stutzman, W.L. and Thiele, G. (2013). *Antenna Theory and Design*, 3e. Hoboken, NJ: Wiley.
ViaSat (2017) https://en.wikipedia.org/wiki/ViaSat-1 (accessed 30 December 2017).
Williams, W.F. (1965). High efficiency antenna reflector. *Microwave Journal* 8: 79–82.
Yagi, H. (1928). Beam transmission of ultra-shortwaves. *Proceedings of the IRE* 16: 715–740.

Acknowledgment

The authors wish to thank Dr D.G. Sweeney of Virginia Tech (retired) who suggested the inclusion of an appendix on antennas in this book because many ECE students have a poor understanding of the topic. Dr Sweeney's review and comments on the appendix were very helpful.

Appendix C

Complementary Error Function erfc(*x*) and Q Function Q(*z*)

C.1 Equivalence Formulas and Tables of Values

The complementary error function erfc (*x*) and the Q function Q(*z*) both give the area under the tail of a Gaussian distribution. The parameters *x* and *z* define the lower limit of integration of the Gaussian function, with an upper limit of infinity. The functions are important in digital communications because they define the probability that additive white Gaussian noise with a normalized rms value of 1 V exceeds a threshold set at *x* or *z* volts, giving the probability of a symbol error due to noise (see Chapter 5 for details).

A useful approximation to erfc(*x*) for $x > 1.5$ is

$$\text{erfc}(x) = \frac{|\exp(-u^2)}{\sqrt{\pi}\, u}$$

where $u = \frac{x}{\sigma\sqrt{2}}$ and σ is the rms value of the Gaussian variable (Haykin 1988).

An approximation for Q(*z*) with $\sigma = 1$ for $z > 3$ is (Couch 1990)

$$Q(z) = \frac{1}{\sqrt{2\pi}\, z}\ e^{-z^2/2}$$

The equivalence between erfc(*x*) and Q(*z*) is

$$\text{erfc}(x) = 2Q\left(\sqrt{2}\, z\right)$$

$$Q(z) = \frac{1}{2}\,\text{erfc}\left(\frac{x}{\sqrt{2}}\right)$$

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

Table of Q function Q(*z*)

z	Q(*z*)	*z*	Q(*z*)
0	0.500	5.0	2.872 E-7
2.0	2.280 E-2	5.1	1.701 E-7
2.1	1.791 E-2	5.2	9.981 E-8
2.2	1.394 E-2	5.3	5.799 E-8
2.3	1.075 E-2	5.4	3.372 E-8
2.4	8.220 E-3	5.5	1.902 E-8
2.5	6.227 E-3	5.6	1.073 E-8
2.6	4.674 E-3	5.7	6.000 E-9
2.7	3.476 E-3	5.8	3.320 E-9
2.8	2.562 E-3	5.9	1.820 E-9
2.9	1.871 E-3	6.0	9.979 E-10
3.0	1.354 E-3	6.1	5.310 E-10
3.1	9.702 E-4	6.2	2.827 E-10
3.2	6.889 E-4	6.3	1.490 E-10
3.3	4.847 E-4	6.4	7.778 E-11
3.4	3.378 E-4	6.5	4.021 E-11
3.5	2.332 E-4	6.6	2.058 E-11
3.6	1.595 E-4	6.7	1.057 E-11
3.7	1.081 E-4	6.8	5.236 E-12
3.8	7.252 E-5	6.9	2.603 E-12
3.9	4.821 E-5	7.0	1.281 E-12
4.0	3.174 E-5	7.1	6.244 E-13
4.1	2.070 E-5	7.2	3.014 E-13
4.2	1.337 E-5	7.3	1.440 E-13
4.3	8.558 E-6	7.4	6.816 E-14
4.4	5.423 E-6	7.5	3.194 E-14
4.5	3.404 E-6	7.6	1.482 E-14
4.6	2.117 E-6	7.7	6.709 E-15
4.7	1.303 E-6	7.8	3.098 E-15
4.8	7.948 E-7	7.9	2.396 E-15
4.9	4.800 E-7	8.0	6.226 E-16

Table of function erfc(*x*)

x	erfc(*x*)	*x*	erfc(*x*)	*x*	erfc(*x*)	*x*	erfc(*x*)
0.0	1.00000	1.00	0.15730	2.0	0.167E-3	4.0	1.587E-8
0.05	0.94363	1.05	0.13776	2.1	3.267E-3	4.1	6.889E-9
0.10	0.88754	1.10	0.11979	2.2	2.029E-3	4.2	2.932E-9
0.15	0.83200	1.15	0.10388	2.3	1.237E-3	4.3	1.224E-9
0.20	0.77730	1.20	0.08969	2.4	7.408E-4	4.4	5.012E-10
0.25	0.72367	1.25	0.07710	2.5	4.357E-4	4.5	2.013E-10
0.30	0.67137	1.30	0.06599	2.6	2.515E-4	4.6	7.925E-11
0.35	0.62062	1.35	0.05624	2.7	1.426E-4	4.7	3.060E-11
0.40	0.57161	1.40	0.04771	2.8	7.932E-5	4.8	1.159E-11
0.45	0.52452	1.45	0.04030	2.9	4.331E-5	4.9	4.303E-12
0.50	0.47950	1.50	0.03389	3.0	2.321E-5	5.0	1.567E-12
0.55	0.43668	1.55	0.02838	3.1	1.220E-5	5.1	5.596E-13
0.60	0.39614	1.60	0.02363	3.2	6.297E-6	5.2	1.959E-13
0.65	0.35797	1.65	0.01692	3.3	3.187E-6	5.3	6.727E-14
0.70	0.32220	1.70	0.01621	3.4	1.583E-7	5.4	2.265E-14
0.75	0.28884	1.75	0.01333	3.5	7.713E-7	5.5	7.476E-15
0.80	0.25790	1.80	0.01091	3.6	3.687E-7	5.6	2.420E-15
0.85	0.22933	1.85	0.00889	3.7	1.729E-7	5.7	7.680E-16
0.90	0.20309	1.90	0.00721	3.8	7.951E-8	5.8	2.390E-16
0.95	0.17911	1.95	0.00582	3.9	3.587E-7	5.9	7.291E-17

References

Couch, L.W. (1990). *Digital and Analog Communication Systems*. New York: Macmillan.
Haykin, S. (1988). *Digital Communications*. New York: Wiley.

Appendix D

Digital Transmission of Analog Signals

The basic processes in digital transmission of analog information are sampling, quantizing, and encoding. The principles underlying sampling and recovery of analog signals are routinely presented in beginning courses in communications theory, but often in a mathematical manner that may be difficult to understand.

D.1 Sampling

Nyquist (Nyquist 1924, 1928) devised the sampling theorem, which can be restated as follows: Any analog signal can be completely recovered without distortion with a low pass filter from a series of uniformly spaced samples of the signal, provided that the samples are made at a rate exceeding twice the highest frequency present in the signal.

Surprisingly, there are excellent texts on communication theory that do not have this statement in words, but instead present several pages of mathematics to explain a process that is much more easily understood by a series of pictures. One picture is indeed worth a thousand words in explaining this topic.

