TECHNOLOGY AND APPLICATIONS SERIES

ELLIOTT D. KAPLAN
CHRISTOPHER J. HEGARTY
EDITORS

UNDERSTANDING GPS/GNSS

PRINCIPLES AND APPLICATIONS

THIRD EDITION

Understanding GPS/GNSS

Principles and Applications

Third Edition

For a listing of recent titles in the
Artech House GNSS Technology and Applications Series,
turn to the back of this book.

Understanding GPS/GNSS

Principles and Applications

Third Edition

Elliott D. Kaplan
Christopher J. Hegarty

Editors

ARTECH
HOUSE

BOSTON | LONDON
artechhouse.com

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the U.S. Library of Congress.

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library.

ISBN-13: 978-1-63081-058-0

Cover design by John Gomes

10 9 8 7 6 5 4 3 2 1

To my children and grandchildren—those that are here and those that may come.

—Elliott D. Kaplan

To my family—Patti, Michelle, David, and Megan—for all their encouragement and support.

—Christopher J. Hegarty

Contributors

John W. Betz, *The MITRE Corporation*
Sunil Bisnath, *York University*
Daniel Blonski, *European Space Agency*
Blake Bullock, *sensewhere Limited*
John Burke, *Air Force Research Laboratory*
Ron Cosentino, *Consultant*
Maarten Uijt DeHaag, *Ohio University*
David W. Diggle, *Ohio University*
Arthur J. Dorsey, *Lockheed Martin Corporation*
Scott Feairheller, *U.S. Air Force*
Peter M. Fyfe, *The Boeing Company*
Christopher J. Hegarty, *The MITRE Corporation*
Len Jacobson, *Global Systems and Marketing*
Elliott D. Kaplan, *The MITRE Corporation*
Mike King, *General Dynamics*
Joseph J. Leva, *Consultant*
Sylvain Loddo, *European Space Agency*
Minquan Lu, *Tsinghua University*
Willard A. Marquis, *Lockheed Martin Corporation*
Dennis Milbert, *NOAA (ret.)*
Samuel J. Parisi, *The MITRE Corporation*
Michael S. Pavloff, *RUAG*
Jun Shen, *Beijing Unistrong Science and Technology Corporation*
Igor Stojkovic, *European Space Agency*
Nadejda Stoyanova, *U.S. Air Force*
Brian Terrill, *U.S. Air Force*
Karen Van Dyke, *Department of Transportation*
Todd Walter, *Stanford University*
Phillip W. Ward, *Navward GPS Consulting*
Lawrence F. Wiederholt, *Consultant*

Contents

APPENDIX C

Preface to the Third Edition

It is hard to believe that it has been 21 years since the publication of the first edition of this book, and 11 years since the publication of the second edition. In the intervening years, the progress of the Global Navigation Satellite System (GNSS) has been staggering. GNSS usage is nearly ubiquitous, providing the position, velocity, and timing (PVT) information that enables applications and functions that permeate our daily lives.

In 1996, when the first edition of this book was published, GNSS included two fully operational satellite navigation systems: the U.S. Global Positioning System (GPS) and the Russian GLONASS. By the time the second edition was published in 2006, GNSS had regressed with respect to the total number of operational satellites due to a decline in size of the GLONASS constellation.

Today, not only is GLONASS back to full strength, but GPS and GLONASS are also being modernized and further GNSS users worldwide are benefitting from the deployment of two more global satellite navigation systems: the Chinese BeiDou and the European Galileo. One regional system—Navigation with Indian Constellation (NavIC)—has been fully deployed, and another is in development, the Japanese Quasi-Zenith Satellite System (QZSS). A myriad of GNSS augmentations are available and provide enhanced performance for those users who require more than the GNSS constellations alone can provide.

The objective of this third edition is to provide the reader with a complete systems engineering treatment of GNSS. The authors are a multidisciplinary team of experts with practical experience in the areas that are addressed within this text. They provide a thorough, in-depth treatment of each topic.

Within this text, updated information on GPS and GLONASS is presented. In particular, descriptions of new satellites, such as GPS III and GLONASS K2 and their respective signal sets (e.g., GPS III L1C and GLONASS L3OC), are included.

New to this edition are in-depth technical descriptions of each emerging satellite navigation system: BeiDou, Galileo, QZSS, and NavIC. Dedicated chapters cover each system's constellation configuration, satellites, ground control system and user equipment. Detailed satellite signal characteristics are also provided.

Over the past two decades, we've heard from many engineers that they learned how GPS receivers work from prior editions of this book. For the third edition, the treatment of receivers is updated and expanded in several important ways. New material has been added on important receiver components, such as antennas and front-end electronics. The increased complexity of multiconstellation,

multifrequency receivers, which are rapidly becoming the norm today, is addressed in detail. Other added features of this edition are the clear step-by-step design process and associated trades required to develop a GNSS receiver, depending on the specific receiver application. This subject will be of great value to those readers who need to understand these concepts, either for their own design tasks or to aid their satellite navigation system engineering knowledge. To round out the discussion of receivers, updated treatments of interference, ionospheric scintillation, and multipath are provided along with new material on blockage from foliage, terrain, and man-made structures.

Since the second edition was published, there have been major developments in GNSS augmentations, including differential GNSS (DGNSS) systems, Precise Point Positioning (PPP) techniques, and the use of external sensors/networks. The numerous deployed or planned satellite-based augmentation system (SBAS) networks are detailed, including WAAS, EGNOS, MSAS, GAGAN, and SDCM, as are ground-based differential systems used for various applications. The use of PPP techniques has greatly increased in recent years, and the treatment in the third edition has been expanded accordingly. Material addressing integration of GNSS with other sensors has been thoroughly revamped, as has the treatment of network assistance as needed to reflect the evolution from 2G/3G to 4G cellular systems that now rely on multiconstellation GNSS receiver engines.

While the book has generally been written for the engineering/scientific community, one full chapter is devoted to GNSS markets and applications. Marketing projections (and the challenge thereof) are enumerated and discussion of the major applications is provided.

As in the previous editions, the book is structured such that a reader with a general science background can learn the basics of GNSS. The reader with a stronger engineering/scientific background will be able to delve deeper and benefit from the more in-depth technical material. It is this ramp-up of mathematical/technical complexity along with the treatment of key topics that enables this publication to serve as a student text as well as a reference source.

Over 18,000 copies of the first and second edition have been sold throughout the world. We hope that the third edition will build upon the success of these, and that this text will prove to be of value to the rapidly increasing number of engineers and scientists working on systems and applications involving GNSS. We wish you, the reader, the very best in your GNSS endeavors!

Elliott D. Kaplan
Christopher J. Hegarty
The MITRE Corporation
Bedford, Massachusetts
May 2017

Third Edition Acknowledgments

Much appreciation is extended to the following individuals for their contributions to this effort. Our apologies to anyone whom we may have inadvertently missed. We thank Don Benson, Sue Carpenito, P. Carril, Frank Czopek, Nigel Davies, A. J. Van Dierendonck, Eddy Emile, Marco Falcone, J. P. Fernow, Jeff Geier, Paul Groves, Kazuma Gunning, Ranwa Haddad, Darrell Judd, Sandy Kennedy, Molly Klemarczyk, Young Lee, Evan Lewis, LaTonya Lofton-Collins, Mike Lombardi, Gary McGraw, Sean McKenna, P. J. Mendicki, Craig O'Grady, Brady O'Hanlon, Gary Okerson, Ed Powers, J. D. Quartararo, Tom Roberts, Jeffry Ross, Logan Scott, Stephen Solomon, Aileen Storry, Sarah O'Rourke, Brian Terrill, Nathan Vary, Mark Walsh, and Jackie Webb.

Elliott D. Kaplan
Christopher J. Hegarty
Editors
Bedford, Massachusetts
May 2017

CHAPTER 1

Introduction

Elliott D. Kaplan

1.1 Introduction

Navigation is defined as the science of getting a craft or person from one place to another. Each one of us conducts some form of navigation in our daily lives. Driving to work or walking to a store requires that we employ fundamental navigational skills. For most of us, these skills necessitate utilizing our eyes, common sense, and landmarks. However, in some cases where a more accurate knowledge of our position, intended course, and/or transit time to a desired destination is needed, navigation aids other than landmarks are used. These may be in the form of a simple clock to determine the velocity over a known distance or the odometer in our car to keep track of the distance traveled. Other navigation aids transmit electronic signals and therefore, are more complex. These are referred to as *radionavigation aids*.

Signals from one or more radionavigation aids enable a person (herein referred to as the *user*) to compute their position. (Some radionavigation aids provide the capability for velocity determination and time dissemination as well.) It is important to note that it is the user's radionavigation receiver that processes these signals and computes the position fix. The receiver performs the necessary computations (e.g., range, bearing, and estimated time of arrival) for the user to navigate to a desired location. In some applications, the receiver may only partially process the received signals with the navigation computations performed at another location.

Various types of radionavigation aids exist, and for the purposes of this text, they are categorized as either ground-based or space-based. For the most part, the accuracy of ground-based radionavigation aids is proportional to their operating frequency. Highly accurate systems generally transmit at relatively short wavelengths and the user must remain within line of sight, whereas systems broadcasting at lower frequencies (longer wavelengths) are not limited to line of sight but are less accurate. The satellite navigation (SATNAV) systems that exist at the time of this writing utilize relatively short wavelengths and are generally highly accurate and line-of-sight-limited. These systems can be augmented to provide enhanced performance as well as to overcome line-of-sight limitations.

1.2 GNSS Overview

Today, there are numerous SATNAV systems operating around the world. Some are global and others only provide service within a certain region. The term *Global Navigation Satellite System* (GNSS) is defined as the collection of all SATNAV systems and their augmentations. (Unfortunately, the term GNSS is also widely used today to refer to any individual global SATNAV system. This book utilizes the original definition, but the reader should be aware of the second definition.) The SATNAV systems discussed within this book are the Chinese BeiDou Navigation Satellite System (BDS), the European Galileo system, the Russian Federation GLObal Navigation Satellite System (GLONASS), the U.S. Global Positioning System (GPS), India's Navigation with Indian Constellation (NavIC), and Japan's Quasi-Zenith Satellite System (QZSS).

The GNSS provides accurate, continuous, worldwide, three-dimensional position and velocity information to users with the appropriate receiving equipment; it also disseminates time within the Coordinated Universal Time (UTC) timescale. Global constellations within the GNSS, sometimes referred to as core constellations, nominally consist of 24 or more medium Earth orbit (MEO) satellites arranged in 3 or 6 orbital planes with four or more satellites per plane. A ground control/monitoring network monitors the health and status of the satellites. This network also uploads navigation and other data to the satellites. With the exception of the radiodetermination service (RDSS) provided by a portion of the BDS, which relies on active ranging to geostationary satellites for positioning, the SATNAV systems discussed within this book provide service to an unlimited number of users since the user receivers operate passively (i.e., receive only). These SATNAV systems utilize the concept of one-way time of arrival (TOA) ranging. Satellite transmissions are referenced to highly accurate atomic frequency standards onboard the satellites, which are in synchronism with an internal system time base. All of the SATNAV systems discussed within this book broadcast ranging codes and navigation data on two or more frequencies using a technique called direct-sequence spread spectrum. Each satellite transmits signals with the ranging code component precisely synchronized to a common timescale. The navigation data provides the means for the receiver to determine the location of the satellite at the time of signal transmission, whereas the ranging code enables the user's receiver to determine the transit (i.e., propagation) time of the signal and thereby determine the satellite-to-user range. This technique requires that the user receiver also contain a clock. Utilizing this technique to measure the receiver's three-dimensional location requires that TOA ranging measurements be made to four satellites. If the receiver clock was synchronized with the satellite clocks, only three range measurements would be required. However, a crystal clock is usually employed in navigation receivers to minimize the cost, complexity, and size of the receiver. Thus, four measurements are required to determine user latitude, longitude, height, and receiver clock offset from internal system time. If either system time or altitude is accurately known, less than four satellites are required. Chapter 2 provides elaboration on TOA ranging as well as user position, velocity, and time (PVT) determination. Present-day commercial user equipment utilizes measurements from multiple SATNAV constellations to form the PVT solution. This ensures signal availability if problems are experienced with one or more SATNAV systems.

Regional SATNAV systems are comprised of the same three segments as the global systems: space, control, and user. The key difference is that the space segment utilizes satellites in geostationary and/or inclined geostationary orbits that provide coverage over the region of interest. The Chinese BDS, NavIC [formerly called the Indian Regional Navigation Satellite System (IRNSS)], and QZSS utilize satellites in these orbital configurations. While the BDS incorporates geostationary and inclined geostationary satellites, it will also have 27 MEO satellites when fully deployed so will provide both a global service and enhanced service within the region surrounding China. (Section 2.3.2 describes these various orbit types.)

1.3 Global Positioning System

Since its inception in the 1970s, the U.S. Global Positioning System (GPS) has continually evolved. System performance has improved in terms of accuracy, availability and integrity. This is attributed to not only major technological enhancements of the three segments: space, control and user but also to increased experience of the U.S. Air Force operational community. Chapter 3 provides details on GPS.

GPS provides two primary services: Precise Positioning Service (PPS) and Standard Positioning Service (SPS). The PPS is an encrypted service intended for military and other authorized Government users. The SPS is free of direct user fees and is in use by billions of civil and commercial users worldwide [1]. Both services provide navigation signals for a user receiver to determine position, velocity and UTC referenced to the U.S. Naval Observatory (USNO).

For the space segment, seven satellite blocks have been developed to date, with each block providing increased capability. At the time of this writing, the GPS constellation consisted of Block IIR, Block IIR-M, and Block IIF satellites. By February 2016, all Block IIF satellites had been launched. The first GPS III satellite was planned for launch in the 2018 timeframe [2]. Figures 1.1 and 1.2 are artist depictions of the GPS Block IIF and GPS III satellites on orbit.

The nominal GPS constellation consists of 24 satellites in 6 MEO orbital planes, known as the baseline 24-slot constellation. For many years, the U.S. Air Force (USAF) has been operating the constellation with more than the baseline number of satellites. In June 2011, the U.S. Air Force formally updated the GPS constellation design to be expandable to accommodate up to 27 satellites in defined slots. This formalized reconfiguration of up to 27 satellites has resulted in improved coverage and geometric properties in most parts of the world [3]. Additional satellites (beyond 27) are typically located next to satellites that are expected to need replacement in the near future.

Improvements have been made to the control and space segments such that the root mean square (rms) value of the space and control segment contribution to ranging error from all satellites in the constellation is approximately 0.5m. The control segment continues to evolve with the Next Generation Operational Control Segment known as OCX planned to become operational prior to 2025.

In terms of user equipment, civil SPS users have a choice of various types of receivers in multiple form factors (e.g., wristwatch, handheld, or mobile phone application). The majority of these utilize signals from GPS and other GNSS constellations.

Figure 1.1 GPS Block IIF satellite. (Courtesy of The Boeing Company.)

Figure 1.2 GPS III satellite. (Courtesy of Lockheed-Martin.)

At the time of this writing, the GPS Directorate continued to oversee the development and production of new satellites, ground control equipment, and the majority of U.S. military user receivers.

1.4 Russian GLONASS System

The Global Navigation Satellite System (GLONASS) is the Russian counterpart to GPS. GLONASS provides military and civil multifrequency L-band navigation services for PVT solutions for maritime, air, land, and space applications both inside Russia and internationally. The form of time provided to users is UTC(SU). GLONASS consists of a constellation of satellites in MEO, a ground control segment, and user equipment. GLONASS is described in detail in Chapter 4. At the time of this writing, there were 24 active satellites and 2 spares. The number of spare satellites is planned to increase to 6. Under the 24-satellite concept, the performance of all 30 satellites will be determined by GLONASS controllers and the

best 24 will be activated. The remaining six will be held for backup or in reserve. Periodically, the mix will be evaluated and, if necessary, a new best set of 24 will be defined. At the beginning of 2017, the GLONASS constellation was populated with two types of spacecraft: Glonass-M, which is a modernized version of the original legacy spacecraft launched from 1982 through 2005, and the newer Glonass-K1 spacecraft design, first launched in 2011. Russia planned to introduce the next generation of spacecraft, Glonass-K2, starting in 2018. Figures 1.3 and 1.4 depict the Glonass-M and Glonass-K1 satellites, respectively.

Both Glonass-M and Glonass-K1 satellites broadcast short- and long-ranging codes and navigation data using frequency division multiple access (FDMA). These satellites also broadcast a code division multiple access (CDMA) ranging code with navigation data, which, at the time of this writing, is serving as a test signal. GLONASS signal characteristics and frequency assignments are contained in Section 4.7.

The Glonass-K satellites carry a search-and-rescue payload (SAR). The payload relays the 406-MHz SAR beacon transmissions that are designed to work with the currently deployed COSPAS-SARSAT system.

GLONASS is supported by a network of ground sites mainly located within the borders of Russia and augmented by monitor sites outside its borders.

GLONASS provides an authorized (military) navigation and a civil navigation service similar to GPS. The Russian government has decreed that the GLONASS open service is available to all national and international users without any limitations. Thus, it is presently incorporated in multiconstellation GNSS single-chip receivers used by millions every day.

1.5 Galileo Satellite System

In 1998, the European Union (EU) decided to pursue a satellite navigation system independent of GPS designed specifically for civilian use worldwide. The development of the Galileo system has followed an incremental approach. Each of the subsequent phases had its own set of objectives. The two major implementation phases

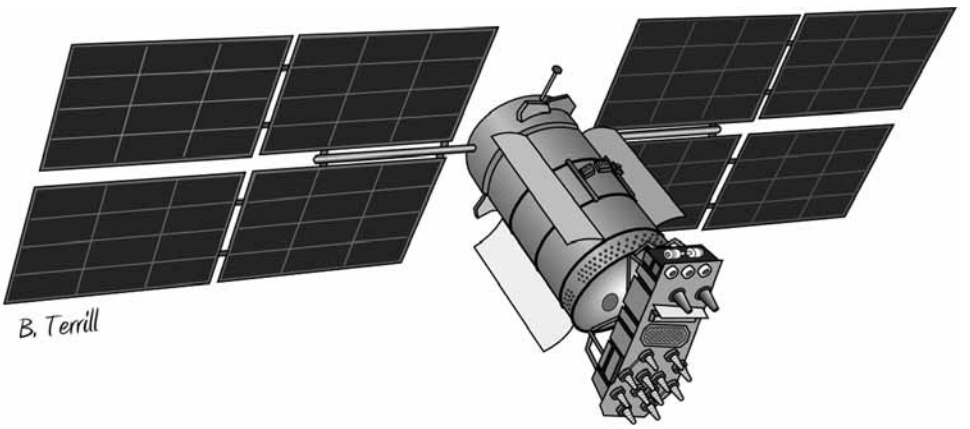

Figure 1.3 Glonass-M satellite. (Courtesy of Brian Terrill.)

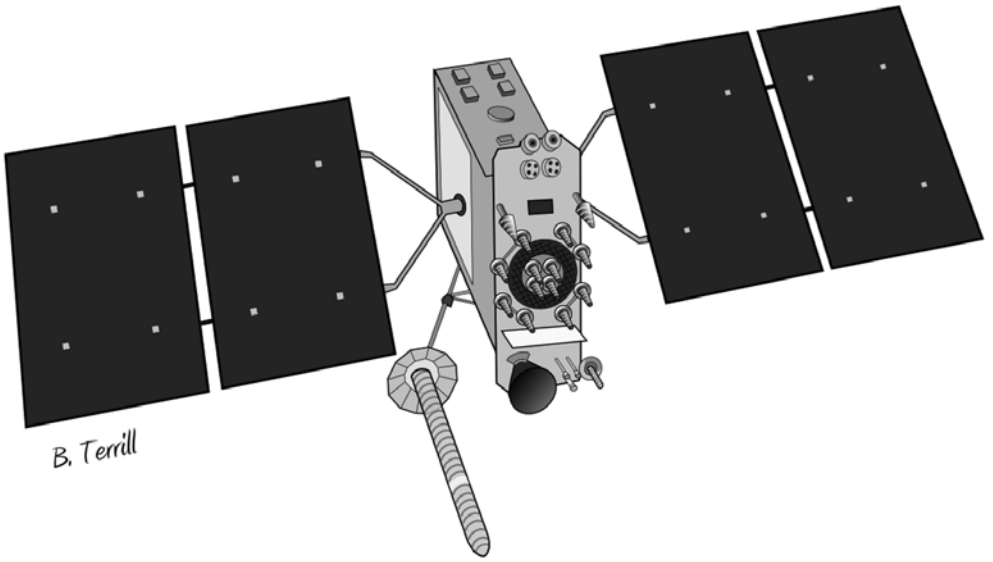

Figure 1.4 Glonass-K1 satellite. (Courtesy of Brian Terrill.)

are the in-orbit validation (IOV) phase and the full operational capability (FOC) phase. The IOV phase has been completed. IOV provided the end-to-end validation of the Galileo system concepts based on an initial constellation of four operational Galileo spacecraft and a first ground segment. Accomplishing a successful service validation campaign, performed throughout 2016, the European Commission (EC) declared the start of the Galileo Initial Services on December 15, 2016.

The system is presently in the FOC phase. FOC will complete the deployment of the Galileo constellation and ground infrastructure and achieve full operational validation and system performance. During the deployment completion, the infrastructure will be integrated and tested in system builds that contain gradually enhanced segment versions, increasing number of remote elements and satellites. The ongoing FOC phase will lead to the fully deployed and validated Galileo system. During this phase, the Galileo system will be handed over in stages to the EC and the European GNSS Agency (GSA)[1] for service provision and exploitation.

When completed, GALILEO will provide multiple levels of service to users throughout the world. Four services are planned: an open service that will be free of direct user charges, a commercial service that will combine value-added data to a high-accuracy positioning service, a public regulated service strictly for government-authorized users requiring a higher level of protection (e.g., increased robustness against interference or jamming), and support for search and rescue.

At the time of this writing, a 30-satellite MEO constellation and a full worldwide ground control segment were in development. Figure 1.5 depicts a Galileo satellite. One key goal is to be interoperable with GPS. Primary interoperability factors being addressed are signal structure, geodetic coordinate reference frame,

1. The European GNSS Agency (GSA) is an agency of the European Union (EU). The GSA's mission is to support EU objectives and achieve the highest return on European GNSS investment, in terms of benefits to users, economic growth, and competitiveness. www.gsa.europa.eu.

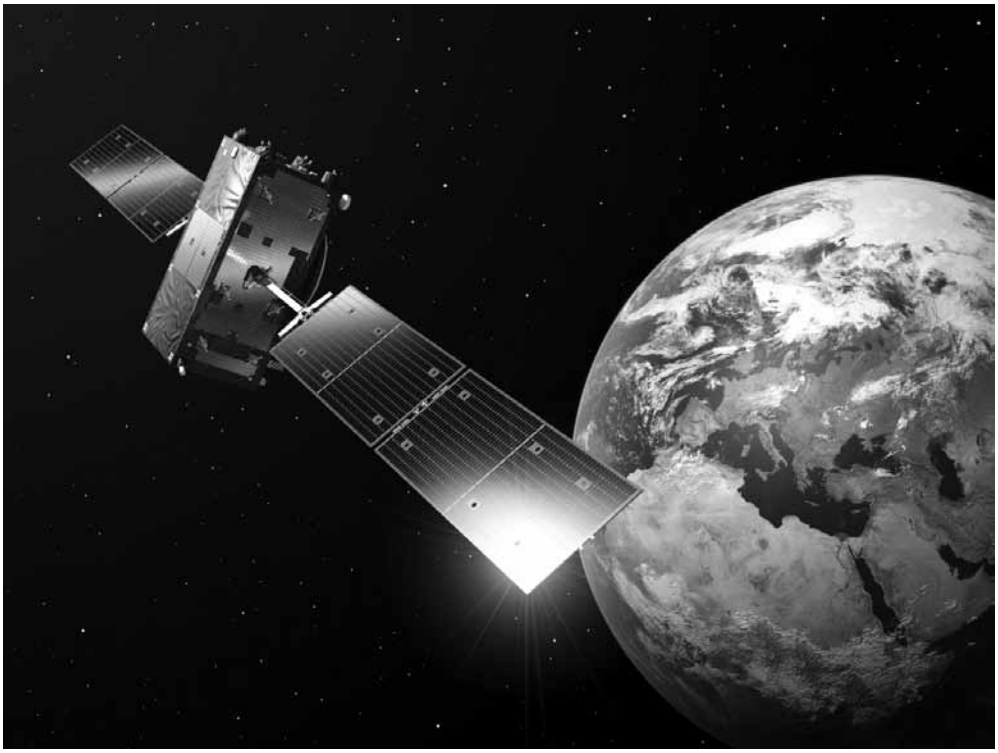

Figure 1.5 Galileo satellite. (©ESA-P. Carill.)

and time reference system. Full operational capability has been planned for 2020. Chapter 5 describes the Galileo system including satellite signal characteristics.