The sampling theorem states that a signal may be reconstructed without error from regularly spaced samples taken at a rate f_s samples/second, which is at least twice the maximum frequency f_{max} present in the signal. Instead of transmitting the continuous analog signal, we can transmit the samples. For example, the international standard for telephony voice signals is to filter the signal at baseband to limit its spectrum to the range 300–3400 Hz. (Telephone practice in the United States uses the frequency range 300–3100 Hz.) Thus, one voice channel could theoretically be transmitted with samples taken at 6800 times per second, called the *Nyquist rate*, or as it is usually expressed, with a minimum sampling frequency of 6800 Hz. However, in practice, we must always sample signals at a rate that is 10–25% higher than the Nyquist rate because we use real filters to recover the signal. Speech with $f_{max} = 3.4$ kHz is sampled at 8 kHz for telephony and music with $f_{max} = 20$ kHz is sampled at 44 kHz for compact disks (CDs).

The process of sampling an analog signal is illustrated in Figures D.1–D.8. In Figure D.1a the switch is driven by a unit square wave $s(t)$ with a frequency f_s and period T_s. The multiplication is achieved by using the switch to connect the signal $v(t)$ to the switch output $v_s(t)$. When the switch is closed, the signal passes through the switch and $v_s(t) = 1 \times v(t)$. When the switch is open, there is no output so $v_s(t) = 0 \times v(t) = 0$. Hence the action of the switch is to multiply the signal by a unit square wave at a frequency f_s.

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

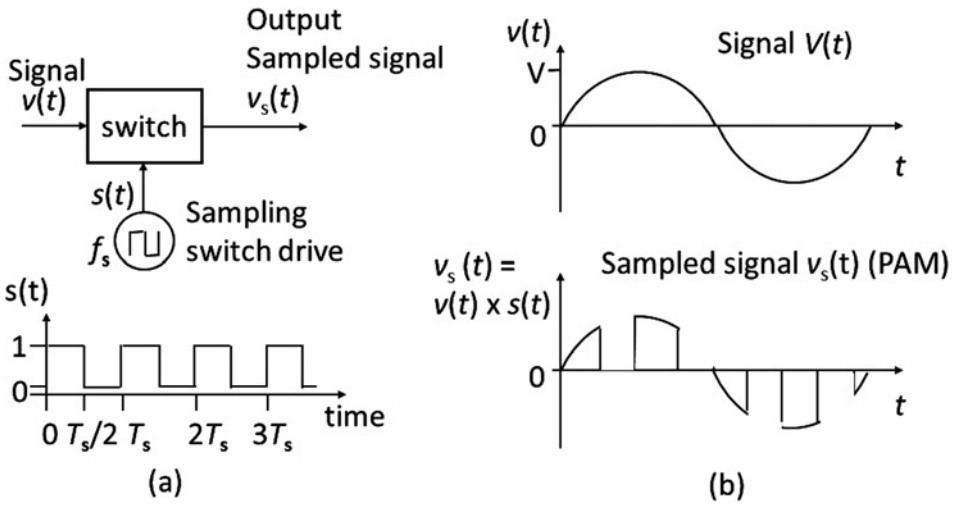


Figure D.1 (a) Sampling of signal $v(t)$ by a square wave $s(t)$. (b) Waveform of signal and sampled signal $v_s(t)$.

Figure D.1b shows the output of the switch when the switch input is a sine wave signal, $v(t) = V \cos \omega_v t$. This produces a chopped waveform with rapid transitions from the sine wave voltage to 0 V, and is termed a pulse amplitude modulated (PAM) waveform. The spectrum of the sampled signal will theoretically extend to infinity, since rapid transitions in any signal always produce a wide spectrum. This process is called *natural sampling* and is not used in practice.

The frequency spectra for the sine wave signal $v(t)$ and the square wave $s(t)$ are shown in Figures D.2a,b. The spectrum of a sine wave is concentrated at a single frequency f_v. The square wave shown in Figure D.1a with amplitude 1.0 V and period T_s, where $T_s = 1/f_s$, has a waveform that can be written as a Fourier series

$$s(t) = 0.5 + \frac{2}{\pi} \times \left(\cos \omega_s t - \frac{1}{3} \cos 3\omega_s t + \frac{1}{5} \cos 5\omega_s t - \cdots \right) \tag{D.1}$$

The spectrum of the square wave consists of single lines in the spectrum corresponding to the odd harmonics of the square wave in Eq. (D.1), as seen in Figure D.2a. Remember that a line in the spectrum is simply a sine wave at that frequency.

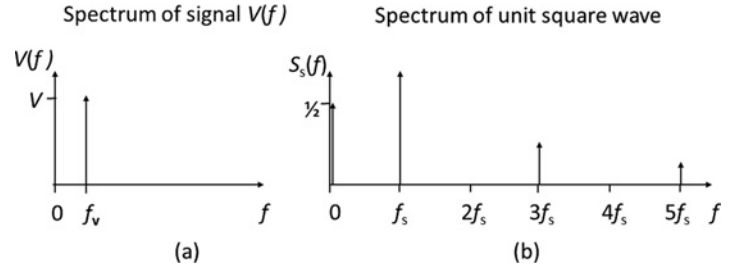


Figure D.2 Spectra of signals in Figure D.1. The narrow line in the spectrum represents a sine wave. The signal has one sine wave at a frequency f_v. The square wave has a DC component at 0 Hz and sine waves at frequencies f_s, $3f_s$, $5f_s$ ….

The output of the switch is the product of the signal $v(t)$ and the square wave $s(t)$ giving the switch output $v_s(t)$ as

$$v_s(t) = v(t) \times s(t) = 0.5\, V \cos \omega_v t + \frac{2}{\pi} V \cos \omega_v t \times \left(\cos \omega_s t - \frac{1}{3} \times \cos 3\omega_s t + \frac{1}{5} \cos 5\omega_s t - \cdots \right) \tag{D.2}$$

Equation (D.2) can be expanded as

$$v_s(t) = 0.5\, V \cos \omega_v t + \frac{2}{\pi} V \times \left(\cos \omega_v t \times \cos \omega_s t - \frac{1}{3} \cos \omega_v t \times \cos 3\omega_s t + \frac{1}{5} \cos \omega_v t \times \cos 5\omega_s t - \cdots \right) \tag{D.3}$$

The product of two sine waves at frequencies f_1 and f_2 consists of two new waveforms at frequencies $f_1 - f_2$ and $f_1 + f_2$, called upper and lower *sidebands*. The spectrum of the sampled signal $S(f)$ consists of the original signal $V(f)$ and pairs of sidebands at frequencies $n f_s \pm f_v$, where n is an odd number from one to infinity.

It is evident from both Eq. (D.2) and Figure D.3 that the original signal $V \cos \omega_v t$ is present in the sampled waveform $v_s(t)$ at the output of the switch. However, there are also sidebands centered at the odd harmonics of the square wave, at frequencies $f_s \pm f_v$, $3f_s \pm f_v$, $5f_s \pm f_v$ … present in the spectrum of the sampled waveform, as shown in Figure D.4.

The DC component present in the spectrum of the square wave preserves the original signal when it is multiplied by the square wave in the switch. The other components of the sampled waveform are unwanted frequencies that are all at higher frequencies than the signal, which suggests that the original signal can be recovered by simply passing the sampled signal through a low pass filter that cuts off all frequencies above the frequency of the original signal. This is illustrated in Figure D.4 where an ideal rectangular low pass filter with a cut off frequency f_c is used to isolate the original signal from the sampled waveform. The condition for undistorted recovery of the original signal $v(t)$ is that f_c must be less than $f_s - f_v$, and that the low pass filter has infinite attenuation above its cut off frequency f_c. To meet this condition requires that the sampling frequency f_s must be equal to or greater than $2f_v$. In the more general case of a real signal that contains many frequencies up to a frequency f_{max}, the condition for undistorted recovery of the original signal is that the sampling frequency must be greater than $2f_{max}$.

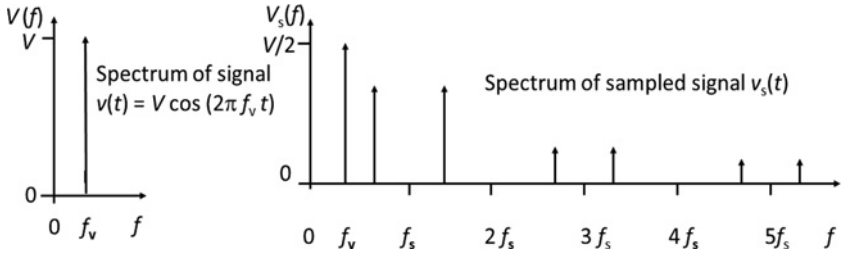


Figure D.3 Spectra of the sine wave signal $v(t)$and the sampled signal $v_s(t)$. The sampled signal contains the sine wave signal $v(t)$ and pairs of sidebands centered on frequencies f_s, $3f_s$, $5f_s$ ….

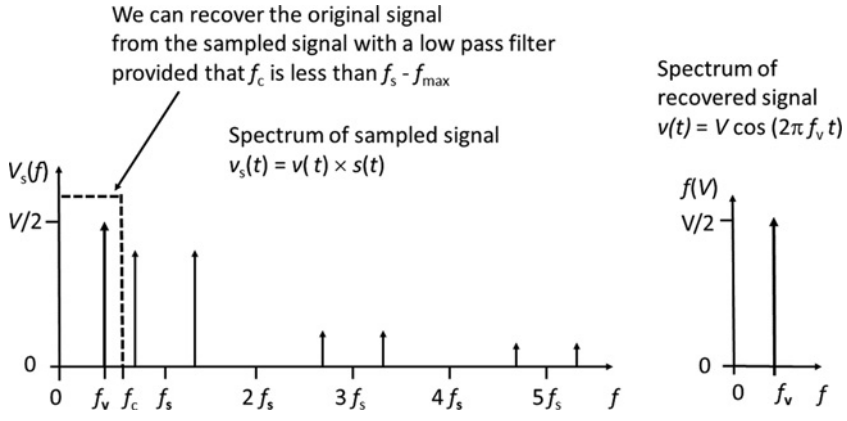


Figure D.4 Recovery of signal $v(t)$ from the sampled signal with a low pass filter. The low pass filter must cut off below a frequency $f_s - f_{max}$, otherwise aliasing will occur.

Mathematically, the condition is that

$$f_{max} < f_s - f_{max} \text{ or } f_s > 2f_{max} \tag{D.4}$$

which proves the sampling theorem.

In real communication systems we do not know the waveform or frequency content of a voice or other analog signal in advance since these parameters are continuously changing. We can require that the spectrum of the signal be restricted to a specific range of frequencies, 300–3400 Hz, for example, for telephony, and enforce the requirement by passing the signal through a filter that removes all frequencies outside the permitted range. Arbitrary signals are represented in the frequency domain by triangles or blocks that indicate the frequency components present in the signal. Figure D.5 shows the spectrum of a voice signal $X(f)$ represented by a triangle that extends from f_{min} to f_{max}. The triangle is not a spectral diagram representing the magnitude of the frequency components present in the signal; it simply indicates the relative location in the spectrum of the low frequency and high frequency portions of the signal.

When we multiply the generic signal with spectrum $X(f)$ by the square wave $s(t)$ we create the spectrum shown in Figure D.5. The spectrum of the sampled signal consists of the original signal that extends from f_{min} to f_{max} and pairs of upper and lower sidebands centered on the sampling frequency and its odd harmonics. For the lowest frequency

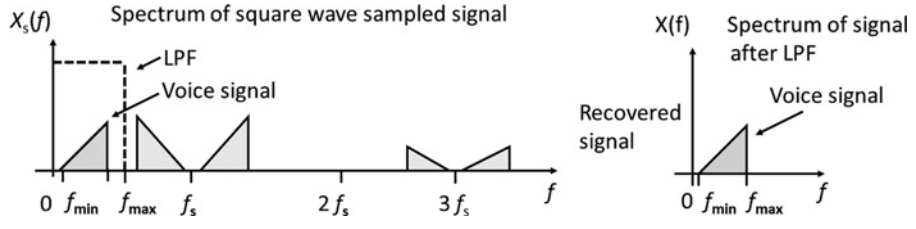


Figure D.5 A generic voice signal, represented by a triangle indicating f_{min} and f_{max}, is sampled with a square wave generating the spectrum on the left of the figure. A low pass filter allows the spectrum of the voice signal to pass, but blocks the harmonic components of the sampling signal. The spectrum on the right of the figure shows that the voice channel is recovered without distortion.

pair of sidebands, centered at the sampling frequency f_s, the lower sideband extends from a low frequency $f_s - f_{max}$ to a high frequency $f_s - f_{min}$, The lower sideband is a spectral replica of the signal spectrum that has been inverted – the lowest frequency is now the highest frequency and the highest frequency the lowest. The bandwidth of the low pass filter that recovers the original signal must be set so that its cut off frequency f_c exceeds f_{max} but does not allow any of the energy in the lower sideband of the sampling frequency at frequency $f_s - f_{max}$ to enter the low pass filter. Thus the criterion for undistorted recovery of the signal with spectrum $X(f)$ from the sampled waveform using a low pass filter with cut off frequency f_c is that f_c must exceed f_{max} but must also be less than $f_s - f_{max}$. Hence the sampling theorem requirement that f_s must be greater than $2f_{max}$.

In practice, we have to use a real low pass filter to recover a signal from a sampled waveform. Real low pass filters do not have an ideal rectangular shape; the attenuation above the cutoff frequency f_c increases rapidly as frequency increases, but not instantaneously. If the sampling frequency is set too low, some of the energy in the lower sideband of the sampled signal will pass through the low pass filter and result in an interfering signal that is present in the recovered waveform as illustrated in Figure D.6. This is known as *aliasing*, and will always occur in practice because no real filter has infinite attenuation in its stop band. In telephony, frequency inversion and shifting of the lower sideband means that the aliasing signal is not intelligible and can treated as noise. Typically, alias components are kept below −40 dB relative to the maximum signal. The description of the sampling process has been discussed here for a voice signal. Music and video signals are sampled and recovered in exactly the same way, but with higher sampling rates.