1.6 Chinese BeiDou System

The BDS is a multifunction SATNAV system that integrates many services. Upon its completion scheduled for 2020, BDS will provide global users with PVT services. It will provide a form of UTC traceable to the National Time Service Center (NTSC) of the Chinese Academy of Science denoted as UTC(NTSC). In addition, it will also provide users in China and surrounding areas with a wide-area differential service with positioning accuracy of better than 1m, as well as a short message service (SMS). Those services can be classified as the following three types [4, 5]:

1. Radionavigation satellite service (RNSS): The RNSS comprise the basic navigation services that all GNSS constellations offer, namely PVT. As with other GNSS constellations, using signals of multiple frequencies, BDS provides users with two kinds of services. The open services are available to global users free of charge. The authorized services are available only to authorized users.
2. RDSS: The RDSS is unique to BDS among the GNSS constellations. These services include rapid positioning, short messaging, and precision timing services via GEO satellites for users in China and surrounding areas. This

was the only service type provided by Phase 1 of BDS deployment, BD-1. This functionality has been incorporated into BDS as the system continues to evolve to FOC. With more in-orbit GEO satellites, the RDSS service performance has been further improved with respect to the two GEO satellites in Phase 1.

Since the BDS RNSS offers better passive positioning and timing performance, the SMS is the most useful feature in the RDSS service family, and is widely used for user communications and position-reporting. From the viewpoint of RDSS services, BDS is actually a satellite communication system with SMS services. A user identification number is required for a user to use the RDSS services; hence, the RDSS services belong to the authorized service category.

3. Wide-area differential services: The augmentation systems of other GNSS systems (see Chapter 12) are built independently from their nominal systems. For example, after GPS was deployed, the United States developed an independent augmentation system, Wide Area Augmentation System (WAAS), to meet the demands of the civil aviation industry. The multiple GEO satellites in the BDS constellation make it possible to have an integrated design to combine the nominal services with the augmentation services. As one of the important BDS services, the space-based augmentation system has been designed and developed in parallel with the nominal system in the BDS development process.

The deployment of the BDS global system with 35 satellites (5 GEO, 3 inclined GEO and 27 MEO) is planned to be completed by around 2020 [6]. Figures 1.6 and 1.7 illustrate the BDS GEO and IGSO/MEO satellites, respectively.

1.7 Regional Systems

1.7.1 Quasi-Zenith Satellite System (QZSS)

QZSS is a regional civil SATNAV system operated by the Japan Aerospace Exploration Agency (JAXA) on behalf of the Japanese government. The QZSS constellation currently consists of one satellite in an inclined-elliptical-geosynchronous orbit (denoted as a quasi-zenith (QZ) orbit), providing high-elevation coverage to

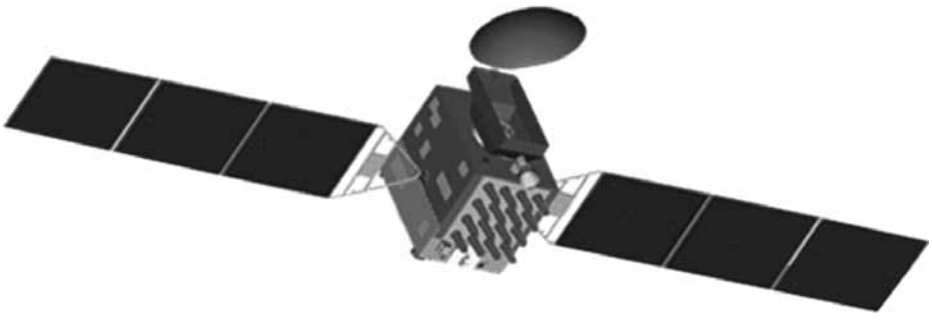

Figure 1.6 BDS GEO satellite [6].

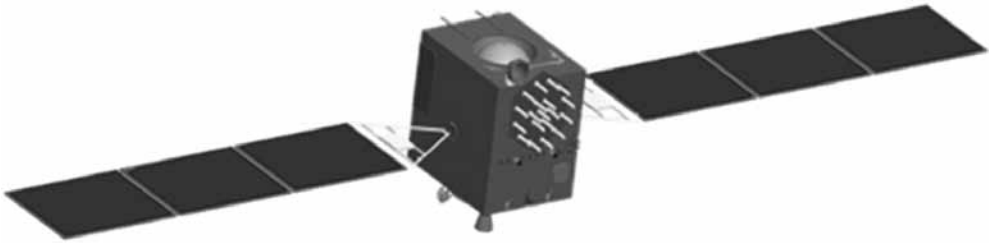

Figure 1.7 BDS IGSO/MEO satellite [6].

complement and augment the U.S. GPS (and potentially other GNSS constellations) over Japan. This QZSS satellite is providing experimental navigation and messaging services. By 2018, plans call for the QZSS constellation to expand to four satellites (one satellite in geostationary orbit and three in QZ orbits), and by 2023 the constellation is planned to consist of seven satellites (one in geostationary orbit, the others in QZ orbits) that will provide independent regional capability in addition to complementing or augmenting other GNSS constellations [7–9]. Figure 1.8 is a depiction of a QZSS satellite.

QZSS is designed to provide three types of services: navigation services to complement GPS, differential GPS augmentation services to improve GPS accuracy, and messaging services for public safety applications during crisis or disasters. As the constellation is completed, QZSS will provide an independent regional navigation capability independent of other GNSS constellations in addition to the current services.

Currently, QZS-1 provides operational services that are being used for a variety of applications in Japan and experimental services which are being tested for future operational use. Planned QZS-2 through QZS-4 satellites will add new experimental augmentation services. Satellites in QZ obits will provide satellite-based augmentation services (SBAS) corrections while the GEO space vehicle (SV) will

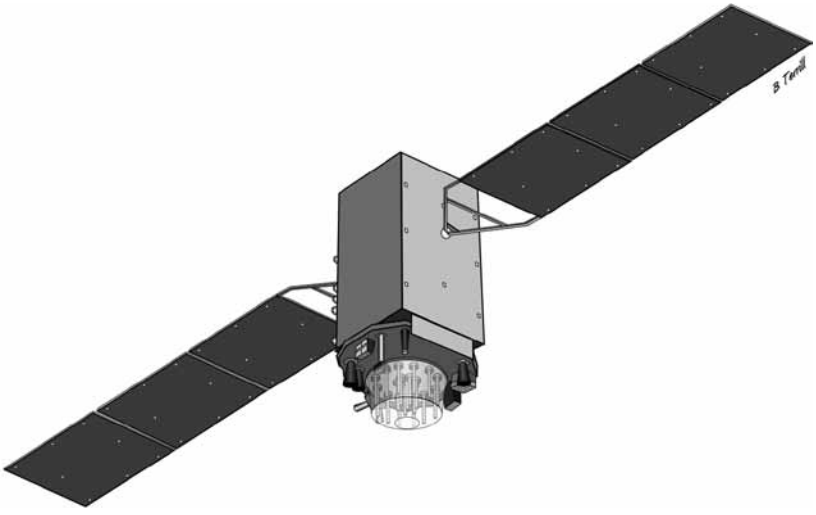

Figure 1.8 QZSS satellite. (Courtesy of Brian Terrill.)

provide S-band messaging services. The navigation and augmentation charges are offered free of any user fees. Section 7.1 provides details on QZSS.

1.7.2 Navigation with Indian Constellation (NavIC)

NavIC is a regional military and civil SATNAV system operated by the Indian Space Research Organization (ISRO) in cooperation with the Indian Defense Research and Development Organization (DRDO) [10, 11]. While other SATNAV systems work primarily in the L-band, NavIC transmits navigation signals in both the L5-band and S-band.

At the time of this writing, NavIC consisted of 3 geostationary and 4 inclined-geosynchronous satellites, ground support segment, and user equipment. The system provides PVT for a region from 30° South Latitude to 50° North Latitude and from 30° East Longitude to 130° East Longitude, which is a region approximately extending about 1500 km around India. A NavIC satellite is depicted in Figure 1.9.

NavIC provides two levels of service, a public Standard Positioning Service (SPS) and an encrypted Restricted Service (RS); both will be available on both L5-band (1176.45 MHz) and S-band (2492.028 MHz) [12–14]. NavIC SPS is designed to support both signal-frequency (L5-band) position fixes using a broadcast ionospheric-correction model and dual-frequency using L5-band and S-band together [15]. A common oscillator provides the timing of both the L5- and S-band signals, thus allowing the receiver to measure the ionospheric delay in real-time and allowing the user equipment to apply corrections. Details of NavIC are contained in Section 7.2.

1.8 Augmentations

Augmentations are available to enhance standalone GNSS performance. These can be space-based such as a geostationary satellite overlay service that provides satellite signals to enhance accuracy, availability, and integrity or ground-based as in a network that assists embedded GNSS receivers in cellular telephones to compute a rapid position fix. The need to provide continuous navigation between the update

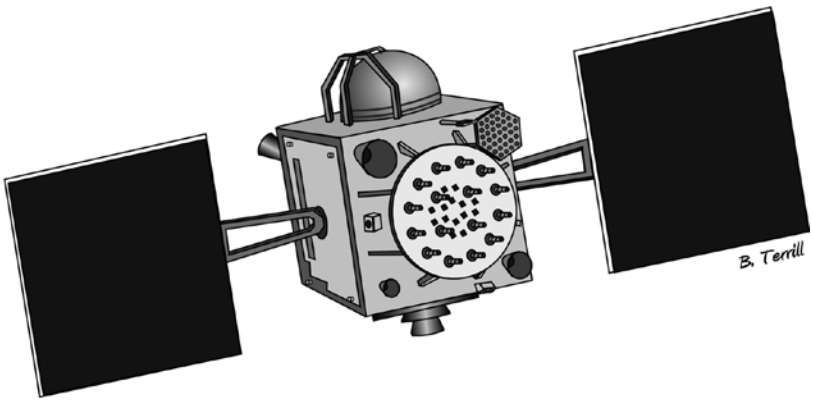

Figure 1.9 NavIC (IRNSS) satellite. (Courtesy of Brian Terrill.)

periods of the GNSS receiver, during periods of shading of the GNSS receiver's antenna, and through periods of interference, is the impetus for integrating GNSS with various additional sensors. The most popular sensors to integrate with GNSS are inertial sensors, but the list also includes dopplerometers (Doppler velocity/altimeters), altimeters, speedometers, and odometers, to name a few. The method most widely used for this integration is the Kalman filter.

In addition to integration with other sensors, it can also be extremely beneficial to integrate a GNSS sensor within a communications network. For example, many cellular handsets now include embedded GNSS engines to locate the user in the event of an emergency, or to support a wide variety of location-based services (LBS). These handsets are often used indoors or in other areas where the GNSS signals are so highly attenuated that demodulation of the GNSS navigation data by the handset takes a long time or is not possible. However, with network assistance, it is possible to track weak GNSS signals and quickly determine the location of the handset. The network can obtain the requisite GNSS navigation data from other GNSS receivers with a clear-sky view or other sources. Further, the network can assist the handset in a number of other ways such as the provision of timing and a coarse position estimate. Such assistance can greatly increase the sensitivity of the GNSS sensor embedded in the handset enabling it to determine position further indoors or in other environments where the GNSS signal is highly attenuated. Chapter 13 covers both integration of GNSS with other sensors and network-assisted GNSS.

Some applications, such as precision farming, aircraft precision approach, and harbor navigation, require far more accuracy than that provided by standalone GNSS. They may also require integrity warning notifications and other data. These applications utilize a technique that dramatically improves standalone system performance, referred to as differential GNSS (DGNSS). DGNSS is a method of improving the positioning or timing performance of GNSS by using one or more reference stations at known locations, each equipped with at least one GNSS receiver to provide accuracy enhancement, integrity or other data to user receivers via a data link.

There are several types of DGNSS techniques and depending on the application, the user can obtain accuracies ranging from millimeters to decimeters. Some DGNSS systems provide service over a local area (10–100 km) from a single reference station, while others service an entire continent. The European Geostationary Navigation Overlay Service (EGNOS) and Indian GAGAN system are examples of wide area DGNSS services. Chapter 12 describes the underlying concepts of DGNSS and details a number of operational and planned DGNSS systems.

1.9 Markets and Applications

Today's 4 billion GNSS deployed devices are projected to grow to over 9 billion by 2023. That is more than one unit for every person on Earth. It is anticipated that while the United States and Europe will grow at 8% per year, Asia and the Pacific Region will grow at 11% per year. The total world market is expected to grow about 8% over the next 5 years due primarily to GNSS use in smart phones and location-based services. Revenues can be broken into core elements like GNSS

hardware/software sales and the enabled revenues created by the applications. With these definitions, annual core revenue is expected to be just over €100 billion ($90 billion) by 2020. Enabled revenue stays fairly flat at €250 billion ($225 billion) over the period, but is estimated to rise dramatically after 2020 as Galileo and BeiDou reach full operational capability [1]. Figure 1.10 shows the projected growth of the installed base of GNSS receivers and Figure 1.11 shows the growth of GNSS devices per capita. The projected global GNSS market size through 2023 is shown in Figure 1.12.

GNSS revenue growth between now and 2023 was estimated to be dominated by both mobile users and location-based services as shown in Figure 1.13.

Applications of GNSS technology are diverse. These range from navigating a drone to providing a player's position on a golf course and distance to the hole. While most applications are land-based such as providing turn-by-turn directions using a smartphone, there are also aviation, maritime, and space-based usages. Further discussion on market projections and applications is contained in Chapter 14.

1.10 Organization of the Book

This book is structured to first familiarize the reader with the fundamentals of PVT determination using GNSS. Once this groundwork has been established, the SATNAV systems mentioned above that comprise the GNSS are described. Each description provides details of the system architecture, geodetic and time references, services and broadcast navigation signals.

Next, the discussion focuses on how a GNSS receiver is actually designed. A step-by-step description of the design process and associated trades required to design a GNSS receiver depending on the specific receiver application is put forth. Each stage of a creating a GNSS receiver is described. Details of receiver signal acquisition and tracking as well as range and velocity measurement processes are provided.

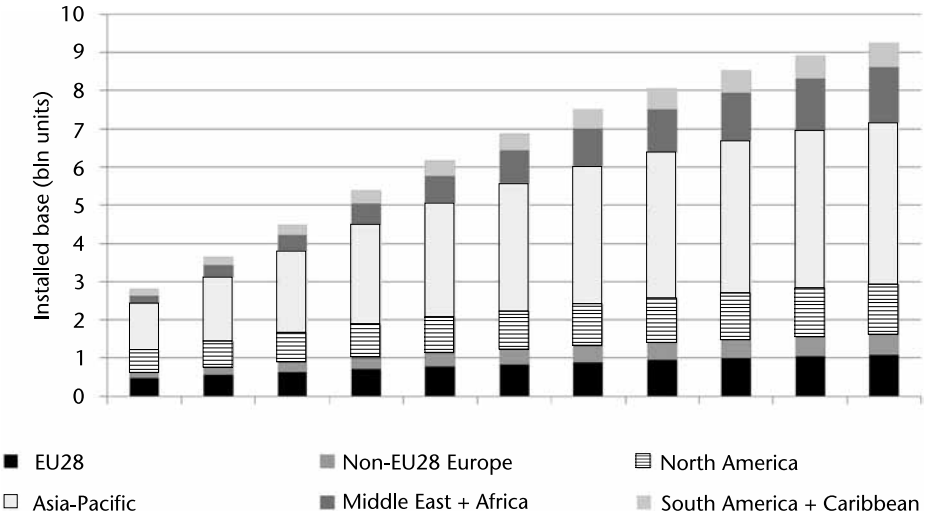

Figure 1.10 Installed base of GNSS devices by region. (Courtesy of GSA.)

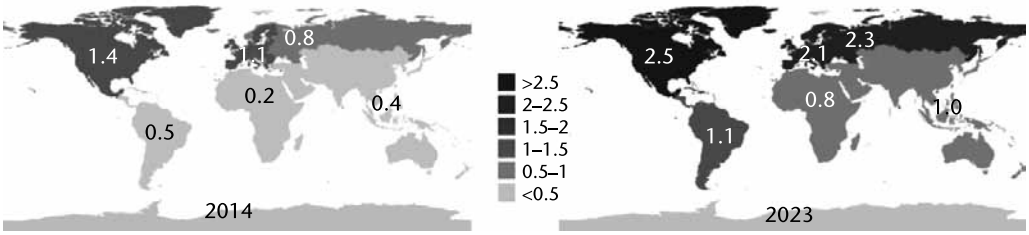

Figure 1.11 GNSS devices per capita: 2014 and 2023. (Courtesy of GSA.)

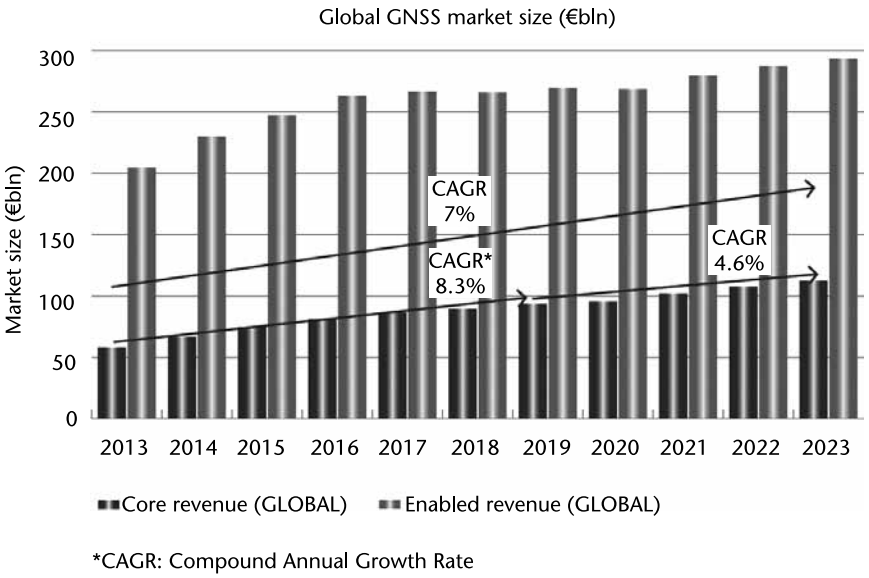

Figure 1.12 Global GNSS market size (billions of Euros). (Courtesy of GSA.)

Signal acquisition and tracking is also analyzed in the presence of interference, multipath and ionospheric scintillation. GNSS error sources are examined followed by an assessment of GNSS performance (accuracy, availability, integrity, and continuity). GNSS differential techniques are then covered. Sensor-aiding techniques including automotive applications and network-assisted GNSS are presented. Finally, information on GNSS applications and their corresponding market projections is discussed. The highlights of each chapter are summarized next.

Chapter 2 provides the fundamentals of user PVT determination. Beginning with the concept of TOA ranging, the chapter develops the principles for obtaining three-dimensional user position and velocity as well as UTC from a SATNAV system. Included in this chapter are primers on GNSS reference coordinate systems, Earth models, satellite orbits, and constellation design. This chapter also provides an overview of GNSS signals including commonly used signal components. Background information on modulations that are useful for satellite radionavigation, multiplexing techniques, and general signal characteristics including autocorrelation functions and power spectra is covered.

In Chapter 3, details of GPS are presented. These include descriptions of the space, control (i.e., worldwide ground control/monitoring network), and user

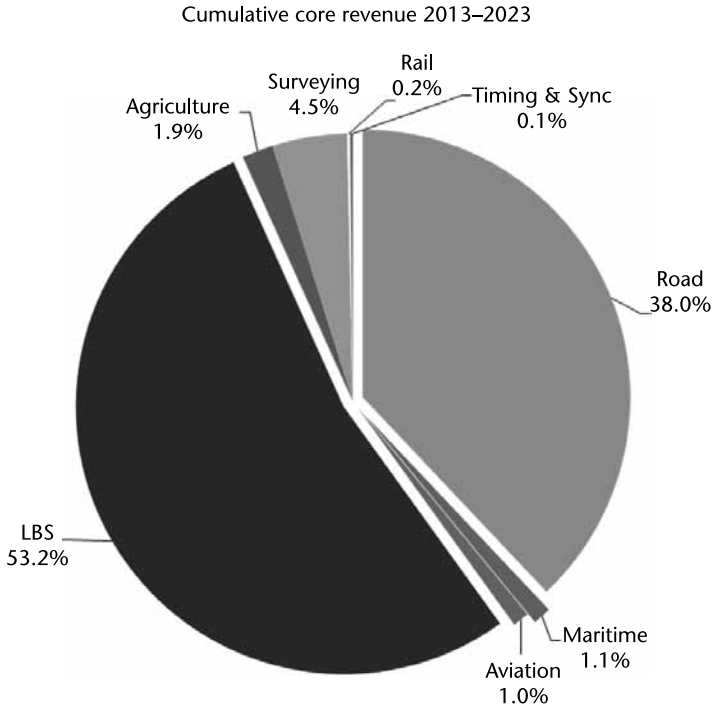

Figure 1.13 Cumulative core revenue 2013 to 2023 by market segment (billions of Euros). (Courtesy of GSA.)

(equipment) segments. Particulars of the constellation are described. Satellite types and corresponding attributes are provided including the Block IIF and GPS III. One will note the increase in the number of transmitted civil and military navigation signals as the various satellite blocks progress. Of considerable interest are interactions between the control segment (CS) and the satellites. This chapter provides a thorough understanding of the measurement processing and building of a navigation data message. A navigation data message provides the user receiver with satellite ephemerides, satellite clock corrections and other information that enable the receiver to compute PVT. An overview of user receiving equipment is presented as well as related selection criteria relevant to both civil and military users.

This chapter also describes the GPS legacy and modernized satellite signals and their generation including frequency assignments, modulation format, navigation data, received power levels, and ranging code generation.

Chapter 4 discusses the Russian GLONASS system. An overview of the system is first presented, accompanied with pertinent historical facts. The constellation and associated orbital plane characteristics are then detailed. This is followed by a description of the ground control/monitoring network and current and planned spacecraft designs. The GLONASS coordinate system, Earth model, time reference, and satellite signal characteristics are also discussed. System performance in terms of accuracy and availability is covered as well as an overview of differential services. (Chapter 12 provides details of differential services.)

Chapter 5 introduces Galileo. The overall program is first discussed followed by details of system services. Next, a detailed technical description of the system architecture is provided along with constellation particulars, satellite design, and

launch vehicle descriptions. Extensive treatment of the downlink satellite signal structure is put forth. Interoperability factors are considered next. In addition to providing navigation services, Galileo will also contribute to the international search and rescue (SAR) architecture. Details of the SAR/Galileo service are contained in Section 5.7.