Natural sampling with a switch driven by a square wave is not used in practice. Instead *instantaneous* sampling is used with a train of very narrow sampling pulses that closes the switch for a short period of time. The very narrow pulse train has a spectrum that differs from that of a square wave only in having a smaller DC component and all harmonics of the sampling frequency. The recovery process with a low pass filter remains the same as for natural sampling. Sampling a baseband waveform with very narrow pulses leaves large time gaps between the pulses, allowing other signals to be inserted. This is not done with analog pulses with varying heights, but after the pulses have been digitized and turned into digital words.

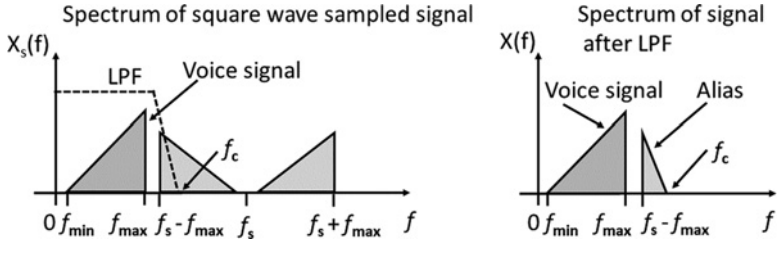


Figure D.6 Aliasing caused by a low pass filter with slow roll off, or by a sampling frequency that is too low for the LPF. The low sampling frequency places the lower edge of the lower sideband of the sampled signal, at a frequency $f_s - f_{max}$, too close to the upper frequency of the signal, f_{max}. Some of the energy in the lower sideband passes through the low pass filter and appears in the right of the figure as an alias component. The sampling frequency needs to be raised in this case, or a better LPF with sharper cut off must be used.

D.2 Bandpass Sampling

The sampling theorem can be extended to signals that are band pass rather than baseband, such as RF or IF waveforms found in radio receivers, leading to a bandpass sampling technique, sometimes called *undersampling*. RF and IF signals are typically centered at a carrier frequency f_c Hz and have a bandwidth B Hz, where $B \ll f_c$. For example, a satellite communications receiver might have an IF frequency of 70 MHz and a signal with a bandwidth of 1 MHz. The sampling theorem requires that the sample rate exceed 140 MHz for accurate reconstruction of the signal to be possible. However, the signal is varying at a rate of 1 MHz, which suggests that it should be possible to sample the signal at a rate greater than 2 MHz. For band pass sampling to be possible, the signal must first be converted to in-phase and quadrature components, which can then be sampled separately. The sampling frequency must be chosen by different criteria that are related to the highest frequency in the signal, and the sample clock has to have much higher stability than for baseband sampling to avoid aliasing. (Glover and Grant 2010) provides a more readable explanation of the process than other texts, and a brief review of undersampling is provided in (undersampling 2018). The sampled signal is recovered with a bandpass filter, rather than the low pass filter used in baseband sampling. Band pass sampling is the basis for digital and software radios, providing the conversion from an analog RF or IF signal to a digital signal.

D.3 Digital Transmission

The second step in the transmission of analog signals in digital form is to convert the samples of the signal waveform into digital words. This is achieved with an *analog to digital converter* (ADC), which produces at its output an N bit word. The N bit word is a binary number that represents the amplitude of the sample at its input. Many different techniques are used in ADCs to generate binary words depending on the speed of operation and the value of N, but will not be discussed here. Search the internet using *ADC* to learn more. The value of N can be as small as three and as large as twenty. Sampling rate can vary from 8 kHz for telephony to 2 GHz for wideband optical communications and radar. For telephony, $N = 8$; with a sampling rate of 8 kHz and the digital output from the ADC is at 64 kbps. The digital words have constant amplitude allowing the signal to be sent over an optical fiber and through a digital computer (*a router*).

Figure D.7 shows a simplified digital transmission system. At the transmitting end of the link, the signal is sampled and converted to digital words. The encoder formats the digital words into packets or frames that allow the receiving end of the link to identify the digital words corresponding to the analog waveform samples. A continuous bit stream consists of ones and zeroes with no identification, so the transmission system must add additional bits that indicate how the data bits are grouped into words; this is achieved by control words in the packets or frames. The encoding process may also add bits that enable error detection and error correction at the receiving end of the link. Chapter 5 discusses the encoding process in more detail. When digital signals are transmitted over a common channel, whether copper wire, fiber optic, or a radio channel, a *protocol* is needed to allow computers to handle the signals. TCP/IP (transmission control

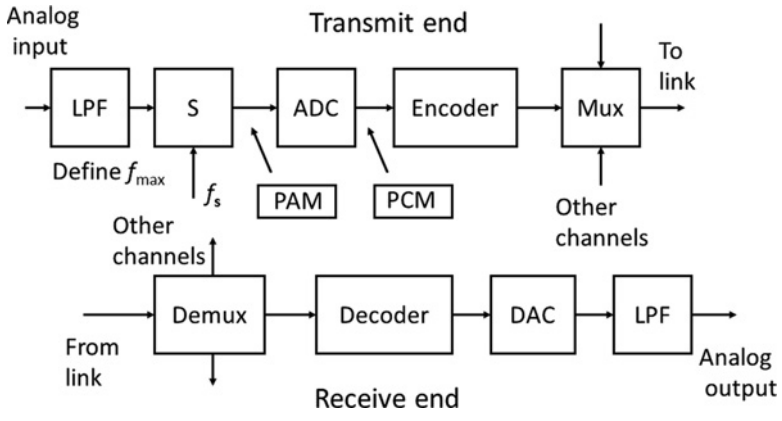


Figure D.7 Simplified diagram of a digital transmission system for analog signals. The low pass filter (LPF) at the analog input limits the signal to a maximum frequency f_{max} and the low pass filter at the end of the receiving system has a cut off frequency f_{max} to recover the analog signal. The sampling frequency must exceed 2 f_{max} to avoid aliasing. The output of the sampler is a pulse amplitude modulated waveform (PAM), which is converted to a series of digital words to form a pulse code modulated (PCM) signal. Other digital signals can be added by the multiplexer (MUX) at the transmit end of the link and recovered at the input of the receiver by the demultiplexer (DeMux).

protocol/internet protocol) is a well known protocol used to send data over the internet. There are many other protocols, designed to meet the requirements of different applications.

Other data streams can be inserted into the link by a process known as *time division multiplexing* (TDM). The insertion device at the transmitting end of the link is called a *multiplexer* (*Mux*) and the device at the receiving end of the link is called a *demultiplexer* (*Demux*). Do not confuse TDM, the process described here, with TDMA, a multiple access method discussed in Chapter 6.

The recovery of the analog signal requires a *digital to analog converter* (DAC) and low pass filter (LPF). The received bit stream is sent to a decoder that removes all the additional bits used for synchronization, control, and error detection and correction, and outputs the N bit words of the original data stream. The DAC converts the digital words back to pulses with amplitudes corresponding to the samples of the analog waveform taken at the transmitter. However, the output of the DAC and LPF is not an exact replica of the analog signal. It a step-wise approximation, which differs from the original analog waveform by a distortion component called *quantization noise*.

The quantization process in the ADC prevents exact reconstruction of the analog waveform in the receiver. The sampling theorem requires that analog (PAM) rather than quantized samples be transmitted to guarantee complete reconstruction of the analog waveform. The error introduced by an ADC is called *quantization error* and a person listening to a reconstructed speech signal perceives the quantization error as an added noise called *quantization noise*. The quantization process is illustrated in Figure D.8 for a much simplified case of an ADC with only six levels that outputs a three bit word.

A uniform quantizer operates with L levels spaced A volts apart. The input signal is amplitude limited to lie between $-A\ (L/2)$ and $+A\ (L/2)$. The quantizer determines in which level an incoming sample falls and puts out the identification number of that level.

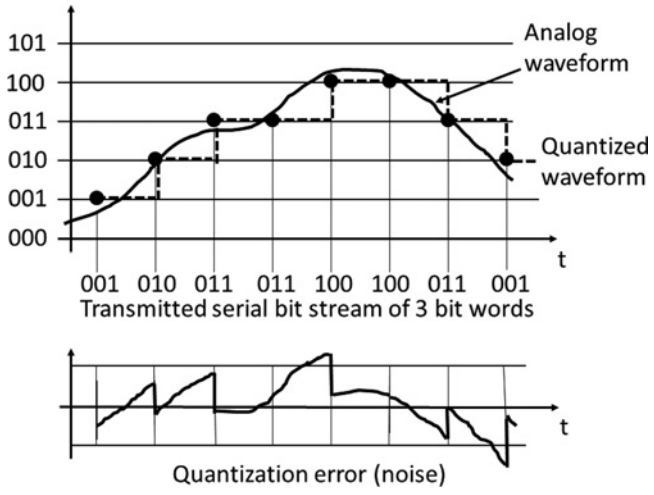


Figure D.8 Illustration of the quantization process for a 3 bit ADC, and typical quantization noise. The ADC in the transmitter selects the nearest digital word to the value of the analog waveform at the sample point, represented by the black dots. The transmitted serial bit sequence is shown below the quantization diagram. The DAC in the receiver outputs the quantized waveform shown by the dotted lines. The difference between the analog waveform and the quantized waveform is quantization error, which is seen as quantization noise at the receiver output added to the analog signal. Quantization noise in an audio signal makes a raspy noise in the background and is quite unlike thermal noise, which makes a hissing sound.

This identification number is the digital word that represents the sample. Transmitting L levels requires N bits where

$$N = \log_2 L \tag{D.5}$$

or

$$L = 2^N \tag{D.6}$$

The levels are normally numbered 0 through $L - 1$. Thus an 8 bit pulse code modulation (PCM) ($N = 8$) system quantizes its incoming samples into one of 256 ($L = 2^N = 256$) levels numbered 0 through 255. Samples of the analog signal are transmitted as binary words ranging from 00000000 (decimal 0) through 11111111 (decimal 255). If the input signal amplitude is uniformly distributed with an rms value of V_{rms}, the signal-to-noise ratio of the reconstructed analog signal (assuming that *only* quantization noise is present) is given by

$$(\text{SNR})_{\text{Q}} = 12\left(\frac{V_{\text{rms}}}{A}\right)^2 \tag{D.7}$$

For uniform quantization and a signal input that has equal probability of any voltage level, the quantization noise added to the recovered analog signal gives a baseband signal to noise ratio of $(\text{SNR})_Q$ where

$$(\text{SNR})_{\text{Q}} \approx 6N \text{ dB} \tag{D.8}$$

Thus a standard digital telephone channel using an 8 bit word and uniform quantization will have an average quantization SNR of 48 dB, using linear quantization.

D.4 Nonuniform Quantization: Compression and Expansion

Uniform quantization introduces more noise when a signal is small and one quantization interval is large in comparison with the signal than it does when the signal is large and one quantization interval is insignificant. The problem is most apparent with the *quiet talker*. The quiet talker produces a voice signal with 30 dB lower power than the design level of the telephone system. As a voltage ratio relative to 1 V, −30 dB is 0.0314 V. A telephone system in which the nominal power level is 0 dBm, and the impedance is 600 Ω (the standard values) has a nominal voltage range of ±0.775 V. With an 8 bit word and 255 quantization levels, the step size is $2 \times 0.775/255 = 6.1$ mV. The quiet talker produces an rms voltage level of 31.4 V rms, with an equivalent peak sine wave level of ±44.4 mV. Thus the quiet talker uses only the lowest 15 steps of the digital quantizer, equivalent to using a 4 bit quantizer. A 4 bit quantizer gives a quantization SNR of 24 dB, so the quiet talker is producing signals that have, at best, a SNR of 24 dB rather than 48 dB.

Improved noise performance can be obtained using nonuniform quantization in which the size of the quantization intervals increases in proportion to the signal value being quantized. The same effect can be obtained from a uniform quantizer if the input signal is compressed before quantization. The distortion introduced by the compressor must be removed at the receiver by an expander. The transfer functions of the compressor and expander are complementary, that is, their product is a constant and the amplitude distribution of a signal that has passed through both a compressor and an expander is unchanged. Compression at the transmitter followed by expanding at the receiver is called *companding*.

Companding was first employed on terrestrial telephone systems using analog compressors that had logarithmic transfer functions. These were the so-called μ-law (North America) and A-law compressors (ITU), which are very similar (Haykin 2001). Later developments in digital technology allowed digital implementation of the compression and expansion functions and permitted the sampling, compression, quantization, and encoding operations to be combined into one integrated circuit called a *coder*. Companding with 8 bit digital voice channels leads to an average SNR of 35 dB (Haykin 2001). Figure D.9 shows the compression characteristic of a typical analog compression circuit. The entire process of converting an analog voice signal to a 64 kbps digital bit stream, and converting a 64 kbps digital voice channel back to an analog voice signal is now done in a single integrated circuit. The telephone wire from a telephone subscriber (the twisted pair of the *last mile*) is taken to a digitizing IC as close to the customer's premises as possible. The digital side of the IC connects to a *four wire circuit,* which has separate

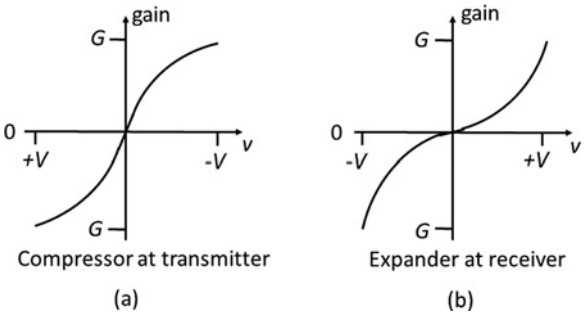


Figure D.9 Gain characteristic of a compressor and an expander in a typical companded link. The compression curve in (a) is logarithmic except close to zero volts where it is linear, because logarithms go to negative infinity with a zero argument. The expander characteristic in (b) is the exact inverse of the compander characteristic, so that the product of the two gains is unity.

go and return pairs. Telephone exchanges (now called *switches*) are digital computers that cannot handle analog signals. All telephone voice signals must be converted to digital form before they can be handled by a switch, so the conversion takes place close to the customer.