Chapter 6 is dedicated to BeiDou. The chapter begins with an overview of the Beidou program, which is denoted as the BeiDou Navigation Satellite System (BDS). Program history and its three-phased evolutionary approach are described. The BDS program began with a regional RDSS and is now expanding to worldwide coverage. The chapter details constellation and satellite design particulars as well as particulars of the ground control segment. Interoperability factors (e.g., geodetic coordinate reference system, time reference system) are covered. This is followed by BDS services and an extensive treatment of satellite signal characteristics. The regional RDSS provides both navigation and messaging services.

In Chapter 7, we describe regional SATNAV systems. There is a growing realization that total dependency on one or more global core constellations for PVT services will not address unique specific regional needs. Without being closely partnered with the core constellation providers, these unique needs may not be met. Among the requirements that a regional service can provide are: guaranteed quality of service within the coverage regions (positioning and timing services to users) and unique messaging requirements for users. In Chapter 7, we discuss the NavIC, a regional service provided by India to support the region of the world centered on the continent of India and the QZSS, the regional service provided by Japan serving the western Pacific region. These constellations improve the coverage of global core constellations in mountainous territories where masking of the core constellation satellites can impact coverage in the mountain valleys and within urban canyons by assuring high-elevation angle satellite availability.

Section 7.1 describes the emerging QZSS. The QZSS program was initiated in 2002 as a government/industry effort. The first satellite was launched in 2010 and the decision to proceed for the initial operating capability came in 2012. In Section 7.1.2, the QZSS space segment is described. Although the QZSS constellation consisted of a single satellite in an inclined geosynchronous orbit at the time of this writing, the remainder of the IOC constellation were planned to be in-orbit before 2023. QZSS will transmit timing signals in the L1, L2, and L5 navigation bands (similar to the U.S. GPS).

Section 7.1.3 focuses on the QZSS control segment (CS). To ensure that the PVT requirements are met, the CS consists of satellite tracking functions (radar and laser ranging), signal monitoring stations, and timing management for the constellation. Section 7.1.4 discusses the geodesy and timing services. Of note is that QZSS plans to be closely synchronized (i.e., very small timing offset) with GPS time. In Section 7.1.5, the QZSS services to military and civil users are described and include specific augmentations for high-precision users as well as crisis and safety messaging services. Given the extremely rugged and mountainous locations in Japan, these services are considered critical for emergency uses. Finally, the specific characteristics of the six QZSS signals are discussed in Section 7.1.6.

Section 7.2, describes the NavIC. In Section 7.2.2, the space segment is discussed. After the initial decision by India to proceed to develop and deploy NavIC in 2006, the first satellite was launched in 2013. At the time of this writing, the

NavIC space segment had seven satellites in a combination of geosynchronous orbits and inclined geosynchronous orbit providing the current operational capability. The current satellites transmit positioning signals in L5 and S bands to provide both civil and military PVT services. The NavIC CS is discussed in Section 7.2.3. The function of the CS is to assure high-accuracy position and timing information and to provide special messaging services to meet the unique civil and military needs. Section 7.2.4 concentrates on the geodesy and time systems while Section 7.2.5 covers the navigation services. Section 7.2.6 covers the NavIC signals and their characteristics and Section 7.2.7 describes the user equipment for military and civil users.

Chapter 8 provides a comprehensive overview at a high level of virtually every GNSS receiver and lays the foundation for how they are designed. This chapter describes in detail every function in a GNSS receiver required to search, acquire and track the SV signals, then extract the code and carrier measurements as well as the navigation message data from the GNSS SVs. The subject matter is so extensive that rigor is often replaced with first principles as a trade-off for conveying the most important objective of this chapter seldom presented elsewhere: how a GNSS receiver is actually designed. Once these extensive design concepts are understood as a whole, the reader will have the basis for understanding or developing new innovations. Numerous references are provided for the reader seeking additional details.

Chapter 9 discusses four general classes of GNSS radio frequency (RF) signal disruptions that can deteriorate GNSS receiver performance. The first class of signal disruptions is interference (the focus of Section 9.2), which may be either unintentional or intentional (commonly referred to as jamming). Section 9.3 discusses the second class of GNSS disruptions called ionospheric scintillation, which is a signal-fading phenomenon caused by irregularities that can arise at times in the ionospheric layer of the Earth's atmosphere. The third class of disruptions is signal blockage, which is discussed in Section 9.4. Signal blockage is manifested when the line-of-sight paths of GNSS RF signals are attenuated excessively by heavy foliage, terrain, or man-made structures. The fourth and final class of GNSS disruptions, discussed in Section 9.5, is multipath. Invariably, there are reflective surfaces between each GNSS spacecraft and the user receiver that result in RF echoes arriving at the receiver after the desired (line-of-sight) signal.

GNSS measurement errors are covered in Chapter 10. A detailed explanation of each pseudorange measurement error source and its contribution to overall error budgets is provided. Spatial and time correlations characteristics are also examined. This treatment lays the groundwork for the reader to better understand DGNSS. All DGNSS systems exploit these correlations to improve overall system performance. (DGNSS system details are discussed in Chapter 12.) The chapter closes with a presentation of representative error budgets for both the single- and dual-frequency GNSS user.

Performance of standalone GNSS is discussed in Chapter 11. This chapter first provides algorithms for estimating PVT using one or more GNSS constellations. A variety of geometry factors are defined that are used in the estimation of the various components (e.g., horizontal, vertical) of the GNSS navigation solution. In Section 11.2.5, usage of additional state variables is discussed including methods to address system time offsets when using measurements from multiple GNSS

constellations. This is especially important if a receiver is tracking satellites from two or more GNSS constellations; then the difference in system times (e.g., GPS System Time, GLONASS System Time, Galileo System Time, BeiDou System Time) needs to be accounted for when blending the measurements to form the PVT solution. Sections 11.3 through 11.5 discuss, respectively, the three other important performance metrics of availability, integrity, and continuity. Each of these metrics is covered within the context of multiconstellation GNSS. It should be noted that the comprehensive treatment of integrity includes a discussion of Advanced Receiver Autonomous Integrity Monitoring (ARAIM).

There are many applications that demand higher levels of accuracy, integrity, availability, and continuity than provided by standalone GNSS. For such applications, augmentation is required. There are several classes of augmentation, which can be used singly or in combination: DGNSS, Precise Point Positioning (PPP), and the use of external sensors. Chapter 12 introduces DGNSS and PPP. Chapter 13 will discuss various external sensors/systems and their integration with GNSS.

Both DGNSS and PPP are methods to improve the positioning or timing performance of GNSS by making use of measurements from one or more reference stations at known locations, each equipped with at least one GNSS receiver. The reference station(s) provides information that is useful to improve PNT performance (accuracy, integrity, continuity, and availability) for the end user.

This chapter describes the underlying concepts of DGNSS and details a number of operational and planned DGNSS systems. The underlying algorithms and performance of code- and carrier-based DGNSS systems are presented in Sections 12.2 and 12.3, respectively. PPP systems are addressed in Section 12.4. Some important DGNSS message standards are introduced in Section 12.5. The final section, Section 12.6, details a number of operational and planned DGNSS and PPP systems.

Chapter 13 focuses on the need to provide continuous navigation between the update periods of the GNSS receiver, during periods of shading of the GNSS receiver's antenna, and through periods of interference. This is the impetus for integrating GNSS with various additional sensors. In Section 13.2, the motivations for GNSS/inertial integration are detailed. The Kalman filter is described, including an example of a typical Kalman filter implementation. Various classes of GNSS/inertial integrations are introduced and discussed. Section 13.3 addresses sensor integration for land vehicles. A description of the sensors, their integration with the Kalman filter, and test data taken during field testing of a practical multisensor system are presented. Section 13.4 discusses methods of enhancing GNSS performance using network assistance. This section includes descriptions of network assistance techniques, performance, and emerging standards. Lastly, Section 13.5 introduces the topic of extending positioning systems into indoor and other areas with GNSS signal blockage using hybrid positioning systems incorporating GNSS, low-cost inertial sensors, and various other RF signals available on mobile devices.

Chapter 14 is dedicated to GNSS markets and applications. This chapter starts with reviews of numerous market projections and continues with the process in which a company would target a specific market segment. Differences between the civil and military markets are discussed. It is of prime importance to understand these differences when targeting a specific segment of either market. The influence of governmental policy on the GNSS market is examined. Numerous civil, government, and military applications are presented.

References

[1] European GNSS Agency, *GNSS Market Report*, Issue 4, 2015.

[2] http://gpsworld.com/us-air-force-releases-gps-iii-3-launch-services-rfp/.

[3] www.gps.gov.

[4] China Satellite Navigation Office, *Development Report of BeiDou Navigation Satellite System*, (v. 2.2), December 2013, http://www.beidou.gov.cn.

[5] Ran, C., "Status Update on the BeiDou Navigation Satellite System (BDS)," *10th Meeting of the International Committee on Global Navigation Satellite Systems* (ICG), Boulder, CO, November 2–6, 2015, http://www.unoosa.org/oosa/en/ourwork/icg/meetings/icg-10/presentations.html.

[6] Fan, B., Z. Li, and T. Liu, "Application and Development Proposition of Beidou Satellite Navigation System in the Rescue of Wenchuan Earthquake [J]," *Spacecraft Engineering*, Vol. 4, 2008, pp. 6–13.

[7] The Quasi-Zenith Satellite System and IRNSS | GEOG 862, https://www.e-education.psu.edu/geog862/node/1880. Accessed January 1, 2015.

[8] Quasi-Zenith Satellite System, Presentation to ICG-9, Prague, 2014, http://www.unoosa.org/oosa/en/ourwork/icg/meetings/icg-09/presentations.html under 1140 20141109_ICG9_Presentation of QZSS_final2.pptx. Accessed January 1, 2015.

[9] Status Update on the Quasi-Zenith Satellite System Presentation to ICG-10, Boulder, CO, 2015. http://www.unoosa.org/pdf/icg/2015/icg10/06.pdf. Accessed on January 1, 2015.

[10] Indian Regional Navigational Satellite System, *Signal in Space ICD for Standard Positioning Service, Version 1*, ISRO-IRNSS-ICD-SPS-1.0, ISRO, June 2014, pp. 2–3. http://irnss.isro.gov.in.

[11] Vithiyapathy, P., "India's Strategic Guardian of the Sky," Occasional Paper 001-2015, August 25, 2015, Chennai Centre for China Studies, http://www.c3sindia.org/strategicissues/5201. Accessed January 1, 2015.

[12] "IRNSS Is Important for the India's Sovereignty," Interview of Shri Avinash Chander, Secretary Department of Defense R&D, DG R&D and Scientific Advisor to RM, Government of India, *Coordinates Magazine*, http://mycoordinates.org/"rnss-is-important-for-the-india-sovereignty.

[13] Indian Regional Navigational Satellite System, Signal in Space ICD for Standard Positioning Service, Version 1, ISRO-IRNSS-ICD-SPS-1, ISRO, June 2014, p. 5.

[14] Mateu, I., et al., "A Search for Spectrum: GNSS Signals in S-Band Part 1," *Inside GNSS Magazine*, September 2010, p. 67.

[15] Indian SATNAV Program, "Challenges and Opportunities Presentation by Dr. S. V. Kibe, Program Director, SATNAV," ISRO Headquarters, Bangalore, 1st ICG Meeting, UN OOSA, Vienna, Austria, November 1–2, 2006. www.unoosa.org/pdf/sap/2006. Slide 25. Accessed January 1, 2015.

CHAPTER 2

Fundamentals of Satellite Navigation

Elliott D. Kaplan, John W. Betz, Christopher J. Hegarty, Samuel J. Parisi, Dennis Milbert, Michael S. Pavloff, Phillip W. Ward, Joseph J. Leva, and John Burke

2.1 Concept of Ranging Using Time-of-Arrival Measurements

GNSS utilizes the concept of time-of-arrival (TOA) ranging to determine user position. This concept entails measuring the time it takes for a signal transmitted by an emitter (e.g., foghorn, ground-based radionavigation transmitter, satellite) at a known location to reach a user receiver. This time interval, referred to as the signal propagation time, is then multiplied by the speed of the signal propagation (e.g., speed of sound, speed of light) to obtain the emitter to-receiver distance. By measuring the propagation time of signals broadcast from multiple emitters (i.e., navigation aids) at known locations, the receiver can determine its position. An example of two-dimensional positioning is provided next.

2.1.1 Two-Dimensional Position Determination

Consider the case of a mariner at sea determining his or her vessel's position from a foghorn. (This introductory example was originally presented in [1] and is contained herein because it provides an excellent overview of TOA position determination concepts.) Assume that the vessel is equipped with an accurate clock and the mariner has an approximate knowledge of the vessel's position. Also, assume that the foghorn whistle is sounded precisely on the minute mark and that the vessel's clock is synchronized to the foghorn clock. The mariner notes the elapsed time from the minute mark until the foghorn whistle is heard. The foghorn whistle propagation time is the time it took for the foghorn whistle to leave the foghorn and travel to the mariner's ear. This propagation time multiplied by the speed of sound (approximately 335 m/s) is the distance from the foghorn to the mariner. If the foghorn signal took 5 seconds to reach the mariner's ear, then the distance to the foghorn is 1,675m. Let this distance be denoted as R1. Thus, with only one measurement, the mariner knows that the vessel is somewhere on a circle with radius R1 centered about the foghorn, which is denoted as Foghorn 1 in Figure 2.1.

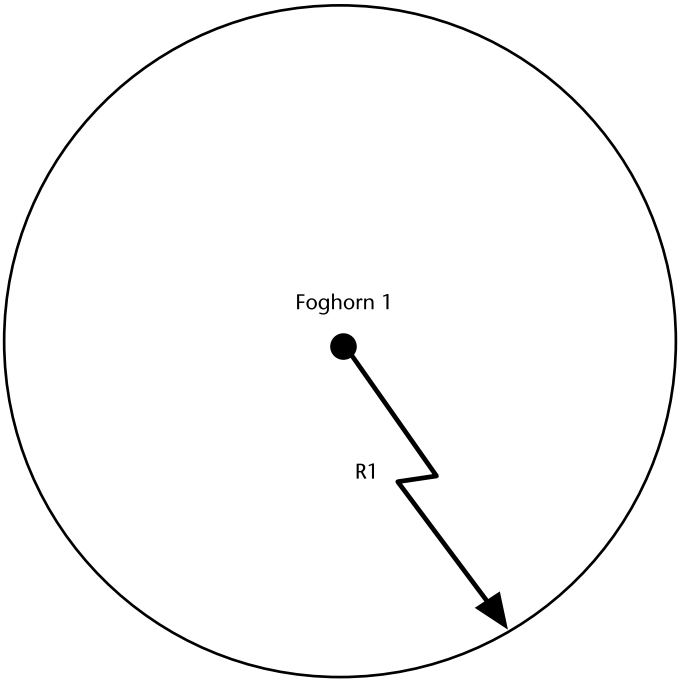

Figure 2.1 Range determination from a single source. (After: [1].)

Hypothetically, if the mariner simultaneously measured the range from a second foghorn in the same way, the vessel would be at range $R1$ from Foghorn 1 and range $R2$ from Foghorn 2, as shown in Figure 2.2. It is assumed that the foghorn transmissions are synchronized to a common time base and the mariner has knowledge of both foghorn whistle transmission times. Therefore, the vessel relative to the foghorns is at one of the intersections of the range circles. Since it was assumed that the mariner has approximate knowledge of the vessel's position, the unlikely fix can be discarded. Resolving the ambiguity can also be achieved by making a range measurement to a third foghorn, as shown in Figure 2.3.

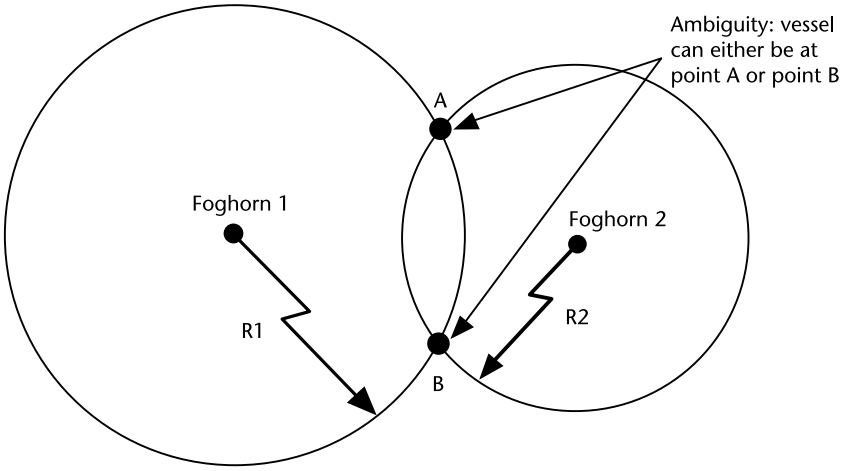

Figure 2.2 Ambiguity resulting from measurements to two sources. (After: [1].)

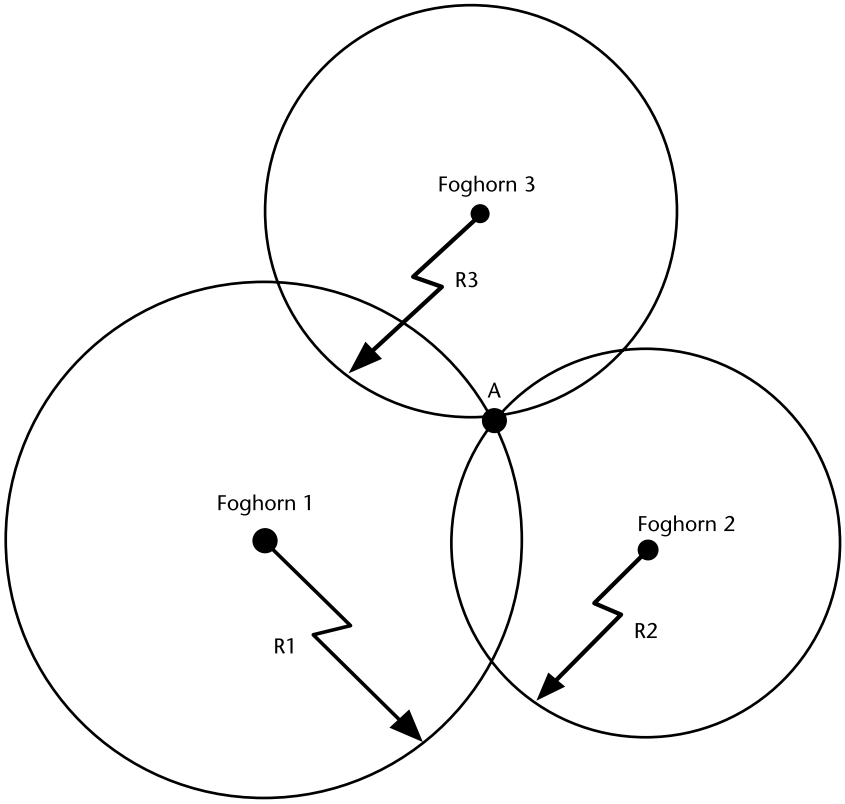

Figure 2.3 Position ambiguity removal by additional measurement. (After: [1].)

2.1.1.1 Common Clock Offset and Compensation

The above development assumed that the vessel's clock was precisely synchronized with the foghorn time base. However, this might not be the case. Let us presume that the vessel's clock is advanced with respect to the foghorn time base by 1 second. That is, the vessel's clock believes the minute mark is occurring 1 second earlier. The propagation intervals measured by the mariner will be larger by 1 second due to the offset. The timing offsets are the same for each measurement (i.e., the offsets are common) because the same incorrect time base is being used for each measurement. The timing offset equates to a range error of 335m and is denoted as ε in Figure 2.4. The separation of intersections C, D, and E from the true vessel position, A, is a function of the vessel's clock offset. If the offset could be removed or compensated for, the range circles would then intersect at point A.

2.1.1.2 Effect of Independent Measurement Errors on Position Certainty

If this hypothetical scenario were realized, the TOA measurements would not be perfect due to errors from atmospheric effects, foghorn clock offset from the foghorn time base, and interfering sounds. Unlike the vessel's clock offset condition cited above, these errors would be generally independent and not common to all measurements. They would affect each measurement in a unique manner and result in inaccurate distance computations. Figure 2.5 shows the effect of independent

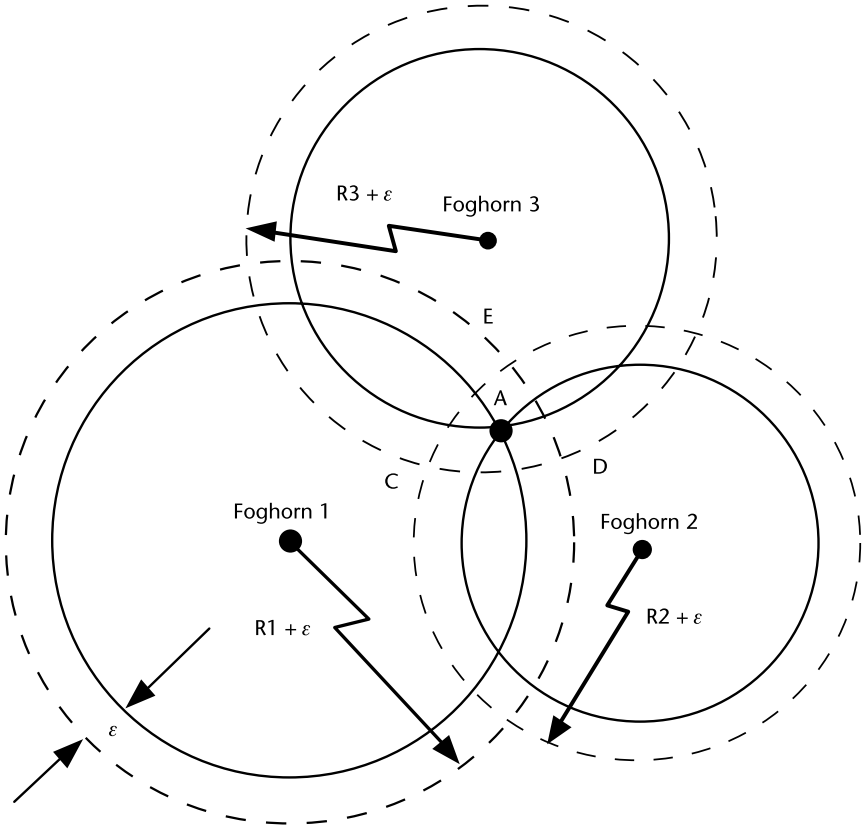

Figure 2.4 Effect of receiver clock offset on TOA measurements. (After: [1].)

errors (i.e., ε_1, ε_2, and ε_3) on position determination assuming foghorn time base/mariner clock synchronization. Instead of the three range circles intersecting at a single point, the vessel location is somewhere within the triangular error space.

2.1.2 Principle of Position Determination via Satellite-Generated Ranging Codes

GNSS employs TOA ranging for user position determination. By making TOA measurements to multiple satellites, three-dimensional positioning is achieved. We will observe that this technique is analogous to the preceding foghorn example; however, satellite ranging codes travel at the speed of light, which is approximately 3×10^8 m/s. It is assumed that the satellite ephemerides are accurate (i.e., the satellite locations are precisely known).