The companding process improves the perceived SNR in the baseband channel for the quiet talker by increasing the number of steps in the quantizer at small signal levels. However, with a fixed number of steps (typically 255 in an 8 bit system) the steps must be larger for large signals. This increases the quantization noise that is present with large signals, and therefore lowers the SNR for the loud talker. The effect of companding is therefore to even out the impact of quantization noise over the dynamic range of the baseband signal. When baseband SNR is calculated from signal and noise powers taking companding into account, the SNR is relatively constant with signal level across the whole baseband. However, this is not what the listener perceives. When presented to a human ear, loud sounds appear to mask the increase in quantization noise at high signal levels, and the perceived SNR is much better than the calculated values would indicate.

The reduction in quantization noise for small signals when non-linear quantization is employed is illustrated in Figure D.10. In Figure D.10a, there is a large signal and linear quantization, resulting in large quantization noise steps. In Figure D.10b, the upper diagrams show a small signal with linear quantization. The quantization noise is large compared to the signal, with peak to peak noise almost equal to signal magnitude. The lower diagram shows the effect of non-linear quantization, where small signals have more levels than large signals and quantization noise is much reduced for the small signal. The combination of a linear ADC and the companding curves in Figure D.9 produces the results shown in Figure D.10.

The last mile of copper wire that traditionally has been used to connect telephones to exchanges is steadily being replaced by optical fiber cables in urban areas. Optical fibers

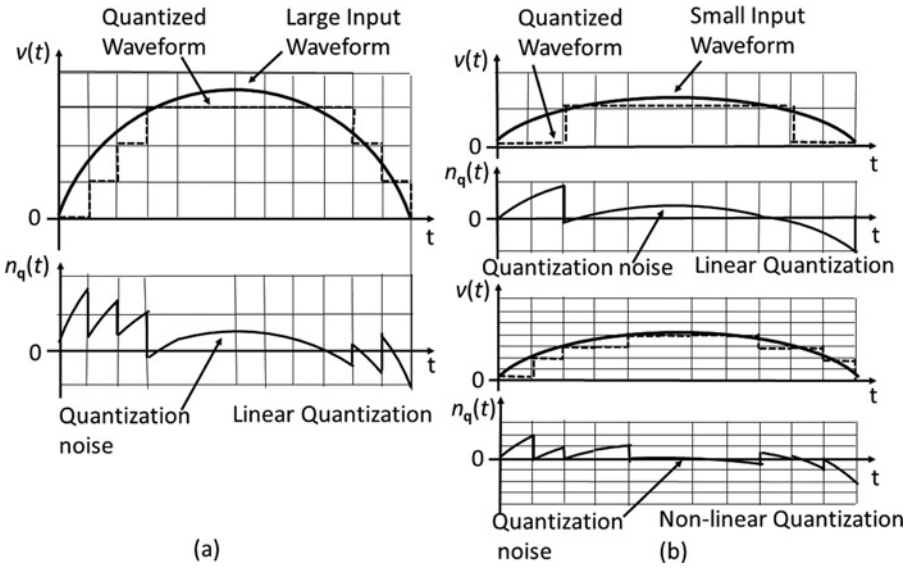


Figure D.10 Illustration of the reduction in quantization noise for small signals when non-linear quantization is employed. (a) Large analog signal. (b) Small analog signal.

have very wide bandwidth, hundreds of megahertz or more, and can carry a wide range of digital signals that can include voice telephone, internet access, and digital (cable) television. The conversion of analog signals to digital signals and digital signals to analog signals takes place in the telephone handset.

The compact disk CD used for sound recordings is another example of a digital audio system. The CD is intended to reproduce music with high quality and therefore requires a much better quantization SNR in the analog sound output than a telephone channel. When a CD is recorded, a 16 bit linear quantizer is used, giving a baseband quantization SNR of 96 dB. The dynamic range of the human ear is about 120 dB, but most sound reproduction systems have thermal noise SNRs of less than 100 dB, so the quantization noise in a CD is inaudible. Linear encoding is more accurate than companding, where a match is needed between the compressor and the expander. Consequently companding is not used when high quality sound reproduction is required. CDs are recorded in stereo using 44 kHz sampling and 16 bit words, giving a bit rate of 1.408 Mbps. The actual bit rate of the recorded material on the CD is much higher because error detection and correction coding is applied to the digital data stream before it is recorded. Digital telephone signals are transmitted at 64 kbps or lower rates, because this is sufficient to achieve speaker recognition and intelligible speech. There is no attempt to make a telephone channel a hi-fi sound system because this would require the transmission of a much higher bit rate, and therefore fewer channels per megabit.

D.5 Reducing the Bandwidth of Digital Signals

The process of converting analog signals to digital signals by sampling, quantization, and serial transmission of bits results in signals that have bandwidths much larger than the analog signal. For example, a voice telephony signal occupies a band from 300 to 3400 Hz. When converted to a standard PCM signal with 8 kHz sampling, the bit rate is 64 kbps. Baseband transmission requires a minimum bandwidth of 40 kHz with low pass SRRC filters having $\alpha = 0.25$. RF transmission using QPSK requires the same bandwidth with $\alpha = 0.25$ band pass SRRC filters. In satellite communication systems and all mobile radio systems, bandwidth is a precious commodity. To conserve bandwidth, *compression* techniques have been developed that reduce the bit rate of the digital signals. Cellular telephone systems use a variety of compression techniques based on *codebook excited linear prediction* (CELP) to reduce digital voice transmission rates to the range 4.8–9.6 kbps, and video transmissions use *motion picture experts group* (MPEG) techniques to reduce television bit rates to the range 2–5 Mbps (Rappaport 2002; Glover and Grant 2010). MPEG 3 is a widely used music compression technique that reduces the high bit rate of a CD recording to 128 kbps for transmission over the internet, with some loss of quality. A brief discussion of compression techniques is included in Chapter 5 of this text.

The conventional calculations of signal to noise ratio in a communication link cannot be applied to compressed digital signals, which generate distortions specific to the signal processing. The quality of a telephone or mobile radio link is described by a *mean opinion score* (MOS) between 0 and 5 that attempts to measure the quality of speaker recognition and intelligibility of the spoken words. A mean opinion score is obtained by having a panel of typical users listen to male and female voices reading lists of words and phrases. The lists include words that are easily confused; for example, *ban* and *van*, *cam*

Table D.1 Mean opinion score (MOS)

Score	Quality rating	Typical application
5	Excellent	Face to face speaking
4 to 5	Good	Wireline telephony using PCM and DPCM
3 to 4	Fair	Cellular telephones using CELP at 4.8 to 9.6 kbps, VOIP
2 to 3	Poor	2.4 kbps compressed speech
1 to 2	Bad	

and *can*, where the difference between the consonants *b* and *v*, *m* and *n* can be difficult to discern when the speaker is speaking in English. Table D.1 provides a summary of MOS quality ratings for several voice communication techniques. Satellite voice communication links carrying international telephone calls should achieve a MOS rating between 4.2 and 4.5. Mobile satcom voice links are likely to have quality similar to cellular telephones with MOS rating in the 3.2–3.8 range. Voice over internet (VOIP) has a wider variation of MOS because of loss of quality when routes are busy and appears to be much less reliable in MOS terms than dedicated voice links. Some test equipment is available to measure MOS on voice links, but there is some disagreement whether the results are comparable to using a listening panel.

The quality of digital video signals can be assessed using the *structural similarity image quality measurement* (SSIM) or the *visual information fidelity* (VIF), a core element of the Netflix *Video Multimethod Assessment Fusion* (VMAF) (Video quality 2018).

References

Glover, I.A. and Grant, P.M. (2010). *Digital Communications*, 3e. Harlow, England: Pearson Education Ltd.

Haykin, S.S. (2001). *Digital Communications*, 4e. Hoboken, NJ: Wiley. Chapter 4.

Nyquist, H. (1924). Certain factors affecting telegraph speed. *Journal of the A. I. E. E.* 43: 124. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1924.tb01361.x (accessed 27 June 2018).

Nyquist, H. (1928). Certain topics in telegraph transmission theory. *AIEE Transactions* 47 (2): 617–644.

Rappaport, T.S. (2002). *Wireless Communications*, 2e. Upper Saddle River, NJ: Prentice Hall.

Undersampling (2018). https://en.wikipedia.org/wiki/Undersampling (accessed 28 June 2018).

Video quality (2018). https://en.wikipedia.org/wiki/Video_quality (accessed 28 June 2018).

Index

Satellite Communications, Third Edition. Timothy Pratt and Jeremy Allnutt.

b

d

e

f

g

m

q

t

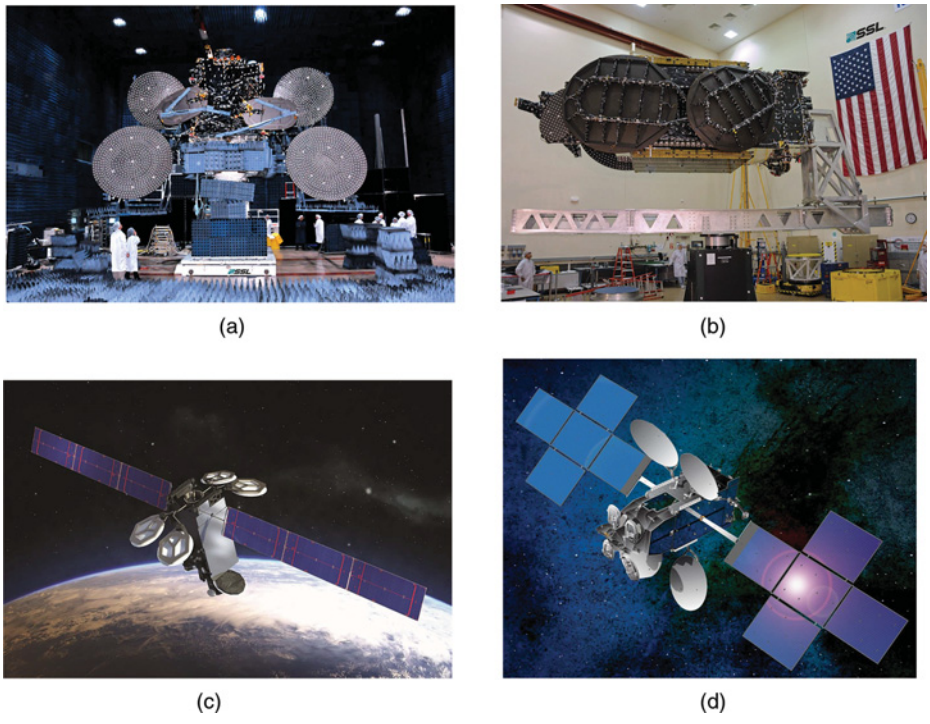


Figure 3.3c Examples of three-axis stabilized communication satellites. (a) A large GEO direct broadcast television satellite built by SSL, under test. (b) Same satellite as (a) folded for launch. The solar cells are folded onto the top and bottom of the body and the antennas are folded against the sides of the body, as viewed in this photograph. (c) Intelsat 35e satellite. (d) ViaSat 1 satellite. Source: Image credits: (a) and (b) Courtesy of SSL, © SSL 2018; (c) © Intelsat, S.A. 2018 and its affiliates. All rights reserved; (d) Courtesy of ViaSat, © ViaSat 2018.

Figure 10.1 Virginia Tech earth station. The two Cassegrain antennas in the left of the photograph are a 9 m C-band steerable antenna and a 5 m Ku-band steerable antenna with dual polarization uplink transmitters. The parabolic torus antenna in the center of the picture is a Simulsat antenna that can receive signals from seven satellites simultaneously. At right is a repositionable Ku-band Cassegrain antenna. Source: Photo credit: Tim Pratt.

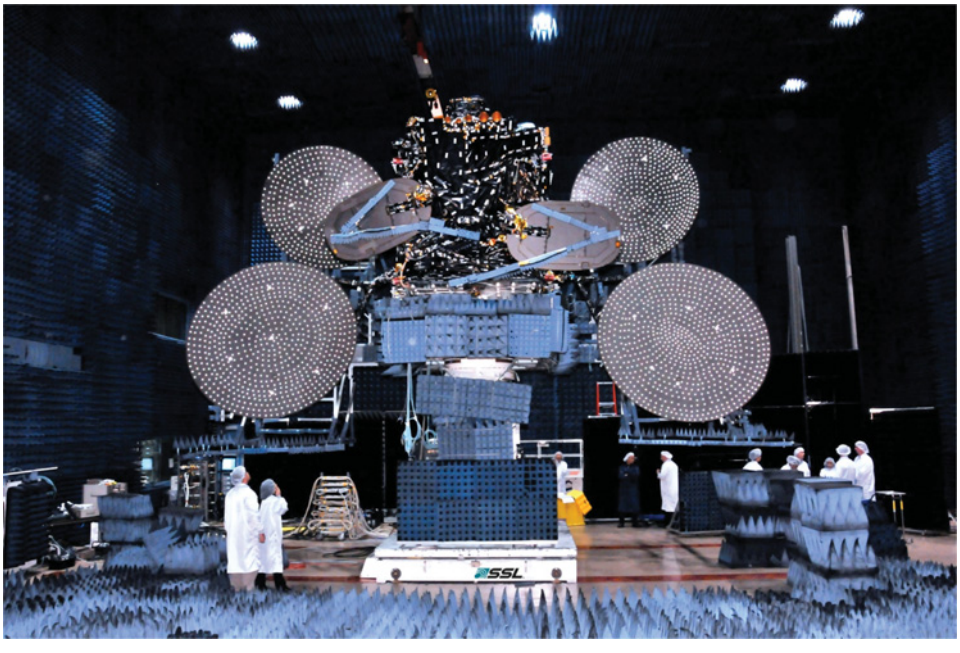


Figure 10.4 A large GEO direct broadcast television satellite under test prior to launch. The four large reflectors create conus beams and spot beams. Source: Image courtesy of SLS, © SLS 2018.