2.1.2.1 Three-Dimensional Position Location Via Intersection of Multiple Spheres

Assume that there is a single satellite transmitting a ranging signal. A clock onboard the satellite controls the timing of the ranging signal broadcast. This clock and others onboard each of the satellites within a particular SATNAV constellation are effectively synchronized to an internal system time scale herein referred to as system time (e.g., GPS system time). The user's receiver also contains a clock that (for

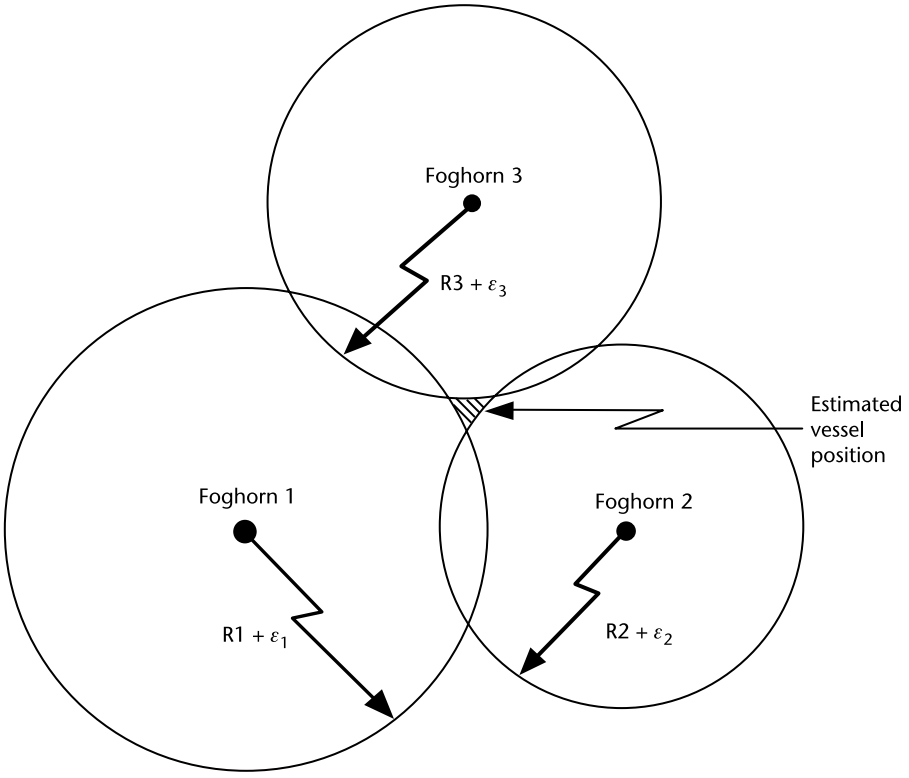

Figure 2.5 Effect of independent measurement errors on position certainty.

the moment) we assume to be synchronized to system time. Timing information is embedded within the satellite ranging signal that enables the receiver to calculate when the signal left the satellite based on the satellite clock time. This is discussed in more detail in Section 2.5.1. By noting the time when the signal was received, the satellite-to-user propagation time can be computed. The product of the satellite-to-user propagation time and the speed of light yields the satellite-to-user range, R. As a result of this measurement process, the user would be located somewhere on the surface of a sphere centered about the satellite as shown in Figure 2.6(a). If a measurement was simultaneously made using the ranging code of a second satellite, the user would also be located on the surface of a second sphere that is concentric about the second satellite. Thus, the user would then be somewhere on the surface of both spheres, which could be either on the perimeter of the shaded circle in Figure 2.6(b) that denotes the plane of intersection of these spheres or at a single point tangent to both spheres (i.e., where the spheres just touch). This latter case could only occur if the user was collinear with the satellites, which is not the typical case. The plane of intersection is perpendicular to a line connecting the satellites, as shown in Figure 2.6(c).

Repeating the measurement process using a third satellite, the user is at the intersection of the perimeter of the circle and the surface of the third sphere. This third sphere intersects the shaded circle perimeter at two points; however, only one of the points is the correct user position, as shown in Figure 2.6(d). A view of the intersection is shown in Figure 2.6(e). It can be observed that the candidate

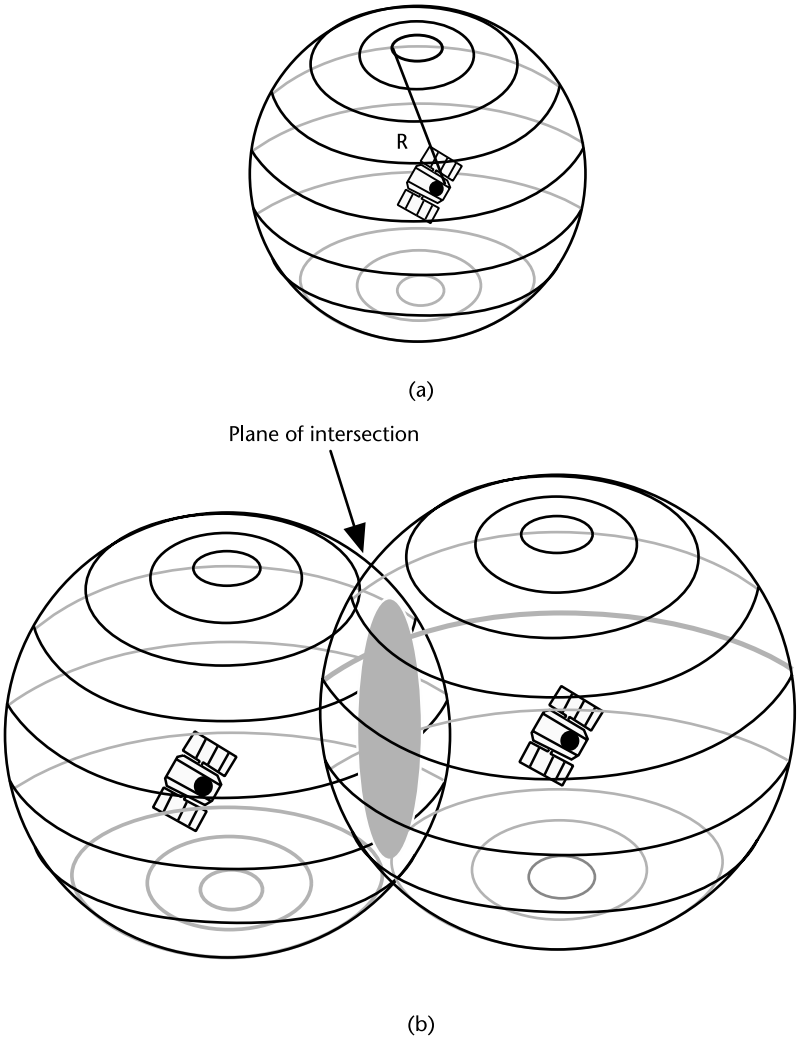

Figure 2.6 (a) User located on surface of sphere; (b) user located on perimeter of shaded circle (source: [2], reprinted with permission); (c) plane of intersection; (d) user located at one of two points on shaded circle (source: [2], reprinted with permission); and (e) user located at one of two points on circle perimeter.

locations are mirror images of one another with respect to the plane of the satellites. For a user on the Earth's surface, it is apparent that the lower point will be the true position. However, users that are above the Earth's surface may employ measurements from satellites at negative elevation angles. This complicates the determination of an unambiguous solution. Airborne/spaceborne receiver solutions may be above or below the plane containing the satellites, and it may not be clear which point to select unless the user has ancillary information.

2.2 Reference Coordinate Systems

To formulate the mathematics of the satellite navigation problem, it is necessary to choose a reference coordinate system in which the states of both the satellite and

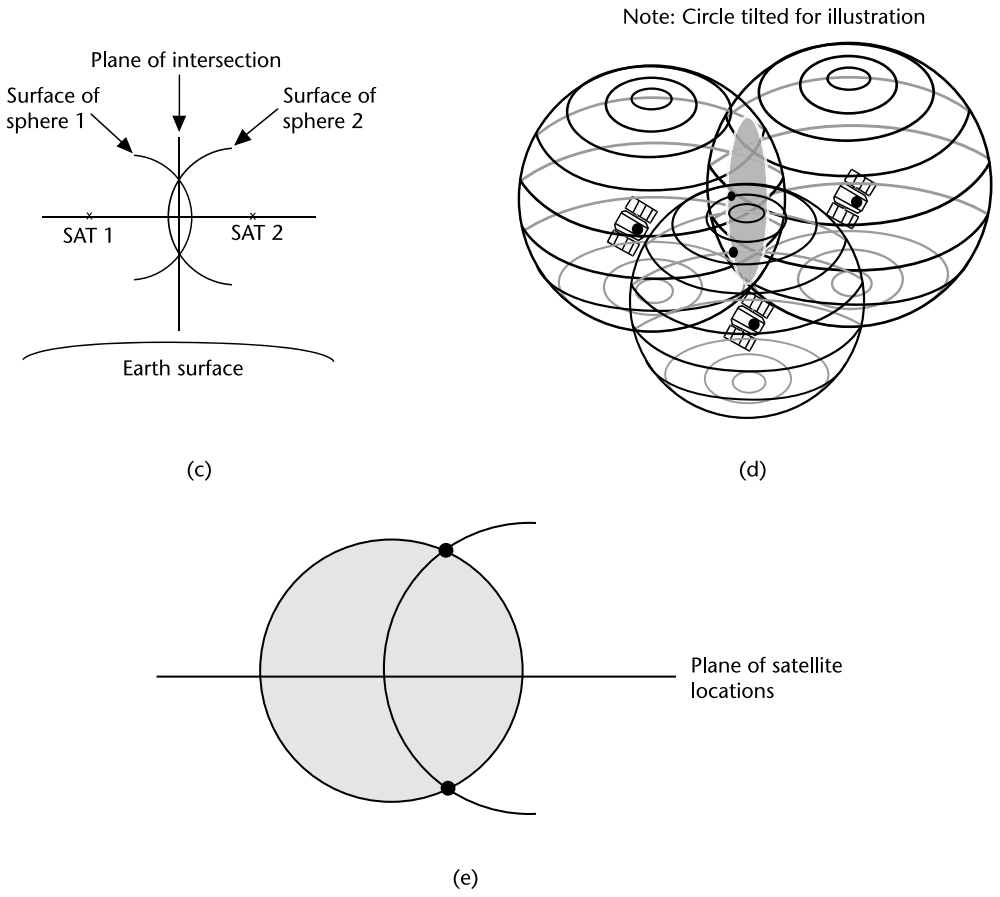

Figure 2.6 (continued)

the receiver can be represented. In this formulation, it is typical to describe satellite and receiver states in terms of position and velocity vectors measured in a Cartesian coordinate system. The Cartesian coordinate systems can be categorized as inertial and rotating systems, and as Earth-centered and local (topocentric) systems. In this section, an overview is provided of the coordinate systems used in conjunction with GNSS.

2.2.1 Earth-Centered Inertial (ECI) Coordinate System

For the purposes of measuring and determining the orbits of satellites, it is convenient to use an Earth-centered inertial (ECI) coordinate system, in which the origin is at the center of mass of the Earth and whose axes are pointing in fixed directions with respect to the stars. A satellite's position and velocity may be modeled with Newton's laws of motion and gravitation in an ECI coordinate system. In typical ECI coordinate systems, the xy-plane is taken to coincide with the Earth's equatorial plane, the +x-axis is permanently fixed in a particular direction relative to the celestial sphere, the +z-axis is taken normal to the xy-plane in the direction of the North Pole, and the +y-axis is chosen so as to form a right-handed coordinate

system. Determination and subsequent prediction of satellite orbits are carried out in an ECI coordinate system.

There is an inherent problem in defining an ECI system in terms of the Earth's equatorial plane. The Earth is subject to motions of precession, nutation, and polar motion. The Earth's shape is oblate, and due largely to the gravitational pull of the Sun and the Moon on the Earth's equatorial bulge, the equatorial plane moves with respect to the celestial sphere. Because the z-axis is defined relative to the equatorial plane, the Earth's motions would cause the ECI system as defined above to have an orientation which changes in time. The solution to this problem is to define the orientation of the axes at a particular instant in time or epoch.

It is customary to define an ECI coordinate system with the orientation of the equatorial plane at 1200 hr TT on January 1, 2000, denoted as the J2000 system. The $+x$-axis is taken to point from the center of mass of the Earth to the direction of vernal equinox, and the y- and z-axes are defined above, all at the aforementioned epoch. Terrestrial time (TT) is a uniform time system representing an idealized clock on the Earth's geoid. TT has replaced the old Ephemeris Time (ET), and TT is ahead of International Atomic Time (TAI) by 32.184 seconds.

2.2.2 Earth-Centered Earth-Fixed (ECEF) Coordinate System

For the purpose of computing the position of a GNSS receiver, it is more convenient to use a coordinate system that rotates with the Earth, known as an Earth-centered Earth-fixed (ECEF) system. In such a coordinate system, it is easier to compute the latitude, longitude, and height. The ECEF coordinate system used for GNSS has its xy-plane coincident with the Earth's equatorial plane. In the ECEF system, the +x-axis points in the direction of 0° longitude and the +y-axis points in the direction of 90° East longitude. The x- and y-axes rotate with the Earth and no longer describe fixed directions in inertial space. The +z-axis is chosen to be normal to the instantaneous equatorial plane in the direction of the geographical North Pole (i.e., where the lines of longitude meet in the northern hemisphere), forming a right-handed coordinate system. The z-axis will trace a path across the celestial sphere due to the Earth's precession, nutation, and polar motion.

Agencies that perform precision GNSS orbit computation include the transformations between the ECI and the ECEF coordinate systems to very high degrees of accuracy. Such transformations are accomplished by the application of rotation matrices to the satellite position and velocity vectors originally computed in the ECI coordinate system, as described below. By contrast, broadcast orbit computations (see [3] for a GPS example) typically generate satellite position and velocity directly in the ECEF frame. Precise orbits from numerous computation centers also express satellite position and velocity in ECEF. The Earth motions of precession, nutation, UT1 difference, and polar motion are small for a short time interval (e.g., interval of a navigation message). Thus, with one provision, we may usually proceed to formulate a GNSS navigation problem in the ECEF system without discussing the details of the orbit determination or the transformation to the ECEF system.

The exception is the average rotation of the Earth. Earth rotation is not negligible for the signal transit interval from satellite to Earth surface. When formulating signal propagation in a rotating, noninertial, ECEF system, a correction is needed. This is known as the Sagnac effect and is further described in Section

10.2.3. Alternatively, one must compute geometric range from the ECI coordinates for satellite and receiver.

As a result of the navigation computation process, the Cartesian coordinates (x_u, y_u, z_u) of the user's receiver are computed in the ECEF system, as described in Section 2.5.2. It is common to transform these Cartesian coordinates to latitude, longitude, and height of the receiver, as detailed in Section 2.2.5.

2.2.2.1 Rotation Matrices

It is useful to consider a coordinate set or a vector $\mathbf{u} = (x_u, y_u, z_u)$ not only in an ECEF system, but also transformed into an arbitrary system, including the ECI system. Such a vector transformation can be computed by multiplication with the rotation matrices [3–5]:

$$\mathbf{R}_1(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix} \quad \mathbf{R}_2(\theta) = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \quad \mathbf{R}_3(\theta) = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Here, $\mathbf{R}_1(\theta)$, $\mathbf{R}_2(\theta)$, and $\mathbf{R}_3(\theta)$, denote rotation by an angle, θ, about the x, y, and z axes, respectively. A positive θ denotes a counterclockwise rotation of the respective axis when the origin is viewed from the positive end of that axis. An example of an $\mathbf{R}_1(\theta)$ rotation is portrayed in Figure 2.7.

An arbitrary rotation, $\mathbf{R}$, is constructed by successive application of elementary axial rotations. Multiplication by the rotation matrices will not change the handedness of the new coordinate system. Rotation matrices and their products are orthogonal, $\mathbf{R}^{-1}(\alpha) = \mathbf{R}^{t}(\alpha)$. Due to the contents of a rotation matrix, $\mathbf{R}^{-1}(\alpha) = \mathbf{R}(-\alpha)$. So, for example, if $\mathbf{R} = \mathbf{R}_1(\alpha)\,\mathbf{R}_2(\beta)$, then

$$\mathbf{R}^{-1} = \left(\mathbf{R}_1(\alpha)\mathbf{R}_2(\beta)\right)^{-1} = \left(\mathbf{R}_2^{-1}(\beta)\mathbf{R}_1^{-1}(\alpha)\right) = \left(\mathbf{R}_2^{t}(\beta)\mathbf{R}_1^{t}(\alpha)\right) = \left(\mathbf{R}_2(-\beta)\mathbf{R}_1(-\alpha)\right)$$

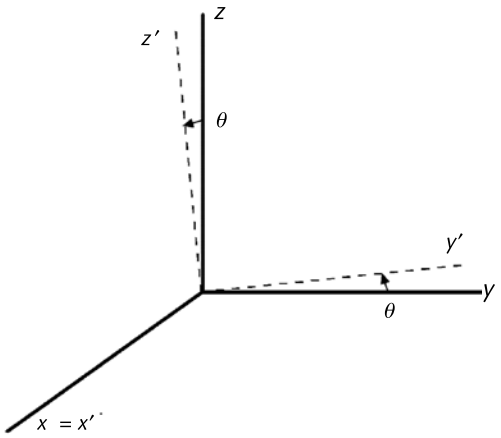

Figure 2.7 Example axial rotation, $\mathbf{R}_1(\theta)$ (x axis, positive θ).

2.2.2.2 Transformation Between ECEF and ECI

Applications seldom require access to the base ECI coordinate system or the complete ECEF-ECI transformation. It is sufficient to merely sketch the transformation. Following [5],

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_{ECEF} = \mathbf{R}_M \mathbf{R}_S \mathbf{R}_N \mathbf{R}_P \begin{bmatrix} x \\ y \\ z \end{bmatrix}_{ECI}$$

where the composite rotation transformation matrices are:

Precession $\mathbf{R}_P = \mathbf{R}_3(-Z)\, \mathbf{R}_2(\theta)\, \mathbf{R}_3(-\zeta)$
Nutation $\mathbf{R}_N = \mathbf{R}_1(\varepsilon - \Delta\varepsilon)\, \mathbf{R}_3(-\Delta\psi)\, \mathbf{R}_1(\varepsilon)$
Earth Rotation $\mathbf{R}_S = \mathbf{R}_3(\text{GAST})$
Polar Motion $\mathbf{R}_M = \mathbf{R}_2(-y_p)\, \mathbf{R}_1(-x_p)$

and where the precession parameters (Z, θ, ζ) and the nutation parameters (ε, $\Delta\varepsilon$, $\Delta\psi$) are computed by power series [5]. GAST symbolizes Greenwich Apparent Sidereal Time, which is computed from a few elements that include the UT1-UTC difference, ΔUT1. The x-axis and y-axis polar motion is x_p and y_p, respectively. Note that the precession and nutation parameters are documented as part of J2000 and are functions of time. However, the Earth orientation components (ΔUT1, x_p, y_p) vary with time and are not accurately predictable. Various agencies monitor Earth orientation components and provide them to the public. Some GNSS navigation messages transmit the Earth orientation components.

Since rotation matrices are orthogonal, we may immediately write the ECEF-to-ECI transformation as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_{ECI} = \mathbf{R}_P^t \mathbf{R}_N^t \mathbf{R}_S^t \mathbf{R}_M^t \begin{bmatrix} x \\ y \\ z \end{bmatrix}_{ECEF}$$

2.2.3 Local Tangent Plane (Local Level) Coordinate Systems

Local tangent plane systems form a useful category of coordinate systems. Refer to Figure 2.8, which displays both ECEF and local tangent systems.

Local tangent systems have their origin, P, at or near the Earth's surface, Q, and have a horizontal plane (the *en*-plane) approximately coincident with local level. Thus, they easily model the experience of an observer. The vertical axis may be aligned with the geocentric radius vector, aligned with the ellipsoidal normal, **u** (portrayed in Figure 2.8), or aligned with the local gravity vector. Without loss

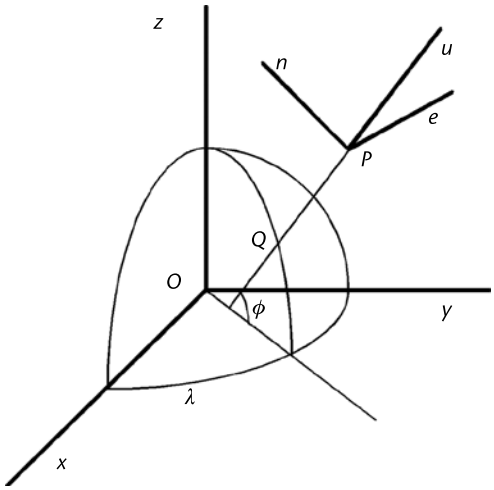

Figure 2.8 Relationships of ECEF and local tangent plane coordinate systems.

of generality, we focus on the ellipsoidal tangent plane system, portrayed in Figure 2.8.

The principal alignments are the vertical (up-down) along the ellipsoidal normal, the North-South axis tangent to the geodetic meridian expressed in an Earth-fixed realization, and the East-West axis perpendicular to these other two axes. In practice, a variety of local ellipsoidal tangent systems are defined. They vary with the choices between Up-Down, North-South, and East-West, and with the ordering of axes to express coordinates. Both right-hand and left-hand tangent systems are found in use.

As an illustrative example, consider the ENU (East-North-Up) ellipsoidal tangent plane system. This is a right-handed system. Let the origin of the ENU system, (x_o, y_o, z_o) at point P, have geodetic latitude and longitude (φ, λ). Latitude is reckoned positive North, and longitude is positive East. Denote the local level coordinate system components with (e, n, u). The Cartesian ECEF system can be brought into the tangent plane system by a translation and a combined rotation. The translation is obtained by subtraction of the local level origin, (x_o, y_o, z_o). The combined rotation is a rotation of $\pi/2 + \lambda$ about the z axis, followed by rotation of $\pi/2 - \varphi$ about the new x-axis. This is expressed formally through rotation matrices and through their explicit product:

$$\begin{bmatrix} e \\ n \\ u \end{bmatrix} = \mathbf{R}_1\left(\frac{\pi}{2} - \varphi\right)\mathbf{R}_3\left(\frac{\pi}{2} + \lambda\right)\begin{bmatrix} x - x_o \\ y - y_o \\ z - z_o \end{bmatrix} = \begin{bmatrix} -\sin\lambda & \cos\lambda & 0 \\ -\sin\varphi\cos\lambda & -\sin\varphi\sin\lambda & \cos\varphi \\ \cos\varphi\cos\lambda & \cos\varphi\sin\lambda & \sin\varphi \end{bmatrix}\begin{bmatrix} x - x_o \\ y - y_o \\ z - z_o \end{bmatrix}$$

Note that the matrix multiplications do not commute. They are applied right to left in the specified order. Rotation matrices and their products are orthogonal. Hence, the inverse transformation is merely the transpose of the explicit product.

Now, as a second example, consider the left-handed system, NEU (North-East-Up), with ellipsoidal tangent plane coordinates (u, v, w). Exchange of any two axes

of a three-dimensional Cartesian system will reverse the handedness of the system. Thus, exchange of the East and North axes will convert the right-handed ENU system into the left-handed NEU system. The explicit transformation is immediately obtained by row exchange:

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} -\sin\varphi\cos\lambda & -\sin\varphi\sin\lambda & \cos\varphi \\ -\sin\lambda & \cos\lambda & 0 \\ \cos\varphi\cos\lambda & \cos\varphi\sin\lambda & \sin\varphi \end{bmatrix} \begin{bmatrix} x - x_o \\ y - y_o \\ z - z_o \end{bmatrix}$$

This section is closed with a sample application of the NEU system. The geocentric vectors in an ECEF system to an observer, $\mathbf{u}_o$, and a satellite, $\mathbf{u}_s$, may be differenced to obtain a relative, observer-to-satellite vector, $\mathbf{u} = (x, y, z)$. The matrix expression above will immediately convert the observer-satellite vector into the local ellipsoidal tangent NEU system. One may then write simple expressions for azimuth, α, and vertical angle, σ, as:

$$\alpha = \tan^{-1}\left(\frac{v}{u}\right) \quad \sigma = \tan^{-1}\left(\frac{w}{\sqrt{u^2 + v^2}}\right)$$

where azimuth is reckoned clockwise from the North, and vertical angle is positive upwards. These would be the look angles from an observer to a satellite.