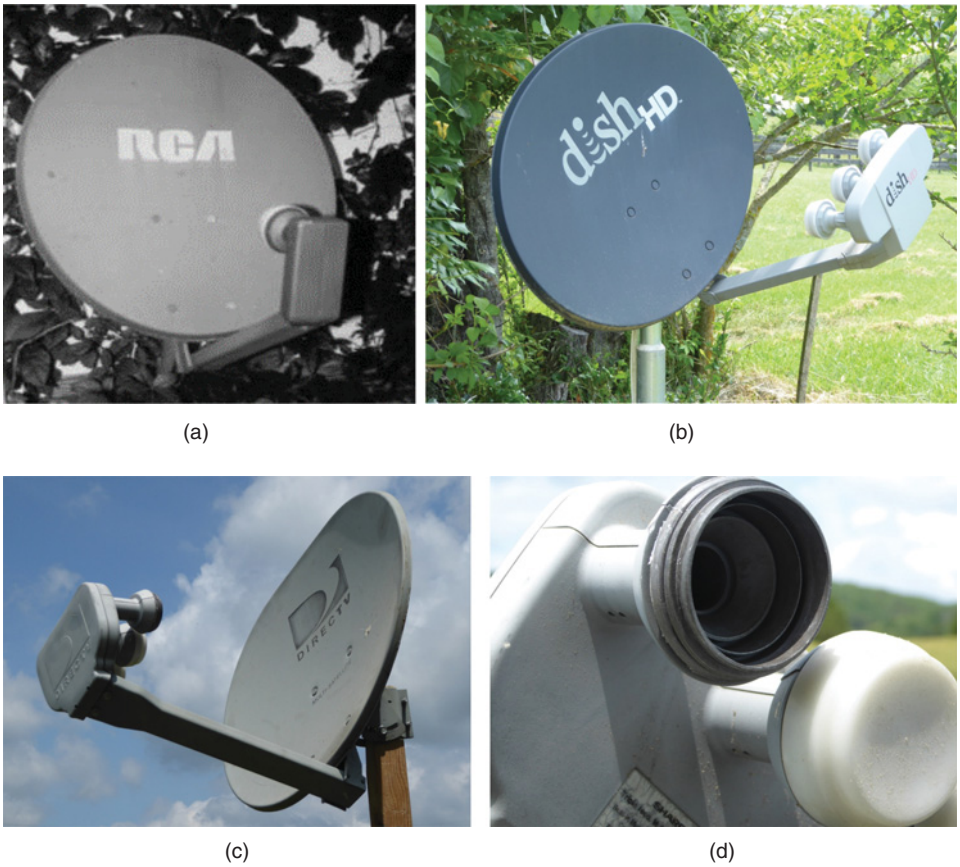


Figure 10.5 Examples of DBS-TV receiving antennas at the author's home (TP) in Blacksburg, Virginia. (a) Early DirecTV antenna from 1996 with a single feed. (b) Dish Network antenna with three feeds, circa 2016. (c) DirecTV antenna with three feeds circa 2013. (d) Corrugated horn of one of the feeds of antenna in (c) with protective cover removed. This is a conical version of the scalar feed illustrated in Figures 10.6a and 10.6b. Note that the single feed dish in (a) has a circular aperture whereas the antennas in (b) and (c) have elliptical reflectors. The wider dimension in the horizontal plane is needed to accommodate the multiple feeds. Source: Photo credit: Tim Pratt.

Figure 10.13a Antennas at the studios of the WDBJ television station in Roanoke, Virginia. The reflector antennas are all used for satellite communications. The center antenna is a Simulsat design capable of receiving signals from multiple satellites. The large antenna at the left of the photograph has a Gregorian configuration with a shaped subreflector. The two antennas on the mast link to WDBJ's broadcast antenna on Poor Mountain.

Figure 10.13b Satellite truck used by the WDBJ television station in Roanoke, Virginia, for outside broadcasts. The main satellite communication antenna is shown in its operating position. Source: Photographs courtesy of WDBJ-TV, © WDBJ-TV 2018.

(a)

Figure 11.2 (a) Illustration of ViaSat 1.

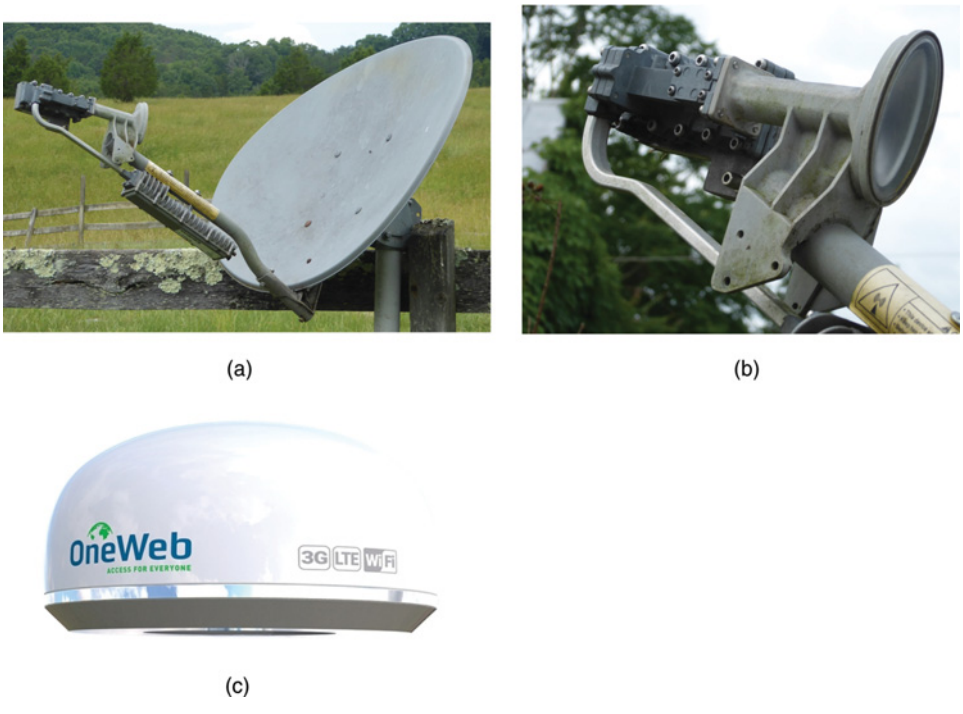


Figure 11.16 Antennas for user terminals and gateways (a) Mid 2000s Hughesnet Ku band user terminal. The HPA is in a finned case below the feed support. (b) Feed horn and LNA for Hughesnet antenna in (a). Note that the elliptical feed is wide in the vertical plane to create a narrow beam in the horizontal plane of the reflector. (c) OneWeb user terminal antenna. Source: Photo credits: (a) and (b) Tim Pratt. (c) Courtesy of OneWeb, © OneWeb 2018.

Figure 12.1 Block IIF GPS satellite. Source: US government.

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.