2.2.4 Local Body Frame Coordinate Systems

Coordinate systems affixed to vehicles or objects are needed for numerous applications. They may be used to establish object attitude, orientation of a sensor package, modeling of effects such as atmospheric drag, or fusion of on-board systems, such as inertial and GNSS.

As with the local tangent plane systems, a variety of local body frame systems have been defined. The origin may be the center of mass of a vehicle, although that is not a strict requirement. The body frame coordinate axes can correspond to the principal axes of the vehicle. However, once again, variations occur in how the body frame axes are associated with a vehicle's axes of symmetry.

Following the example of [6], a right-hand coordinate system is constructed. The positive y' axis points along the nose of the vehicle. The positive z' axis points through the top of the vehicle. The third axis, the x' axis, extends to the right of the vehicle. This arrangement is displayed in Figure 2.9.

The transformation of coordinates from a vehicle-centered ENU tangent plane system into the local body frame system is obtained by a combined rotation formed from three elementary axial rotations that lead from the ENU system into the desired local body frame.

The transformation is visualized most easily by imagining a starting vehicle as level and aligned to the North in the ENU system. The first rotation is around the

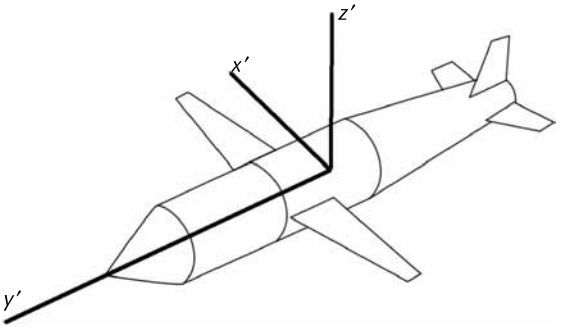

Figure 2.9 Example local body frame coordinate system.

z' axis, and is called yaw, y. In this starting condition, the z' axis equals the e axis. The second rotation is around the new x' axis, and is called pitch, p. The final rotation is around an even newer y' axis, and is called roll, r. (The use of the symbol y for yaw is for mnemonic reasons, and should not be confused with an ECF or ECEF y axis.)

The combined rotation from the ENU ellipsoidal tangent plane system into the local body frame coordinate system is obtained by multiplication of the elementary rotation matrices:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \mathbf{R}_2(r)\mathbf{R}_1(p)\mathbf{R}_3(y)\begin{bmatrix} e \\ n \\ u \end{bmatrix}$$

The coordinate transformation is written explicitly as:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos r\cos y - \sin r\sin p\sin y & \cos r\sin y + \sin r\sin p\cos y & -\sin r\cos p \\ -\cos p\sin y & \cos p\cos y & \sin p \\ \sin r\cos y + \cos r\sin p\sin y & \sin r\sin y - \cos r\sin p\cos y & \cos r\cos p \end{bmatrix}\begin{bmatrix} e \\ n \\ u \end{bmatrix}$$

As before, the rotation matrices and their products are orthogonal. The inverse transformation is merely the transpose of the explicit product.

2.2.5 Geodetic (Ellipsoidal) Coordinates

We are concerned here with estimating the latitude, longitude, and height of a GNSS receiver. This is accomplished with an ellipsoidal model of the Earth's shape, as shown in Figure 2.10. In this model, cross-sections of the Earth parallel to the equatorial plane are circular. The cross-sections of the Earth normal to the equatorial plane are ellipsoidal. The ellipsoidal cross-section has a semimajor axis length, a, and a semiminor axis length, b. The eccentricity of the Earth ellipsoid, e, can be determined by

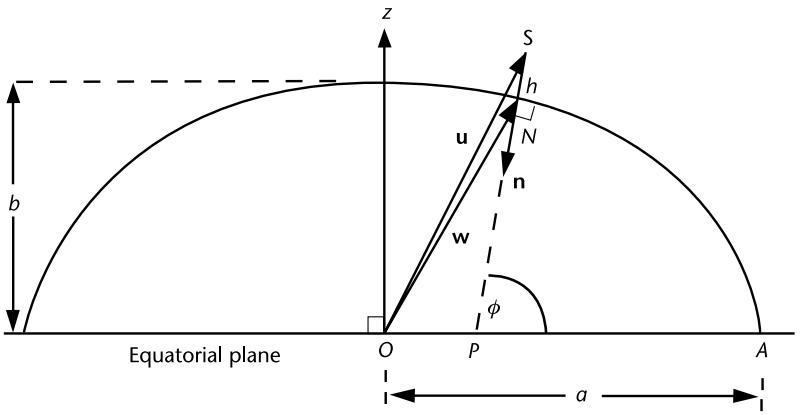

Figure 2.10 Ellipsoidal model of Earth (cross-section normal to equatorial plane).

$$e = \sqrt{1 - \frac{b^2}{a^2}}$$

Another parameter sometimes used to characterize the reference ellipsoid is the second eccentricity, e', which is defined as follows:

$$e' = \sqrt{\frac{a^2}{b^2} - 1} = \frac{a}{b} e$$

2.2.5.1 Determination of User Geodetic Coordinates: Latitude, Longitude, and Height

The ECEF coordinate system is affixed to the reference ellipsoid, as shown in Figure 2.10, with the point O corresponding to the center of the Earth. We can now define the parameters of latitude, longitude, and height with respect to the reference ellipsoid. When defined in this manner, these parameters are called *geodetic*. Given a user receiver's position vector of $\mathbf{u} = (x_u, y_u, z_u)$ in the ECEF system, we can compute the geodetic longitude, λ, as the angle between the user and the x-axis, measured in the xy-plane

$$\lambda = \begin{cases} \arctan\left(\dfrac{y_u}{x_u}\right), & x_u \geq 0 \\ 180° + \arctan\left(\dfrac{y_u}{x_u}\right), & x_u < 0 \text{ and } y_u \geq 0 \\ -180° + \arctan\left(\dfrac{y_u}{x_u}\right), & x_u < 0 \text{ and } y_u < 0 \end{cases} \tag{2.1}$$

In (2.1), negative angles correspond to degrees West longitude. The geodetic parameters of latitude, ϕ, and height, h, are defined in terms of the ellipsoid normal at the user's receiver. The ellipsoid normal is depicted by the unit vector **n** in Figure 2.10. Notice that unless the user is on the poles or the equator, the ellipsoid normal does not point exactly toward the center of the Earth. A GNSS receiver computes height relative to the reference ellipsoid. However, the height above sea level given on a map can be quite different from GNSS-derived height due to the difference between the reference ellipsoid and the geoid (local mean sea level). In the horizontal plane, differences between a local datum [e.g., North American Datum 1983 (NAD 83) and European Datum 1950 (ED 50)], and GNSS-based horizontal position can also be significant.

Geodetic height, h, is simply the minimum distance between the user S (at the endpoint of the vector **u**) and the reference ellipsoid. Notice that the direction of minimum distance from the user to the surface of the reference ellipsoid will be in the direction of the vector **n**. Notice, also, that S may be below the surface of the ellipsoid, and that the ellipsoidal height, h, will be negative in those cases.

Geodetic latitude, ϕ, is the angle between the ellipsoid normal vector **n** and the projection of **n** into the equatorial xy-plane. Conventionally, ϕ is taken to be positive if $z_u > 0$ (i.e., if the user is in the northern hemisphere) and ϕ is taken to be negative if $z_u < 0$. With respect to Figure 2.10, geodetic latitude is the angle NPA. N is the closest point on the reference ellipsoid to the user. P is the point where a line in the direction of **n** intersects the equatorial plane. Numerous solutions, both closed-form and iterative, have been devised for the computation of geodetic curvilinear coordinates (ϕ, λ, h) from Cartesian coordinates (x, y, z). A popular and highly convergent iterative method by Bowring [7] is described in Table 2.1. For the computations shown in Table 2.1, a, b, e^2, and e'^2 are the geodetic quantities described previously. Note that the use of N in Table 2.1 follows Bowring [7] and does not refer to geoid height described in Section 2.2.6.

2.2.5.2 Conversion from Geodetic Coordinates to Cartesian Coordinates in an ECEF System

For completeness, equations for transforming from geodetic coordinates back to Cartesian coordinates in the ECEF system are provided next. Given the geodetic parameters λ, ϕ, and h, we can compute $\mathbf{u} = (x_u, y_u, z_u)$ in a closed form as follows:

$$\mathbf{u} = \begin{bmatrix} \dfrac{a\cos\lambda}{\sqrt{1+\left(1-e^2\right)\tan^2\phi}} + h\cos\lambda\cos\phi \\ \dfrac{a\sin\lambda}{\sqrt{1+\left(1-e^2\right)\tan^2\phi}} + h\sin\lambda\cos\phi \\ \dfrac{a\left(1-e^2\sin\phi\right)}{\sqrt{1-e^2\sin^2\phi}} + h\sin\phi \end{bmatrix}$$

Table 2.1 Determination of Geodetic Height and Latitude in Terms of ECEF Parameters

$$p = \sqrt{x^2 + y^2}$$

$$\tan u = \left(\frac{z}{p}\right)\left(\frac{a}{b}\right)$$

Iteration Loop

$$\cos^2 u = \frac{1}{1 + \tan^2 u}$$

$$\sin^2 u = 1 - \cos^2 u$$

$$\tan\phi = \frac{z + e'^2 b \sin^3 u}{p - e^2 a \cos^3 u}$$

$$\tan u = \left(\frac{b}{a}\right)\tan\phi$$

until tan u converges, then

$$N = \frac{a}{\sqrt{1 - e^2 \sin^2 \phi}}$$

$$h = \frac{p}{\cos\phi} - N \quad \phi \neq \pm 90°$$

otherwise

$$h = \frac{z}{\sin\phi} - N + e^2 N \quad \phi \neq 0$$

2.2.6 Height Coordinates and the Geoid

The ellipsoid height, h, is the height of a point, P, above the surface of the ellipsoid, E, as described in Section 2.2.5.1. This corresponds to the directed line segment EP in Figure 2.11, where positive sign denotes a point P further from the center of the Earth than point E. Note that P need not be on the surface of the Earth, but could also be above or below the Earth's surface. As discussed in the previous sections, ellipsoid height is easily computed from Cartesian ECEF coordinates.

Historically, heights have not been measured relative to the ellipsoid but, instead, relative to a surface called the geoid. The geoid is that surface of constant geopotential, $W = W_0$, which corresponds to global mean sea level in a least squares sense. Heights measured relative to the geoid are called orthometric heights, or, less formally, heights above the mean sea level. Orthometric heights are important, because these are the types of height found on innumerable topographic maps and in paper and digital data sets.

The geoid height, N, is the height of a point, G, above the ellipsoid, E. This corresponds to the directed line segment EG in Figure 2.11, where positive sign denotes point G further from the center of the Earth than point E. The orthometric height, H, is the height of a point P, above the geoid, G. Hence, we can immediately write the equation

$$h = H + N \tag{2.2}$$

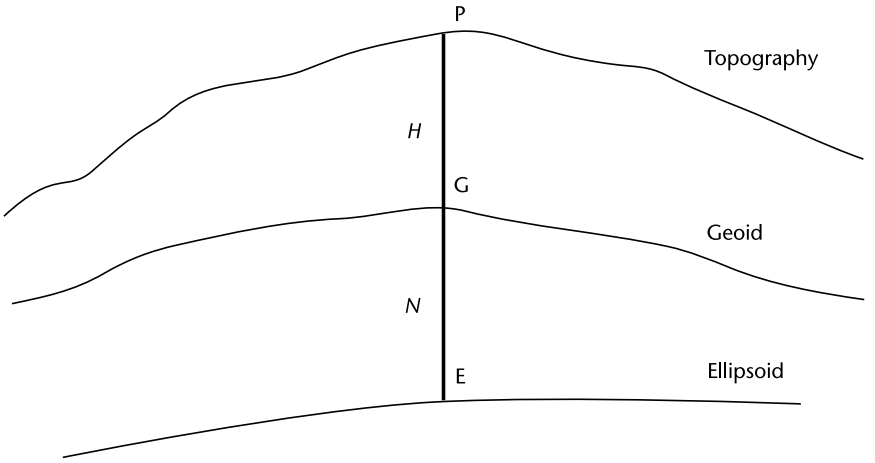

Figure 2.11 Relationships between topography, geoid, and ellipsoid.

Note that Figure 2.11 is illustrative, and that G and/or P may be below point E. Similarly, any or all terms of (2.2) may be positive or negative. For example, in the conterminous United States, the geoid height, *N*, is negative.

The geoid is a complex surface, with undulations that reflect topographic, bathymetric (i.e., measurements derived from bodies of water), and geologic density variations of the Earth. The magnitude of geoid height can be several tens of meters. Geoid height ranges from a low of about –105m at the southern tip of India, to a high of about +85m at New Guinea. Thus, for many applications, the geoid is not a negligible quantity, and one must avoid mistaking an orthometric height for an ellipsoidal height.

In contrast to the ellipsoid, the geoid is a natural feature of the Earth. Like topography, there is no simple equation to describe the spatial variation of geoid height. Geoid height is modeled and tabulated by several geodetic agencies. Global geoid height models are represented by sets of spherical harmonic coefficients, and, also, by regular grids of geoid height values. Regional geoid height models can span large areas, such as the entire conterminous United States, and are invariably expressed as regular grids. Recent global models contain harmonic coefficients to degree and order 2190. As such, their resolution is 5 arc-minutes, and their accuracy is limited by truncation error. Regional models, by contrast, are computed to a much higher resolution. One arc-minute resolution is not uncommon, and truncation error is seldom encountered.

The best-known global geoid model is the NGA (National Geospatial-Intelligence Agency) EGM2008 - WGS 84 version Geopotential Model, hereafter referred to as EGM2008 [8]. This product is a set of coefficients to degree and order 2190 and a companion set of correction coefficients needed to compute geoid height over land. EGM2008 replaces EGM96, complete to degree and order 360, and WGS 84 (180,180), complete up to degree and order 180. Most of that latter WGS 84 coefficient set was originally classified in 1985, and only coefficients through degree and order 18 were released. Hence, the first public distributions of WGS 84 geoid height only had a 10 arc-degree resolution and suffered many meters of truncation

error. Therefore, historical references to WGS 84 geoid values must be used with caution.

Within the conterminous United States, the current high-resolution geoid height grid is GEOID12B, developed by the National Geodetic Survey, NOAA. This product is a grid of geoid heights, at 1 arc-minute resolution, and has an accuracy of 2–4 cm, one sigma. Work is underway on a series of test models (e.g. xGEOID14B) that span a region of 80° of latitude and 180° of longitude. It is anticipated this new geoid model will be declared operational in 2022.

When height accuracy requirements approach the meter level, then one must also become aware of the datum differences between height coordinates. For example, the origin of the NAD 83 reference frame is offset about 2.2m from the center of the Earth, causing about 0.5–1.5m differences in ellipsoidal heights. Estimates place the origin of the U.S. orthometric height datum, NAVD 88, about 30 to 50 cm below the EGM 96 reference geopotential surface. Because of these two datum offsets, GEOID12B was constructed to accommodate these origin differences, and directly convert between NAD 83 and NAVD 88, rather than express a region of an idealized global geoid. In addition, offsets of one half meter or more in national height data are common, as tabulated in [9]. For these reasons, (2.2) is valid as a conceptual model, but may be problematic in actual precision applications. Detailed treatment of height systems is beyond the scope of this text. However, more information may be found in [10, 11].

2.2.7 International Terrestrial Reference Frame (ITRF)

The foregoing material outlines the theory of reference systems applicable to GNSS. Following the nomenclature of the International Earth Rotation and Reference Systems Service (IERS) [12], a sharp distinction is now made between reference systems and reference frames. Briefly, a reference system provides the theory to obtain coordinates, whereas a reference frame is an actual materialization of coordinates. A reference frame is needed to conduct practical GNSS applications.

The fundamental ECEF reference frame is the International Terrestrial Reference Frame (ITRF). The ITRF is maintained through the international cooperation of scientists through the IERS. The IERS is established by the International Astronomical Union and the International Union of Geodesy and Geophysics, and operates as a service under the International Association of Geodesy (IAG). The IERS provides reference systems and reference frames in both ECI and ECEF forms, Earth orientation parameters to convert between ECI and ECEF, and recommended theory and practices in establishing reference systems and reference frames [12–14].

The work of the IERS is not restricted to GNSS. Rather, the IERS incorporates every suitable technology in establishing an ITRF. The IERS Techniques Centers are the International GNSS Service (IGS), the International Laser Ranging Service, the International Very Long Baseline Interferometry (VLBI) Service, and the International DORIS Service. The different measurement technologies complement one another and serve as checks against systematic errors in the ITRF combination solutions.

The ITRF realizations are issued on a regular basis. These realizations include coordinates and velocities of permanent ground stations. Each combination uses the latest theory and methods, and includes the newest measurements from both

legacy and modernized systems. The progression of longer and improved data sets and theory insures a continual improvement in the ITRF. Past materializations include ITRF94, ITRF96, ITRF97, ITRF2000, ITRF2005, and ITRF2008. Since January 21, 2016, the newest ITRF frame is ITRF2014 [15].

ITRF realizations are in ECEF Cartesian coordinates. The IERS does not establish an ellipsoidal figure of the Earth. However, the International Association of Geodesy (IAG) adopted a figure called the Geodetic Reference System 1980 (GRS 80) ellipsoid, which is in widespread use. Quantities suitable for use with coordinate conversion by Table 2.2 are provided next.

For GNSS applications, access to the ITRF is obtained through the products of the IGS. The IGS is a voluntary federation of over 200 organizations throughout the world. The IGS objective is to provide GNSS satellite orbits and clock models of the highest accuracy. This is achieved with a global network of over 400 reference stations [16].

The principal IGS products are satellite orbit and clock error values in an ECEF frame denoted IGS14. This frame is aligned with the ITRF2014, and carries a different designation due to its method of computation. As of this edition, IGS routinely distributes ultrarapid, rapid, and final orbits and clocks for GPS, and final orbits for GLONASS. In addition, IGS provides station coordinates and velocities, GNSS receiver and satellite antenna models, and tropospheric, ionospheric, and Earth orientation parameters. With these products and suitable GNSS receiver data, it is possible to obtain ITRF2014 coordinates at the highest levels of accuracy.

IGS products were initially developed to support postprocessing applications. In time, the products grew to include near-real-time and real-time needs. However, from the beginning, SATNAV systems were engineered to function in a standalone mode, without the presence of supporting Internet data streams. The standalone mode entails satellite orbit and clock data transmitted in navigation messages as part of a GNSS signal. Also, various SATNAV systems can maintain their own tracking networks, and establish their own versions of an ECEF reference frame. Such SATNAV system reference frames may or may not have a close coincidence with ITRF2014. Further description of specific SATNAV system reference frames and their relationships with ITRF are found in subsequent chapters detailing these various GNSS components.

2.3 Fundamentals of Satellite Orbits

2.3.1 Orbital Mechanics

As described in Section 2.1, a GNSS user needs accurate information about the positions of GNSS satellites to determine his or her position. Therefore, it is important

Table 2.2 Quantities for the GRS 80 Ellipsoid

Parameter	*Value*
Semimajor axis, a	6,378.137 km
Semiminor axis, b	6,356.7523141 km
Square eccentricity, e^2	0.00669438002290
Square second eccentricity, e'^2	0.00673949677548

to understand how GNSS orbits are characterized. We begin by describing the forces acting on a satellite, the most significant of which is the Earth's gravitation. If the Earth were perfectly spherical and of uniform density, then the Earth's gravitation would behave as if the Earth were a point mass. Let an object of mass m be located at position vector $\mathbf{r}$ in an ECI coordinate system. If G is the universal gravitational constant, M is the mass of the Earth, and the Earth's gravitation acts as a point mass, then, according to Newton's laws, the force, $\mathbf{F}$, acting on the object would be given by

$$\mathbf{F} = m\mathbf{a} = -G\frac{mM}{r^3}\mathbf{r} \tag{2.3}$$

where a is the acceleration of the object, and $r = |\mathbf{r}|$. The minus sign on the right side of (2.3) results from the fact that gravitational forces are always attractive. Since acceleration is the second time derivative of position, (2.3) can be rewritten as follows:

$$\frac{d^2\mathbf{r}}{dt^2} = -\frac{\mu}{r^3}\mathbf{r} \tag{2.4}$$

where μ is the product of the universal gravitation constant and the mass of the Earth. Equation (2.4) is the expression of two-body or Keplerian satellite motion, in which the only force acting on the satellite is the point-mass Earth. Because the Earth is not spherical and has an uneven distribution of mass, (2.4) does not model the true acceleration due to the Earth's gravitation. If the function V measures the true gravitational potential of the Earth at an arbitrary point in space, then (2.4) may be rewritten as follows:

$$\frac{d^2\mathbf{r}}{dt^2} = \nabla V \tag{2.5}$$

where ∇ is the gradient operator, defined as follows:

$$\nabla V \underset{\text{def}}{=} \begin{bmatrix} \dfrac{\partial V}{\partial x} \\ \dfrac{\partial V}{\partial y} \\ \dfrac{\partial V}{\partial z} \end{bmatrix}$$

Notice that for two-body motion, $V = \mu/r$:

$$\nabla(\mu/r) = \mu\begin{bmatrix}\frac{\partial}{\partial x}\left(r^{-1}\right)\\ \frac{\partial}{\partial y}\left(r^{-1}\right)\\ \frac{\partial}{\partial z}\left(r^{-1}\right)\end{bmatrix} = -\frac{\mu}{r^2}\begin{bmatrix}\frac{\partial}{\partial x}\left(x^2+y^2+z^2\right)^{1/2}\\ \frac{\partial}{\partial y}\left(x^2+y^2+z^2\right)^{1/2}\\ \frac{\partial}{\partial z}\left(x^2+y^2+z^2\right)^{1/2}\end{bmatrix}$$

$$= -\frac{\mu}{2r^2}\left(x^2+y^2+z^2\right)^{-1/2}\begin{bmatrix}2x\\ 2y\\ 2z\end{bmatrix} = -\frac{\mu}{r^3}\begin{bmatrix}x\\ y\\ z\end{bmatrix} = -\frac{\mu}{r^3}\mathbf{r}$$

Therefore, with $V = \mu/r$, (2.5) is equivalent to (2.4) for two-body motion. In the case of true satellite motion, the Earth's gravitational potential is modeled by a spherical harmonic series. In such a representation, the gravitational potential at a point P is defined in terms of the point's spherical coordinates (r, ϕ', α), where $r = |\mathbf{r}|$, ϕ' is the geocentric latitude of the point P (i.e., the angle between $\mathbf{r}$ and the xy-plane), and α is the right ascension of P (i.e., the angle measured in the xy-plane between the x-axis and the projection of P into the xy-plane). The geometry is illustrated in Figure 2.12. Note that geocentric latitude is defined differently from geodetic latitude, as defined in Section 2.2.5.1.

The spherical harmonic series representation of the Earth's gravitational potential as a function of the spherical coordinates of a position vector $\mathbf{r} = (r, \phi', \alpha)$, is as follows:

$$V = \frac{\mu}{r}\left[1 + \sum_{l=2}^{\infty}\sum_{m=0}^{l}\left(\frac{a}{r}\right)^l P_{lm}\left(\sin\phi'\right)\left(C_{lm}\cos m\alpha + S_{lm}\sin m\alpha\right)\right] \tag{2.6}$$

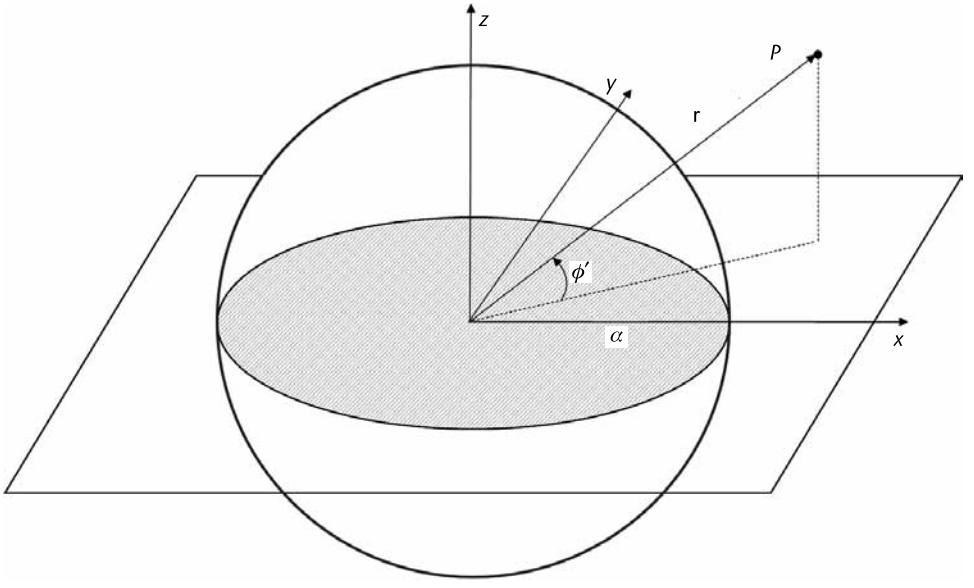

Figure 2.12 Illustration of spherical coordinate geometry

where

r = distance of P from the origin
ϕ' = geocentric latitude of P
α = right ascension of P
a = mean equatorial radius of the Earth
P_{lm} = associated Legendre function
C_{lm} = spherical harmonic cosine coefficient of degree l and order m
S_{lm} = spherical harmonic sine coefficient of degree l and order m

Notice that the first term of (2.6) is the two-body potential function. Additional forces acting on satellites include the third-body gravitation from the Sun and Moon. Modeling third-body gravitation requires knowledge of the solar and lunar positions in the ECI coordinate system as a function of time. Polynomial functions of time are generally used to provide the orbital elements of the Sun and Moon as functions of time. A number of alternative sources and formulations exist for such polynomials with respect to various coordinate systems; for example, see [17]. Another force acting on satellites is solar radiation pressure, which results from momentum transfer from solar photons to a satellite. Solar radiation pressure is a function of the Sun's position, the projected area of the satellite in the plane normal to the solar line of sight, and the mass and reflectivity of the satellite. There are additional forces acting on a satellite, including outgassing (i.e., the slow release of gases trapped in the structure of a satellite), the Earth's tidal variations, and orbital maneuvers. To model a satellite's orbit very accurately, all these perturbations to the Earth's gravitational field must be modeled. For the purposes of this text, we will collect all these perturbing accelerations in a term $\mathbf{a}_d$, so that the equations of motion can be written as

$$\frac{d^2\mathbf{r}}{dt^2} = \nabla V + \mathbf{a}_d \tag{2.7}$$

There are various methods of representing the orbital parameters of a satellite. One obvious representation is to define a satellite's position vector, $\mathbf{r}_0 = \mathbf{r}(t_0)$, and velocity vector, $\mathbf{v}_0 = \mathbf{v}(t_0)$, at some reference time, t_0. Given these initial conditions, we could solve the equations of motion (2.7) for the position vector $\mathbf{r}(t)$ and the velocity vector $\mathbf{v}(t)$ at any other time t. Only the two-body equation of motion (2.4) has an analytical solution, and even in that simplified case, the solution cannot be accomplished entirely in a closed form. The computation of orbital parameters from the fully perturbed equations of motion (2.7) requires numerical integration.

Although many applications, including GNSS, require the accuracy provided by the fully perturbed equations of motion, orbital parameters are often defined in terms of the solution to the two-body problem. It can be shown that there are six constants of integration, or integrals for the equation of two-body motion, (2.4). Given six integrals of motion and an initial time, one can find the position and ve-

locity vectors of a satellite on a two-body orbit at any point in time from the initial conditions.

One of the most popular (and oldest) ways to formulate and solve the two-body problem uses a particular set of six integrals or constants of motion known as the Keplerian orbital elements. These Keplerian elements depend on the fact that, for any initial conditions $\mathbf{r}_0$ and $\mathbf{v}_0$ at time t_0, the solution to (2.4) (i.e., the orbit), will be a planar conic section. The first three Keplerian orbital elements, illustrated in Figure 2.13, define the shape of the orbit. Figure 2.13 shows an elliptical orbit that has semimajor axis a and eccentricity e. (Hyperbolic and parabolic trajectories are possible but not relevant for Earth-orbiting satellites, such as in GNSS.) For elliptical orbits, the eccentricity, e, is related to the semimajor axis, a, and the semiminor axis, b, as follows:

$$e = \sqrt{1 - \frac{b^2}{a^2}}$$

In Figure 2.13, the elliptical orbit has a focus at point F, which corresponds to the center of mass of the Earth (and hence the origin of an ECI or ECEF coordinate system). The time t_0 at which the satellite is at some reference point A in its orbit is known as the epoch. The point P where the satellite is closest to the center of the Earth is known as perigee, and the time at which the satellite passes perigee, τ, is another Keplerian orbital parameter. In summary, the three Keplerian orbital elements that define the shape of the elliptical orbit and time relative to perigee are as follows: a = semimajor axis of the ellipse, e = eccentricity of the ellipse, and τ = time of perigee passage.

Although the Keplerian integrals of two-body motion use time of perigee passage as one of the constants of motion, there is an equivalent parameter used in GNSS applications known as the mean anomaly at epoch. Mean anomaly is an angle that is related to the true anomaly at epoch, which is illustrated in Figure 2.13 as the angle ν. After defining true anomaly precisely, the transformation to mean

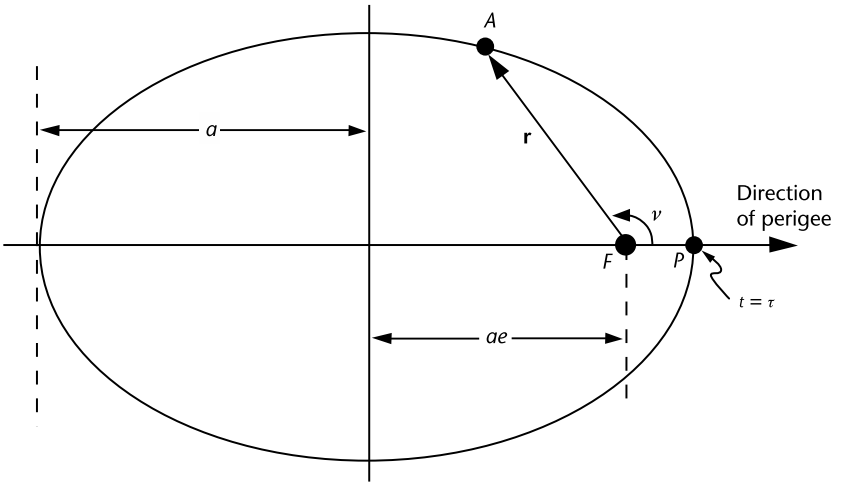

Figure 2.13 The three Keplerian orbital elements defining the shape of the satellite's orbit.

anomaly and the demonstration of equivalence to time of perigee passage will be shown.

True anomaly is the angle in the orbital plane measured counterclockwise from the direction of perigee to the satellite. In Figure 2.13, the true anomaly at epoch is $\nu = \angle PFA$. From Kepler's laws of two-body motion, it is known that true anomaly does not vary linearly in time for noncircular orbits. Because it is desirable to define a parameter that does vary linearly in time, two definitions are made that transform the true anomaly to the mean anomaly, which is linear in time. The first transformation produces the eccentric anomaly, which is illustrated in Figure 2.14 with the true anomaly. Geometrically, the eccentric anomaly is constructed from the true anomaly first by circumscribing a circle around the elliptical orbit. Next, a perpendicular is dropped from the point A on the orbit to the major axis of the orbit, and this perpendicular is extended upward until it intersects the circumscribed circle at point B. The eccentric anomaly is the angle measured at the center of the circle, O, counterclockwise from the direction of perigee to the line segment OB. In other words, $E = \angle POB$. A useful analytical relationship between eccentric anomaly and true anomaly is as follows [17]:

$$E = 2\arctan\left[\sqrt{\frac{1-e}{1+e}}\tan\left(\frac{1}{2}\nu\right)\right] \tag{2.8}$$

Once the eccentric anomaly has been computed, the mean anomaly is given by Kepler's equation

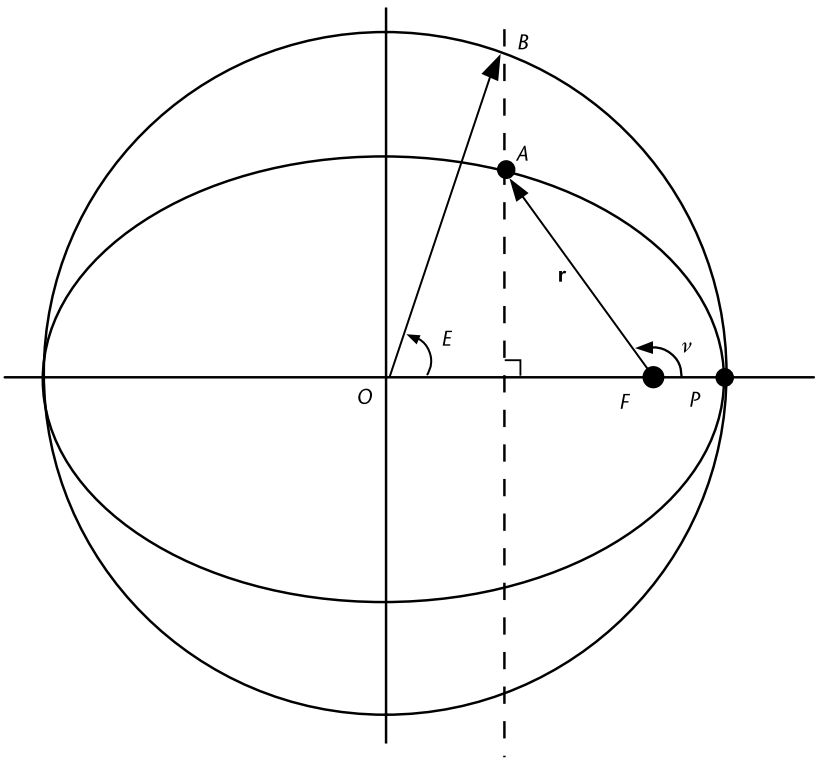

Figure 2.14 Relationship between eccentric anomaly and true anomaly.

$$M = E - e\sin E \tag{2.9}$$

As stated previously, the importance of transforming from the true to the mean anomaly is that time varies linearly with the mean anomaly. That linear relationship is as follows:

$$M - M_0 = \sqrt{\frac{\mu}{a^3}}(t - t_0) \tag{2.10}$$

where M_0 is the mean anomaly at epoch t_0, and M is the mean anomaly at time t. From Figures 2.13 and 2.14 and (2.8) and (2.9), it can be verified that $M = E = \nu = 0$ at the time of perigee passage. Therefore, if we let $t = \tau$, (2.10) provides a transformation between mean anomaly and time of perigee passage:

$$M_0 = -\sqrt{\frac{\mu}{a^3}}(\tau - t_0) \tag{2.11}$$

From (2.11), it is possible to characterize the two-body orbit in terms of the mean anomaly, M_0, at epoch t_0, instead of the time of perigee passage τ.

Another parameter commonly used by GNSS systems to characterize orbits is known as mean motion, which is given the symbol n and is defined to be the time derivative of the mean anomaly. Since the mean anomaly was constructed to be linear in time for two-body orbits, mean motion is a constant. From (2.10), we find the mean motion as follows:

$$n \underset{\text{def}}{=} \frac{dM}{dt} = \sqrt{\frac{\mu}{a^3}}$$

From this definition, (2.10) can be rewritten as $M - M_0 = n(t - t_0)$.

Mean motion can also be used to express the orbital period P of a satellite in two-body motion. Since mean motion is the (constant) rate of change of the mean anomaly, the orbital period is the ratio of the angle subtended by the mean anomaly over one orbital period to the mean motion. It can be verified that the mean anomaly passes through an angle of 2π radians during one orbit. Therefore, the orbital period is calculated as follows:

$$P = \frac{2\pi}{n} = 2\pi\sqrt{\frac{a^3}{\mu}} \tag{2.12}$$

Figure 2.15 illustrates the three additional Keplerian orbital elements that define the orientation of an elliptical orbit. The coordinates in Figure 2.15 could refer either to an ECI or to an ECEF coordinate system, in which the xy-plane is the Earth's equatorial plane. The following three Keplerian orbital elements define the

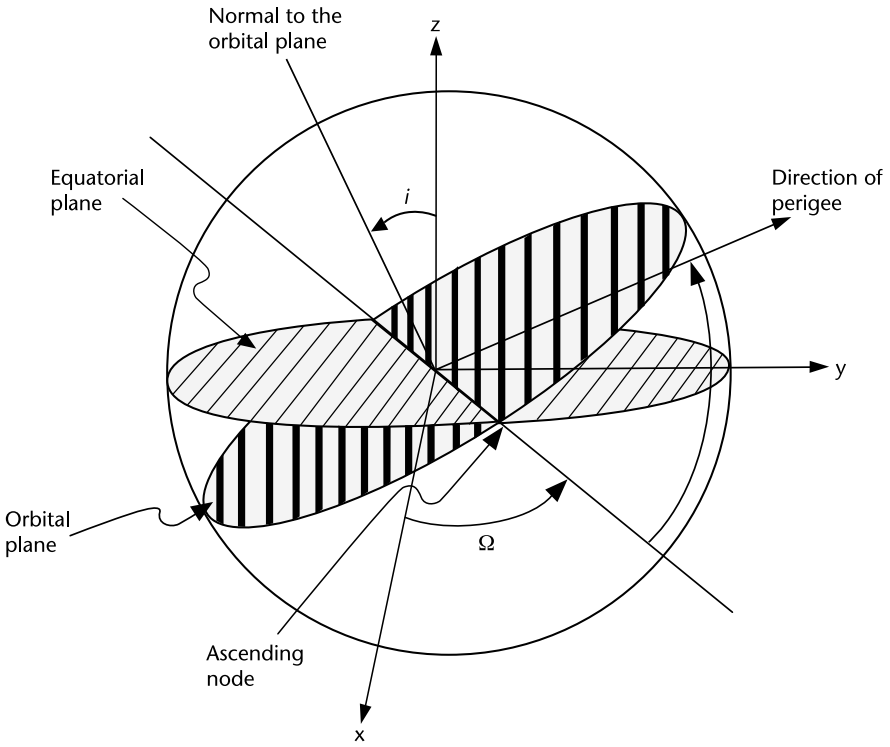

Figure 2.15 The three Keplerian orbital elements defining the orientation of the orbit.

orientation of the orbit in the ECEF coordinate system: i = inclination of orbit, Ω = longitude of the ascending node, and ω = argument of perigee.

Inclination is the dihedral angle between the Earth's equatorial plane and the satellite's orbital plane. The other two Keplerian orbital elements in Figure 2.15 are defined in relation to the ascending node, which is the point in the satellite's orbit where it crosses the equatorial plane with a $+z$ component of velocity (i.e., going from the southern to the northern hemisphere). The orbital element that defines the angle between the $+x$-axis and the direction of the ascending node is called the right ascension of the ascending node, abbreviated as RAAN. Because the $+x$-axis is fixed in the direction of the prime meridian (0° longitude) in the ECEF coordinate system, the right ascension of the ascending node is actually the longitude of the ascending node, Ω, if an ECEF coordinate system is being used. The final orbital element, known as the argument of perigee, ω, measures the angle from the ascending node to the direction of perigee in the orbit. Notice that Ω is measured in the equatorial plane, whereas ω is measured in the orbital plane.

In the case of the fully perturbed equation of motion, (2.7), it is still possible to characterize the orbit in terms of the six integrals of two-body motion, but those six parameters will no longer be constant. A reference time is associated with two-body orbital parameters used to characterize the orbit of a satellite moving under the influence of perturbing forces. At the exact reference time, the reference orbital parameters will describe the true position and velocity vectors of the satellite, but as time progresses beyond (or before) the reference time, the true position

and velocity of the satellite will increasingly deviate from the position and velocity described by the six two-body integrals or parameters.

2.3.2 Constellation Design

A satellite constellation (i.e., group of satellites fulfilling an overall mission) is characterized by the set of orbital parameters for the individual satellites in that constellation. The orbital parameters used are often the Keplerian orbital elements defined in Section 2.3.1. The design of a satellite constellation entails the selection of orbital parameters that optimize some objective function of the constellation [typically to maximize some set of performance parameters at minimum cost (i.e., with the fewest satellites)]. The design of satellite constellations has been the subject of numerous studies and publications, some of which are summarized next. Our purpose here is to provide a general overview of satellite constellation design, to summarize the salient considerations in the design of constellations for satellite navigation, to provide some perspective on the selection of the global (i.e., core) constellations (BeiDou, Galileo, GLONASS, and GPS).

2.3.2.1 Overview of Constellation Design

Given innumerable combinations of satellite orbital parameters in a constellation, it is convenient to segregate orbits into categories. One categorization of orbits is by eccentricity:

- Circular orbits have zero (or nearly zero) eccentricity.
- Highly elliptical orbits (HEO) have large eccentricities (typically with $e > 0.6$).
- Here we will address only circular orbits. Another categorization of orbits is by altitude: Geosynchronous Earth orbit (GEO) is an orbit with period equal to the duration of the sidereal day [substituting $P = 23$ hours, 56 minutes, 4.1 seconds into (2.12) yields $a = 42{,}164.17$ km as the orbital semimajor axis for GEO, or an altitude of 35,786 km].
- Low Earth orbit (LEO) is a class of orbits with altitude typically less than 1,500 km.
- Medium Earth orbit (MEO) is a class of orbits with altitudes below GEO and above LEO, with most practical examples being in the range of roughly 10,000–25,000-km altitude.
- Supersynchronous orbits are those with altitude greater than GEO (greater than 35,786 km).

Note that GEO defines an orbital altitude such that the period of the orbit equals the period of rotation of the Earth in inertial space (the sidereal day). A geostationary orbit is a GEO orbit with zero inclination and zero eccentricity. In this special case, a satellite in geostationary orbit has no apparent motion to an observer on Earth, because the relative position vector from the observer to the satellite (in ECEF coordinates) remains fixed over time. In practice, due to orbital

perturbations, satellites never stay in exactly geostationary orbit; therefore, even geostationary satellites have some small residual motion relative to users on the Earth. Geostationary GEO satellites are used most often for satellite communications. However, it is also sometimes of interest to incline a GEO orbit to provide coverage also of the Earth's poles, but at the expense of the satellite having greater residual motion relative to the earth. As we will see, the Chinese BeiDou constellation and Japanese QZSS specifically make use of such inclined GEO satellites.

Another categorization of orbits is by inclination:

- Equatorial orbits have zero inclination; hence a satellite in equatorial orbit travels in the Earth's equatorial plane.
- Polar orbits have 90° inclination (or close to 90° inclination); hence, a satellite in polar orbit passes through (or near) the Earth's axis of rotation.
- Prograde orbits have nonzero inclination with right ascension of the ascending node less than 180° (and hence have ground tracks that go in general from southwest to northeast).
- Retrograde orbits have nonzero inclination with right ascension of the ascending node greater than 180° (and hence have ground tracks that go in general from northwest to southeast).
- Collectively, prograde and retrograde orbits are known as "inclined."

Finally, there are specialized classes of orbits that combine orbital parameters in specific ways to achieve unique orbital characteristics. One such example is Sun-synchronous orbits, which are used for many optical Earth-observing satellite missions. A Sun-synchronous orbit is one in which the orbit is nearly polar, and the local time (i.e., at the subsatellite point on Earth) when the satellite crosses through the equatorial plane is the same on every orbital pass. In this way, the satellite motion is synchronized relative to the Sun, which is achieved by selecting a specific inclination as a function of desired orbital altitude.

Selection of a class of orbits for a particular application is made based on the requirements of that application. For example, in many high-bandwidth satellite communications applications (e.g., direct broadcast video or high rate data trunking), it is desirable to have a nearly geostationary orbit to maintain a fixed line of sight from the user to the satellite to avoid the need for the user to have an expensive steerable or phased array antenna. However, for lower-bandwidth mobile satellite service applications where lower data latency is desirable, it is preferable to use LEO or MEO satellites to reduce range from the user to the satellite. For satellite navigation applications, it is necessary to have multiple (at least four) satellites in view at all time, usually worldwide.

Apart from orbital geometry, there are several other significant considerations when configuring a satellite constellation. One such consideration is the requirement to maintain orbital parameters within a specified range. Such orbital maintenance is called stationkeeping, and it is desirable to minimize the frequency and magnitude of maneuvers required over the lifetime of a satellite. This is true in all applications because of the life-limiting factor of available fuel on the satellite, and it is particularly true for satellite navigation applications, because satellites are not

immediately available to users after a stationkeeping maneuver while orbital and clock parameters are stabilized and ephemeris messages are updated. Therefore, more frequent stationkeeping maneuvers both reduce the useful lifetime of satellites and reduce the overall availability of the constellation to users. Some orbits have a resonance effect, in which there is an increasing perturbation in a satellite's orbit due to the harmonic effects of (2.6). Such orbits are undesirable because they require more stationkeeping maneuvers to maintain a nominal orbit.

Another consideration in constellation design is radiation environment, caused by the Van Allen radiation belts, in which charged particles are trapped by the Earth's magnetic field. The radiation environment (measured by flux of trapped protons and electrons) is a function of height above the Earth's surface and of the out-of-plane angle relative to the equator. LEO satellites below 1,000 km altitude operate in a relatively benign radiation environment, whereas MEO satellites at 15,000–25,000-km altitude will pass through the radiation environment at every equatorial plane crossing. A high radiation environment drives satellite design in a number of ways, including the need for space-class electronics components, installing redundant equipment, and shielding all the way from component to spacecraft level. These design impacts result in increased mass and cost of the satellite.

2.3.2.2 Inclined Circular Orbits

Theoretical studies of satellite constellations typically focus on some particular subset of orbital categories. For example, Walker extensively studied inclined circular orbits [18], Rider further studied inclined circular orbits to include both global and zonal coverage [19], and Adams and Rider studied circular polar orbits [20]. These studies all focus on determining the set of orbits in their categories that require the fewest satellites to provide a particular level of coverage (i.e., the number of satellites in view from some region on Earth above some minimum elevation angle). The studies determine the optimal orbital parameters for a given category of orbits that minimize the number of satellites required to achieve the desired level of coverage. Satellites in a constellation are segregated into orbital planes, where an orbital plane is defined as a set of orbits with the same right ascension of the ascending node (and hence the satellites travel in the same plane in an ECI coordinate system). In the most general approach, Walker addresses constellations of satellites where each satellite can be in a different orbital plane, or there can be multiple satellites per plane. Rider's work assumes multiple satellites per orbital plane. In each case, the point of the study is to find the particular combination of orbital parameters (how many satellites, in how many planes, in what exact geometrical configuration and phasing) that minimize the number of satellites required to obtain a particular level of coverage. Usually this can be obtained with a Walker constellation with one satellite per orbital plane. However, there are additional considerations beyond just minimizing the total number of satellites in a constellation. For example, since on-orbit spares are usually desired for a constellation, and since maneuvers to change orbital planes consume considerable fuel, it is usually desired to suboptimize by selecting a constellation with multiple satellites per orbital plane, even though this usually requires a few extra satellites to achieve a given level of coverage.

One important result from the studies [18–20] is that the required number of satellites to achieve a desired level of coverage increases significantly the lower the

orbital altitude selected. This effect is illustrated in Figure 2.16, which shows the number of satellites required to achieve single worldwide coverage (above 0° elevation angle) as a function of orbital altitude, as shown by Rider [19]. In general, for every 50% reduction in orbital altitude, the required number of satellites increases by 75%. This becomes important when trading off satellite complexity versus orbital altitude in constellation design, as discussed next.

Practical applications of the theoretical work [18–20] have included the IRIDIUM LEO mobile satellite communications constellation, which was originally planned to be an Adams/Rider 77-satellite polar constellation and ended up as a 66-satellite polar constellation, the Globalstar LEO mobile satellite communications constellation, which was originally planned to be a Walker 48-satellite inclined circular constellation of 8 planes, and most recently a proposed constellation called OneWeb with 648 satellites in polar orbits to provide internet service. In addition, the global constellations (GPS, GLONASS, Galileo, and BeiDou) all employ constellations making use of the principles set forth in [18, 19].

Rider Constellations

As an example of how to use one of these constellation design studies, consider Rider's work [19] on inclined circular orbits. Rider studied the class of orbits that are circular and of equal altitude and inclination. In Rider's work, the constellation is divided into a number of orbital planes, P, with a number of satellites per plane, S. Also, the satellites in this study are assumed to have equal phasing between planes (i.e., satellite 1 in plane 1 passes through its ascending node at the same time as satellite 1 in plane 2). Figure 2.17 illustrates equal versus unequal phasing between planes in the case of two orbital plans with three equally spaced satellites per plane ($P = 2$, $S = 3$). The orbital planes are equally spaced around the equatorial plane, so that the difference in right ascension of ascending node between planes equals 360°/P, and satellites are equally spaced within each orbital plane.

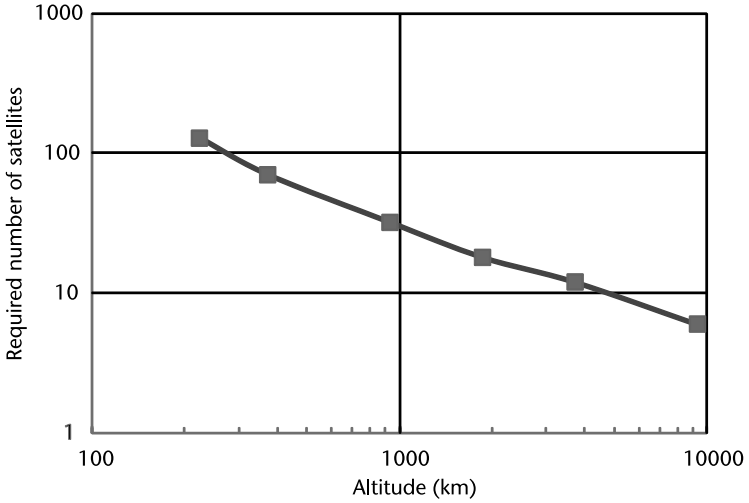

Figure 2.16 Number of satellites required to achieve at least one satellite in view worldwide at all times.

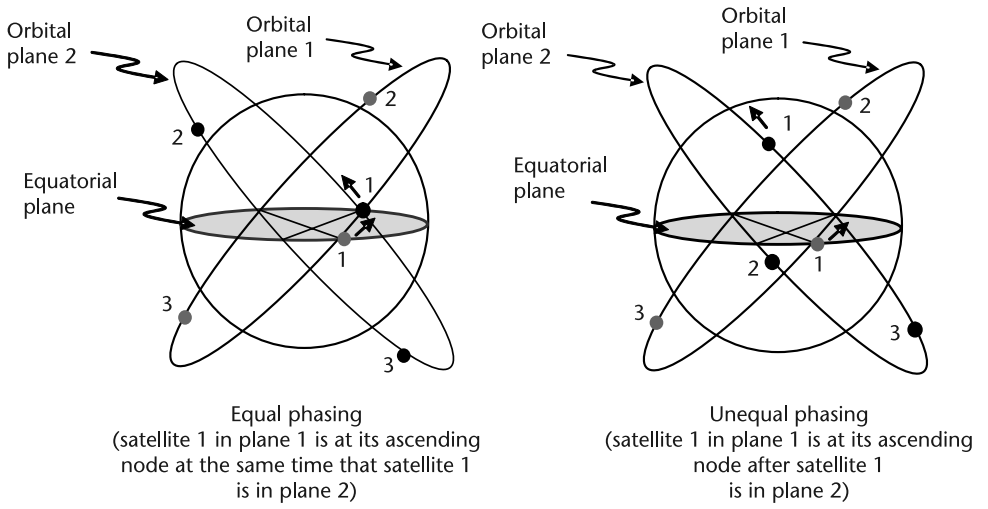

Figure 2.17 Equal versus unequal phasing between orbital planes.

Rider [19] made the following definitions: α = elevation angle, R_e = spherical radius of the Earth (these studies all assume a spherical Earth), and h = orbital altitude of the constellation being studied.

Then the Earth central angle, θ, as shown in Figure 2.18, is related to these parameters as follows:

$$\cos(\theta + \alpha) = \frac{\cos\alpha}{1 + {}^{h}\!/_{Re}} \tag{2.13}$$

From (2.13), given an orbital altitude, h, and a minimum elevation angle, α, the corresponding Earth central angle, θ, can be computed. Rider then defines a half street width parameter, c, which is related to the Earth central angle, θ, and the number of satellites per orbital plane, S, as follows:

$$\cos\theta = (\cos c)\left(\cos\frac{\pi}{S}\right) \tag{2.14}$$

Finally, Rider's analysis produces a number of tables that relate optimal combinations of orbital inclination, i, half street width, c, and number of orbital planes, P, for various desired Earth coverage areas (global versus mid-latitude versus equatorial versus polar) and various levels of coverage (minimum number of satellites in view).

Walker Constellations

It turns out that the more generalized Walker constellations [18] can produce a given level of coverage with fewer satellites in general than the Rider constellations [19]. Walker constellations use circular inclined orbits of equal altitude and inclination, the orbital planes are equally spaced around the equatorial plane, and satellites

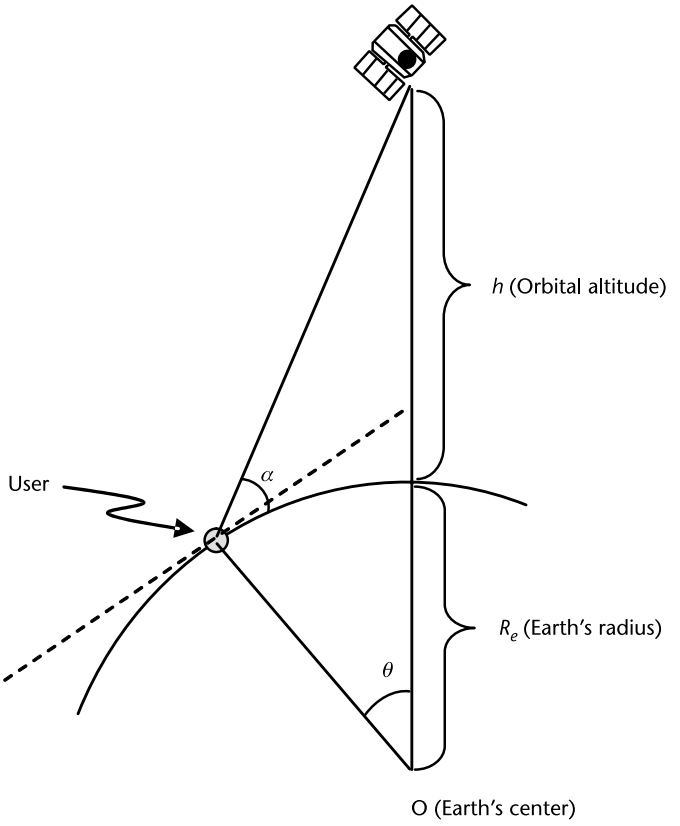

Figure 2.18 Relationship between elevation angle and Earth central angle (θ).

are equally spaced within orbital planes, as with Rider constellations. However, Walker constellations allow more general relationships between the number of satellites per plane and the phasing between planes. To that end, Walker introduced the notation $T/P/F$, where T is the total number of satellites in the constellation, P is the number of orbital planes, and F is the phase offset factor that determines the phasing between adjacent orbital planes (see Figure 2.17 for an illustration of the concept of phasing between orbital planes). With the number of satellites per plane, S, it is obvious that $T = S \times P$. F is an integer such that $0 \leq F \leq P - 1$, and the offset in mean anomaly between the first satellite in each adjacent orbital plane is $360° \times F/P$. That is, when the first satellite in plane 2 is at its ascending node, the first satellite in plane 1 will have covered an orbital distance of $(360° \times F/P)$ degrees within its orbital plane.

Typically, with one satellite per plane, a value of F can be found such that a Walker constellation can provide a given level of coverage with fewer satellites than a Rider constellation. However, such Walker constellations with one satellite per plane are less robust against failure than constellations with multiple satellites per plane, because on-orbit sparing is nearly impossible with only one satellite per plane. In such a sparing scenario, it would be required to reposition the satellite from the spare plane into the plane of a failed satellite, but the cost in fuel is extremely prohibitive to execute such an orbital maneuver. To give an idea, a single plane change would require approximately 30 times the amount of fuel that is

currently budgeted on the Galileo satellites for maneuvers over their entire lifetime. Because satellites can therefore be repositioned realistically only within an orbital plane, there is greater application of Rider-type constellations or Walker constellations with multiple satellites per plane versus Walker constellations with a single satellite per plane.

As a specific example of constellation design using the work of Walker and Rider ([18] and [19]), consider a MEO satellite constellation providing 4-fold worldwide continuous coverage above a minimum 5° elevation angle for the satellite navigation application. In this example, the objective is to minimize the number of satellites providing this level of coverage within the class of Rider orbits. Specifically, consider the case with $h = 20{,}182$ km (corresponding to an orbital period of approximately12 hours). With $\alpha = 5°$, the Earth central angle θ can be computed from (2.13) to be 71.2°.

Rider's results in Table 4 of [19] then show that with 6 orbital planes, the optimal inclination is 55°, and $c = 44.92°$. We now have enough information to solve equation (2.14) for S. This solution is $S = 2.9$, but since satellites come only in integer quantities, one must round up to 3 satellites per plane. Hence, Rider's work indicates that with 6 orbital planes, one must have 3 satellites per plane, for a total of 18 satellites, to produce continuous worldwide coverage with a minimum of 4 satellites above a minimum 5° elevation angle. With 5 orbital planes of the same altitude and with the same coverage requirement, Rider's work shows $c = 55.08°$, and $S = 3.2$, which rounds up to 4 satellites per plane. In this case, 20 total satellites would be required to provide the same level of coverage. Likewise, with 7 orbital planes, the requirement is 3 satellites per plane, for a total of 21 satellites. Therefore, the optimal Rider constellation configuration to provide worldwide 4-fold coverage above 5 degrees elevation angle is a 6×3 constellation ($P = 6$, $S = 3$) for a total of 18 satellites. It turns out that in the early 1980s, the U.S. Air Force was looking at smaller GPS constellation alternatives, consisting of different configurations with 18 total satellites [21]. Note that for the navigation application, where there are more considerations than just the total number of satellites in view, it turns out to be preferable to modify Walker or Rider constellations, for example, by unevenly spacing the satellites in each orbital plane. The details of these additional considerations will be explored more fully in the following section.

2.3.2.3 Constellation Design Considerations for Satellite Navigation

Satellite navigation constellations have very different geometrical constraints from satellite communications systems, the most obvious of which is the need for more multiplicity of coverage (i.e., more required simultaneous satellites in view for the navigation applications). As discussed in Section 2.5.2, the GNSS navigation solution requires a minimum of four satellites to be in view of a user to provide the minimum of four measurements necessary for the user to determine three-dimensional position and time. Therefore, a critical constraint on a GNSS constellation is that it must provide a minimum of 4-fold coverage at all times. In order to ensure this level of coverage robustly, a nominal GNSS constellation is designed to provide more than fourfold coverage so that the minimum of four satellites in view can be maintained even with a satellite failure. Also, more than fourfold coverage is necessary for user equipment to be able to determine autonomously if a GNSS satellite

is experiencing a signal or timing anomaly, and therefore to exclude such a satellite from the navigation solution (this process is known as integrity monitoring); see Section 11.4. Therefore, the practical constraint for coverage of a GNSS constellation is minimum sixfold coverage above a 5° elevation angle.

Constellation design for satellite navigation has the following major constraints and considerations:

1. Coverage needs to be global.
2. At least 6 satellites need to be in view of any user position at all times.
3. To provide the best navigation accuracy, the constellation needs to have good geometric properties, which entail a dispersion of satellites in both azimuth and elevation angle from users anywhere on the Earth (a discussion of the effects of geometric properties on navigation accuracy is provided in Section 11.2).
4. The constellation needs to be robust against single satellite failures.
5. The constellation must be maintainable, that is, it must be relatively inexpensive to reposition satellites within the constellation.
6. Stationkeeping requirements need to be manageable. In other words, it is preferable to minimize the frequency and magnitude of maneuvers required to maintain the satellites within the required range of their orbital parameters.
7. Orbital altitude must be selected to achieve a balance between payload size and complexity versus required constellation size to achieve a minimum sixfold coverage. The higher the orbital altitude, the fewer the satellites required to achieve sixfold coverage, but the larger and more complex the payload and hence satellite. Payload complexity increases at higher altitudes, for example, due to the increased transmitter power and antenna size required to achieve a certain minimum received signal strength on the ground for a user.

2.4 GNSS Signals

This section provides an overview of GNSS signals including commonly used signal components. This discussion is followed by a description of important signal characteristics such as auto-correlation and cross-correlation functions.

2.4.1 Radio Frequency Carrier

Every GNSS signal is generated using one or more radio frequency (RF) carriers, which are nominally perfect sinusoidal voltages produced within the transmitter (see Figure 2.19). As shown in Figure 2.19, one important characteristic of an RF carrier is the time interval, T_0, between recurrences of amplitude (e.g., peak-to-peak) in units of seconds. Such a recurrence in amplitude is referred to as a cycle and the time interval corresponding to one cycle is referred to as the period. More commonly used in practice to characterize RF carriers is the carrier frequency, which is the reciprocal of the period, $f_0 = 1/T_0$, expressed in units of cycles/second

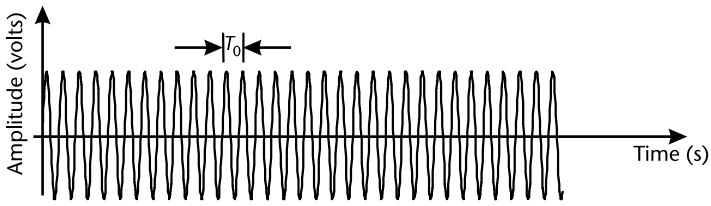

Figure 2.19 Radio frequency carrier.

or equivalently hertz. (By definition 1 Hz is one cycle/second). Metric prefixes are frequently encountered, for example, 1 kHz = 10^3 Hz, 1 MHz = 10^6 Hz, and 1 GHz = 10^9 Hz.

Most GNSS signals today use carrier frequencies in the L-band, which is defined by the Institute of Electrical and Electronics Engineers (IEEE) to be the range of 1 to 2 GHz. L-band offers several advantages for GNSS signals as compared to other bands. At lower frequencies, the Earth's atmosphere results in larger delays and inhomogeneities in the atmosphere cause more severe fading in received signal strength. At greater frequencies, additional satellite power is required and precipitation (e.g., rain) attenuation can be significant. Two L-band frequency subsets have been allocated globally by the International Telecommunication Union (ITU) for radionavigation satellite services (RNSS), which is the name given by the global spectrum management community to the services provided by GNSS constellations. The RNSS allocations in L-band are for 1,164–1,300 MHz and 1,559–1,610 MHz. Two GNSS constellations discussed within this book additionally utilize S-band (2–4 GHz) navigation signals, and several GNSS service providers are considering the future addition of navigation signals in C-band (4–8 GHz).

2.4.2 Modulation

GNSS signals are designed to enable several functions:

- Precise ranging by the user equipment;
- Conveyance of digital information about the location of the GNSS satellites, clock errors, satellite health, and other navigation data;
- For some systems, utilization of a common carrier frequency among multiple satellites broadcasting simultaneously.

To accomplish these functions, some properties of the RF carrier must be varied with time. Such variation of an RF carrier is referred to as modulation. Consider a signal whose voltage is described by

$$s(t) = a(t)\cos\left[2\pi f(t)t + \phi(t)\right] \tag{2.15}$$

If the amplitude, $a(t)$, frequency, $f(t)$, and phase offset, $\phi(t)$, are nominally constant with respect to time, then this equation would describe an unmodulated carrier. Variation of amplitude, frequency, and phase, are referred to as amplitude modulation, frequency modulation, and phase modulation, respectively. If $a(t)$, $f(t)$, or $\phi(t)$ can take on any of an infinite set of values varying continuously over time,

then the modulation is referred to as analog. The GNSS navigation signals broadcast by satellite navigation systems described in this book use digital modulation, meaning that the modulation parameters can only take on a finite set of values that are only permitted to change at specific, discrete epochs of time.

2.4.2.1 Navigation Data

One example of a digital modulation that is frequently used to convey digital navigation data from GNSS satellites to receivers is binary phase shift keying (BPSK). BPSK is a simple digital signaling scheme in which the RF carrier is either transmitted as is or with a 180° phase shift over successive intervals of T_b seconds in time depending on whether a digital 0 or 1 is being conveyed by the transmitter to the receiver (see, e.g., [22]). From this viewpoint, BPSK is a digital phase modulation with two possibilities for the phase offset parameter: $\phi(t) = 0$ or $\phi(t) = \pi$.

A BPSK signal can alternatively be viewed as being created using amplitude modulation, as illustrated in Figure 2.20. Note, as shown in the figure, that the BPSK signal can be formed as the product of two time waveforms: the unmodulated RF carrier and a data waveform that takes on a value of either +1 or –1 for each successive interval of $T_b = 1/R_b$ seconds where R_b is the data rate in bits per second. The data waveform amplitude for the k-th interval of T_b seconds can be generated from the kth data bit to be transmitted using either the mapping [0, 1] → [–1, +1] or [0, 1] → [+1, –1]. Mathematically, the data waveform $d(t)$ can be described as:

$$d(t) = \sum_{k=-\infty}^{\infty} d_k p(t - kT_b) \tag{2.16}$$

where d_k is the kth data bit (in the set [–1,+1]) and $p(t)$ is a pulse shape. The data waveform alone is considered a baseband signal, meaning that its frequency content is concentrated around 0 Hz rather than the carrier frequency. Modulation by the RF carrier centers the frequency content of the signal about the carrier frequency, creating what is known as a bandpass signal.

The BPSK signal shown in Figure 2.20 uses rectangular pulses:

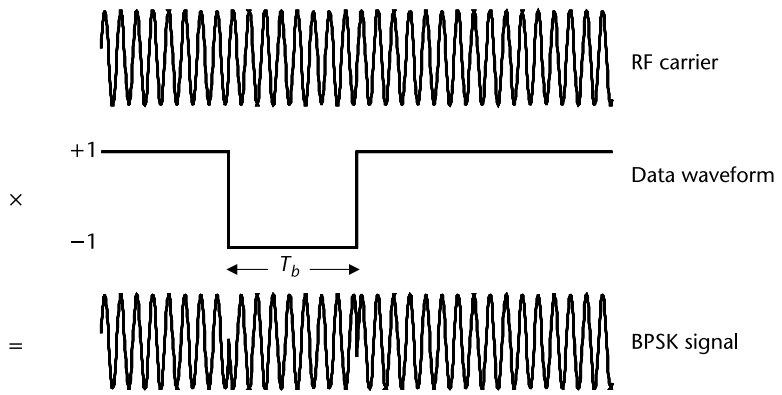

Figure 2.20 BPSK modulation.

$$p(t) = \begin{cases} 1, & 0 \le t < T_b \\ 0, & \text{elsewhere} \end{cases} \tag{2.17}$$

but other pulse shapes may be used. For instance, Manchester encoding is a term that is used to describe BPSK signals that use pulses consisting of one cycle of a square wave.

In many modern GNSS signal designs, forward error correction (FEC) is employed for the navigation data whereby redundant bits (more than the original information bits) are transmitted over the channel according to some prescribed method, enabling the receiver to detect and correct some errors that may be introduced by noise, interference, or fading. When FEC is employed, common convention is to replace T_b with T_s and R_b with R_s to distinguish data symbols (actually transmitted) from data bits (that contain the information before FEC). The coding rate is the ratio R_b/R_s.

2.4.2.2 Direct Sequence Spread Spectrum

To enable precise ranging, all of the GNSS signals described in this book employ direct sequence spread spectrum (DSSS) modulation. As shown in Figure 2.21, DSSS signaling involves the modulation of an RF carrier with a spreading or pseudorandom noise (PRN) waveform, often (as shown in the figure) but not necessarily in addition to modulation of the carrier by a navigation data waveform. The spreading and data waveforms are similar but there are two important differences. First, the spreading waveform is deterministic (i.e., the digital sequence used to produce it is completely known, at least to the intended receivers). Second, the symbol rate of the spreading waveform is much higher than the symbol rate of the navigation data waveform. The digital sequences used to generate spreading waveforms are referred to by various names including ranging code, pseudorandom sequence, and PRN code. An excellent overview of pseudorandom sequences, including their generation, characteristics, and code families with good properties is provided in [23].

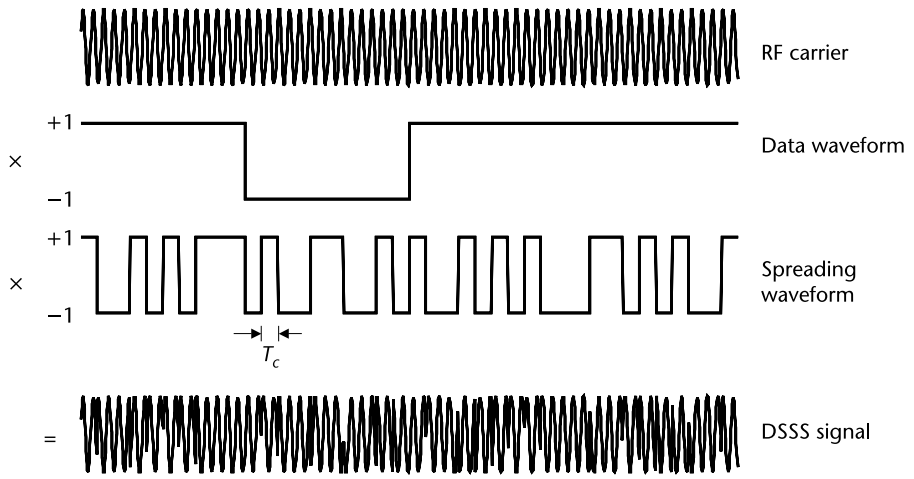

Figure 2.21 DSSS modulation.

GNSS signals that are intended to be used by the general public are referred to as open signals. Open GNSS signals use ranging codes that are unencrypted and periodic, with lengths varying from 511 to 767,250 bits. Some GNSS signals are only intended to be employed by authorized (e.g., military) users. To prevent general public use, authorized or restricted-use GNSS signals use ranging codes that are encrypted and thus aperiodic. Knowledge of the encryption scheme as well as secret numbers known as *private keys* are required to be able to fully process authorized GNSS signals.

To avoid confusion between information-bearing bits within the navigation data and the bits of the ranging code, the latter are often referred to as *chips*, which determine the polarity of the spreading symbols. The time duration of the spreading waveform corresponding to one chip of the ranging code is referred to as the chip period, and the reciprocal of the chip period as the chipping rate, R_c. The independent time parameter for the spreading waveform is often expressed in units of chips and referred to as code phase. The signal is called *spread spectrum*, due to the wider bandwidth occupied by the signal after modulation by the high rate spreading waveform. In general, the bandwidth is proportional to the chipping rate.

There are three primary reasons why DSSS waveforms are employed for satellite navigation. First and most importantly, the frequent phase inversions in the signal introduced by the spreading waveform enable precise ranging by the receiver. Second, the use of different spreading sequences from a well-designed set enables multiple satellites to transmit signals simultaneously and at the same frequency. A receiver can distinguish among these signals, based on their different codes. For this reason, the transmission of multiple DSSS signals having different spreading sequences on a common carrier frequency is referred to as code division multiple access (CDMA). Finally, as detailed in Chapter 9, DSSS provides significant rejection of narrowband interference.

2.4.2.3 Binary Offset Carrier

It should be noted that the spreading symbols in a DSSS signal do not need to be rectangular (i.e., a constant amplitude over the chip period), as shown in Figure 2.21. In principle, any shape could be used and different shapes can be used for different chips. Henceforth, we will denote DSSS signals generated using BPSK signaling with rectangular chips as BPSK-R signals. Several variations of the basic DSSS signal that employ nonrectangular symbols are used for satellite navigation applications. Binary offset carrier (BOC) signals [24] are generated using DSSS techniques, but employ portions of a square wave for the spreading symbols. A generalized treatment of the use of arbitrary binary patterns to generate each spreading symbol is provided in [25]. Spreading symbol shapes, such as raised cosines, whose amplitudes vary over a wide range of values are used extensively in digital communications. These shapes have also been considered for satellite navigation, but to date have not been used for practical reasons. For precise ranging, it is necessary for the satellite and user equipment to be able to faithfully reproduce the spreading waveform, which is facilitated through the use of signals that can be generated using simple digital means. Furthermore, spectral efficiency, which has motivated extensive studies in symbol shaping for communications applications, is generally not a concern for satellite navigation and can be detrimental for precise ranging.

In addition, DSSS signals with a constant envelope (i.e., those that have constant power over time) can be efficiently transmitted using switching-class amplifiers, although there are ways to combine multiple waveforms, not binary-valued, into a constant-envelope signal.

2.4.2.4 Pilot Components

A feature of many modern GNSS signals is that they split the total power in one overall signal between two components that are referred to as the data and pilot (or dataless) components. As the names suggest, the data component is modulated by navigation data and the pilot component is not modulated by the navigation data. Both components are modulated by spreading waveforms and utilize different ranging codes. Typical splits of power when separate data and pilot components are utilized range from 50%-50% (i.e., equal power in each component) to 25%-75% (i.e., power in the pilot component is three times that in the data component). Why are pilot components utilized? The reason is that a receiver can much more robustly track a signal that is not modulated by navigation data, as will be discussed in Chapter 8. Thus, pilot components can allow GNSS signals to be tracked in more challenging environments (e.g., deeper indoors, or in the presence of greater levels of interference) than would be possible without this design feature.

2.4.3 Secondary Codes

Many modern GNSS signals employ both primary ranging codes (discussed in Section 2.4.2.2) and secondary (or synchronization) codes. Secondary codes reduce interference between GNSS signals and also facilitate robust data bit synchronization within GNSS receivers.

A secondary code is a periodic, binary sequence that is generated at the primary code repetition rate. Each bit of the secondary code is modulo-2 summed to one entire period of the primary code. The GNSS constellations described in Chapters 3 through 7 use secondary codes of lengths from 4 to 1,800 for various signals.

To illustrate the concept of a secondary code, consider a hypothetical GNSS signal that uses a primary ranging code that is 1,023 chips in length, with the first 10 chip values of [1 0 0 1 1 0 1 0 1 0]. If a 4-bit secondary code of [1 0 1 0] is applied at the primary ranging code repetition rate (equal to 1/1,023 of the primary code chipping rate), every four repetitions of the primary ranging code would be modified as follows. For the first and third repetitions, the primary ranging code would be inverted, so that it started with the ten chips [0 1 1 0 0 1 0 1 0 1]. For the second and fourth repetitions, the primary ranging code would be unchanged, and this entire pattern would repeat again after the fourth ranging code repetition.

2.4.4 Multiplexing Techniques

In satellite navigation applications, it is frequently required to broadcast multiple signals from a satellite constellation, from a single satellite, and even upon a single carrier frequency. There are a number of techniques to facilitate this sharing of a common transmission channel without the broadcast signals interfering with each other. The use of different carrier frequencies to transmit multiple signals is referred

to as frequency division multiple access (FDMA) or frequency division multiplexing (FDM). Sharing a transmitter over time among two or more signals is referred to as time division multiple access (TDMA) or time division multiplexing (TDM). CDMA, or the use of different spreading codes to allow the sharing of a common carrier frequency, was introduced in Section 2.4.2.2.

When a common transmitter is used to broadcast multiple signals on a single carrier, it is desirable to combine these signals in a manner that forms a composite signal with a constant envelope for the reason discussed in Section 2.4.2.3. Two binary DSSS signals may be combined using quadrature phase shift keying (QPSK). In QPSK, the two signals are generated using RF carriers that are in phase quadrature (i.e., they have a relative phase difference of 90° such as cosine and sine functions of the same time parameter and are simply added together). The two constituents of a QPSK signal are referred to as the in-phase and quadraphase components.

When it is desired to combine more than two signals on a common carrier, more complicated multiplexing techniques are required. Interplexing combines three binary DSSS signals on a common carrier while retaining constant envelope [26]. To accomplish this feat, a fourth signal that is completely determined by the three desired signals, is also transmitted. The overall transmitted signal may be expressed as in the form of a QPSK signal:

$$s(t) = s_I(t)\cos(2\pi f_c t) - s_Q(t)\sin(2\pi f_c t) \tag{2.18}$$

with in-phase and quadraphase components, $s_I(t)$ and $s_Q(t)$, respectively, as:

$$\begin{aligned} s_I(t) &= \sqrt{2P_I}\,s_1(t)\cos(m) - \sqrt{2P_Q}\,s_2(t)\sin(m) \\ s_Q(t) &= \sqrt{2P_Q}\,s_3(t)\cos(m) + \sqrt{2P_I}\,s_1(t)s_2(t)s_3(t)\sin(m) \end{aligned} \tag{2.19}$$

where $s_1(t)$, $s_2(t)$, and $s_3(t)$, are the three desired signals, f_c is the carrier frequency and m is an index that is set in conjunction with the power parameters P_I and P_Q to achieve the desired power levels for the four multiplexed (three desired plus one additional) signals.

Other techniques for multiplexing more than two binary DSSS signals while retaining constant envelope include majority vote [27] and intervoting [28]. In the majority vote, an odd number of DSSS signals are combined by taking the majority of their underlying PRN sequence values at every instant in time to generate a composite DSSS signal. Intervoting consists of the simultaneous application of interplexing and majority vote.

2.4.5 Signal Models and Characteristics

In addition to the general quadrature signal representation in (2.18) for GNSS signals, we will find it occasionally convenient to use the complex-envelope or lowpass representation, $s_l(t)$, defined by the relation:

$$s(t) = \mathrm{Re}\left\{s_l(t)e^{j2\pi f_c t}\right\} \tag{2.20}$$

where Re{·} denotes the real part of. The in-phase and quadraphase components of the real signal $s(t)$ are related to its complex envelope by:

$$s_l(t) = s_I(t) + js_Q(t) \tag{2.21}$$

Two signal characteristics of great importance for satellite navigation applications are the autocorrelation function and power spectral density. The autocorrelation function for a lowpass signal with constant power is defined as:

$$R(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} s_l^*(t) s_l(t+\tau) dt \tag{2.22}$$

where * denotes complex conjugation. The power spectral density is defined to be the Fourier transform of the autocorrelation function:

$$S(f) = \int_{-\infty}^{\infty} R(\tau) e^{-j2\pi f\tau} dt \tag{2.23}$$

The power spectral density describes the distribution of power within the signal with regards to frequency.

It is often convenient to model some portions of a DSSS signal as being random. For instance, the data symbols and ranging code are often modeled as nonrepeating coin-flip sequences (i.e., they randomly assume values of either +1 or –1 with each outcome occurring with equal probability and with each value being independent of other values). The autocorrelation function for a DSSS signal with random components is generally taken to be the average or expected value of (2.22). The power spectral density remains as defined by (2.23).

As an example, consider a baseband DSSS signal without data employing rectangular chips with a perfectly random binary code as shown in Figure 2.22(a). The autocorrelation function illustrated in Figure 2.22(b) is described in equation form as [29]:

$$\begin{aligned} R(\tau) &= A^2\left(1 - \frac{|\tau|}{T_c}\right) && \text{for } |\tau| \le T_c \\ &= 0 && \text{elsewhere} \end{aligned} \tag{2.24}$$

The power spectrum of this signal shown in Figure 2.4(c) (as a function of angular frequency $\omega = 2\pi f$) may be determined using (2.20) to be:

$$S(f) = A^2 T_c \operatorname{sinc}^2(\pi f T_c) \tag{2.25}$$

where $\operatorname{sinc}(x) = \dfrac{\sin x}{x}$. What is important about a DSSS signal using a random binary code is that it correlates with itself in one and only one place and it is uncorrelated with any other random binary code. Satellite navigation systems employing rectangular chips have similar autocorrelation and power spectrum properties to those described above for the random binary code case, but employ ranging codes that

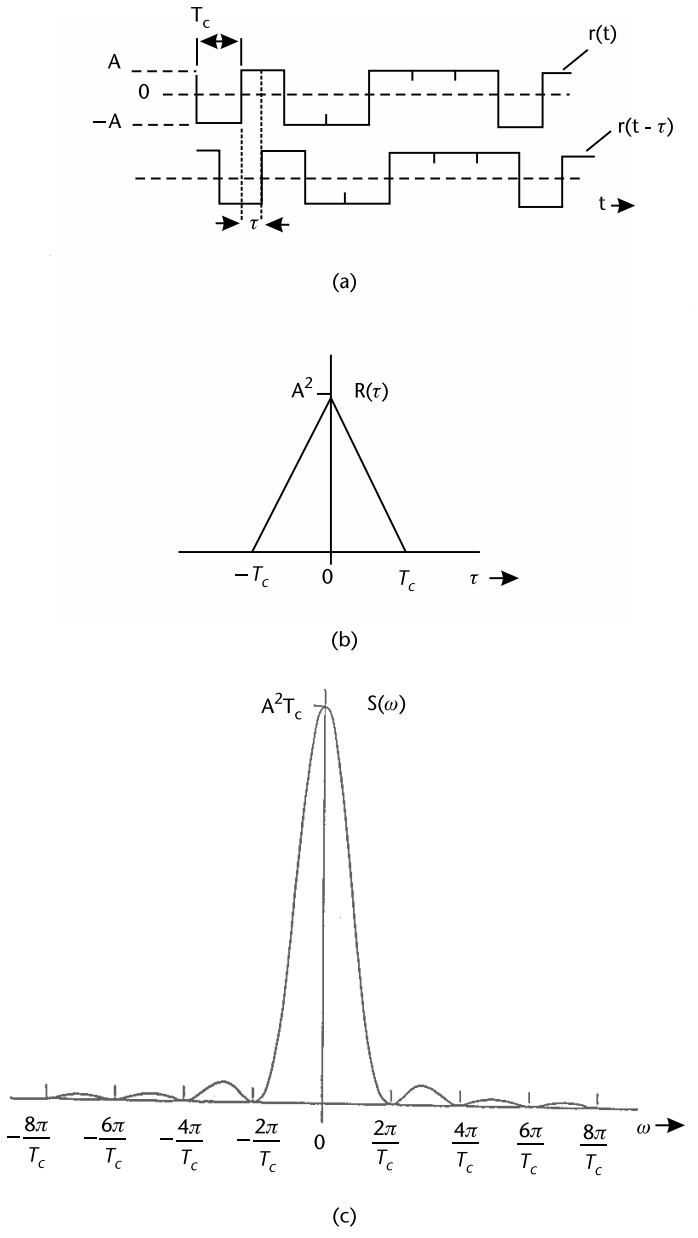

Figure 2.22 (a) A random binary code producing (b) the autocorrelation function and (c) the power spectrum of a DSSS signal.

are perfectly predictable and reproducible. This is why they are called "pseudo" random codes.

To illustrate the effects of finite-length ranging codes, consider a DSSS signal without data employing a pseudorandom sequence that repeats every N bits. Further, let us assume that this sequence is generated using a linear feedback shift register that is of maximum length. A linear feedback shift register is a simple digital circuit that consists of n bits of memory and some feedback logic [23], all clocked at a certain rate. Every clock cycle, the nth bit value is output from the device, the

logical value in bit 1 is moved to bit 2, the value in bit 2 to bit 3, and so on, and finally, a linear function is applied to the prior values of bits 1 to n to create a new input value into bit 1 of the device. With an n-bit linear feedback shift register, the longest length sequence that can be produced before the output repeats is $N = 2^n - 1$. A linear feedback shift register that produces a sequence of this length is referred to as maximum-length. During each period, the n bits within the register pass through all 2^n possible states, except the all-zeros state, since all zeros would result in a constant output value of 0. Because the number of negative values (1s) is always one larger than the number of positive values (0s) in a maximum-length sequence, the autocorrelation function of the spreading waveform $PN(t)$ outside of the correlation interval is $-A^2/N$. Recall that the correlation was 0 (uncorrelated) in this interval for the DSSS signal with random code in the previous example. The autocorrelation function for a maximum length pseudorandom sequence is the infinite series of triangular functions with period NT_c(seconds) shown in Figure 2.23(a). The negative correlation amplitude ($-A^2/N$) is shown in Figure 2.23(a) when the time shift, τ, is greater than $\pm T_c$, or multiples of $\pm T_c(N\pm1)$, and represents a DC term in the series. Expressing the equation for the periodic autocorrelation function mathematically [30] requires the use of the unit impulse function shifted in time by discrete (m) increments of the PRN sequence period NT_c: $\delta(\tau + mNT_c)$. Simply stated, this notation (also called a Dirac delta function) represents a unit impulse with a discrete phase shift of mNT_c seconds. Using this notation, the autocorrelation function can be expressed as the sum of the DC term and an infinite series of the triangle function, $R(\tau)$, defined by (2.24). The infinite series of the triangle function is obtained by the convolution (denoted by $\otimes$)of $R(\tau)$ with an infinite series of the phase shifted unit impulse functions as follows:

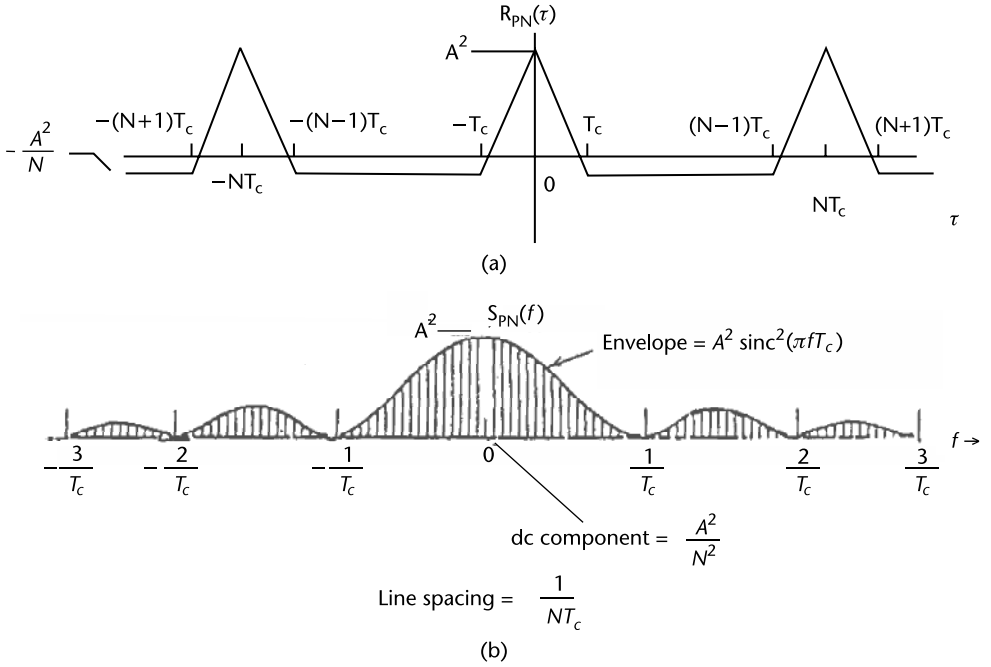

Figure 2.23 (a) The autocorrelation function of a DSSS signal generated from a maximum length pseudorandom sequence and (b) its line spectrum.

$$R_{PN}(\tau) = \frac{-A^2}{N} + \frac{N+1}{N} R(\tau) \otimes \sum_{m=-\infty}^{\infty} \delta(\tau + mNT_c) \tag{2.26}$$

The power spectrum of the DSSS signal generated from a maximum length pseudorandom sequence is derived from the Fourier transform of (2.26) and is the line spectrum shown in Figure 2.23(b). The unit impulse function is also required to express this in equation form as follows:

$$S_{PN}(f) = \frac{A^2}{N^2}\left(\delta(f) + \sum_{m=-\infty \neq 0}^{\infty} (N+1)\operatorname{sinc}^2\left(\frac{m\pi}{N}\right)\delta\left(2\pi f + \frac{m2\pi}{NT_c}\right)\right) \tag{2.27}$$

where $m = \pm 1, \pm 2, \pm 3, \ldots$

Observe in Figure 2.23(b) that the envelope of the line spectrum is the same as the continuous power spectrum obtained for the random code except for the small DC term in the line spectrum and the scale factor T_c. As the period, N (chips), of the maximum length sequence increases then the line spacing, $2\pi/NT_c$ (radians/s) or $1/NT_c$ (Hz), of the line spectrum decreases proportionally, so that the power spectrum begins to approach a continuous spectrum.

Next consider the general baseband DSSS signal that uses the arbitrary symbol $g(t)$:

$$s(t) = \sum_{k=-\infty}^{\infty} a_k g(t - kT_c) \tag{2.28}$$

If the ranging code values $\{a_k\}$ are assumed to be generated as a random coin-flip sequence, then the autocorrelation function for this signal may be found by taking the mean value of (2.22) resulting in:

$$R(\tau) = \int_{-\infty}^{\infty} g(t) g^*(t - \tau) dt \tag{2.29}$$

Although data was neglected in (2.28), its introduction does not change the result for a nonrepeating coin-flip sequence. Using this result, along with (2.23) for power spectral density, we can express the autocorrelation function and power spectrum for unit-power BPSK-R signals, for which

$$g_{BPSK-R}(t) = \begin{cases} 1/\sqrt{T_c}, & 0 \le t < T_c \\ 0, & \text{elsewhere} \end{cases} \tag{2.30}$$

as

$$\begin{aligned} R_{BPSK-R}(\tau) &= \begin{cases} 1 - |\tau|/T_c, & |\tau| < T_c \\ 0, & \text{elsewhere} \end{cases} \\ S_{BPSK-R}(f) &= T_c \operatorname{sinc}^2(\pi f T_c) \end{aligned} \tag{2.31}$$

The notation BPSK-R(n) is often used to denote a BPSK-R signal with $n \times 1.023$ MHz chipping rate. As will be discussed in Chapters 3, 5, 6, and 7, GPS, Galileo, BeiDou, and various regional systems employ frequencies that are multiples of 1.023 MHz. GPS was the first to use chipping rates that are integer multiples of 1.023 MHz (based upon a design choice to use a length-1,023 ranging code for one of the original GPS navigation signals and the desire for the repetition period to be a convenient value of 1 ms). Other systems subsequently adopted chipping rates that are integer multiples of 1.023 MHz to be interoperable with GPS.

A BOC signal may be viewed as being the product of a BPSK-R signal with a square-wave subcarrier. The autocorrelation and power spectrum are dependent on both the chip rate and characteristics of the square wave subcarrier. The number of square wave half-periods in a spreading symbol is typically selected to be an integer:

$$k = \frac{T_c}{T_s} \tag{2.32}$$

where $T_s = 1/(2f_s)$ is the half-period of a square wave generated with frequency f_s. When k is even, a BOC spreading symbol can be described as:

$$g_{BOC}(t) = g_{BPSK-R}(t)\,\mathrm{sgn}\left[\sin\left(\pi t / T_s + \psi\right)\right] \tag{2.33}$$

where sgn is the signum function (1 if the argument is positive, –1 if the argument is negative) and ψ is a selectable phase angle. When k is odd, a BOC signal may be viewed as using two symbols over every two consecutive chip periods, that given in (2.33) for the first spreading symbol in every pair and its inverse for the second. Two common values of ψ are 0° or 90°, for which the resultant BOC signals are referred to as sine-phased or cosine-phased, respectively.

With a perfect coin-flip spreading sequence, the autocorrelation functions for cosine- and sine-phased BOC signals resemble saw teeth, piece-wise linear functions between the peak values as shown in Table 2.3. The expression for the autocorrelation function applies for k odd and k even when a random code is assumed. The notation BOC(m,n) used in the table is shorthand for a BOC modulation generated using a $m \times 1.023$ MHz square wave frequency and a $n \times 1.023$ MHz chipping rate. The subscripts s and c refer to sine-phased and cosine-phased, respectively.

The power spectral density for a sine-phased BOC modulation is [24]:

Table 2.3 Autocorrelation Function Characteristics for BOC Modulations

	Number of Positive and Negative Peaks in Autocorrelation		*Autocorrelation Function Values for Peak at $\tau = jT_S/2$*	
Modulation	*Function*	*Delay Values of Peaks (s)*	*j even*	*j odd*
$BOC_s(m, n)$	$2k - 1$	$\tau = jT_S/2$, $-2k + 2 \le j \le 2k - 2$	$(-1)^{j/2}(k-\lvert j/2\rvert)/k$	$(-1)^{(\lvert j\rvert-1)/2}/(2k)$
$BOC_c(m, n)$	$2k + 1$	$\tau = jT_S/2$, $-2k + 1 \le j \le 2k - 1$	$(-1)^{j/2}(k-\lvert j/2\rvert)/k$	$(-1)^{(\lvert j\rvert+1)/2}/(2k)$

$$S_{BOC_s}(f) = \begin{cases} T_c \text{sinc}^2(\pi f T_c) \tan^2\left(\dfrac{\pi f}{2f_s}\right) & , \ k \text{ even} \\ T_c \dfrac{\cos^2(\pi f T_c)}{(\pi f T_c)^2} \tan^2\left(\dfrac{\pi f}{2f_s}\right) & , \ k \text{ odd} \end{cases} \tag{2.34}$$

and the power spectral density for a cosine-phased BOC modulation is:

$$S_{BOC_c(m,n)}(f) = \begin{cases} 4T_c \text{sinc}^2(\pi f T_c) \left(\dfrac{\sin^2\left(\dfrac{\pi f}{4f_s}\right)}{\cos\left(\dfrac{\pi f}{2f_s}\right)} \right)^2 & , k \text{ even} \\ 4T_c \dfrac{\cos^2(\pi f T_c)}{(\pi f T_c)^2} \left(\dfrac{\sin^2\left(\dfrac{\pi f}{4f_s}\right)}{\cos\left(\dfrac{\pi f}{2f_s}\right)} \right)^2 & , k \text{ odd} \end{cases} \tag{2.35}$$

A binary coded symbol (BCS) modulation [25] uses a spreading symbol defined by an arbitrary bit pattern $\{c_m\}$ of length M as:

$$g_{BCS}(t) = \sum_{m=0}^{M-1} c_m p_{T_c/M}(t - mT_c / M) \tag{2.36}$$

where $p_{T_c/M}(t)$ is a pulse taking on the value $1/\sqrt{T_c}$ over the interval $[0, T_c/M)$ and zero elsewhere. The notation BCS($[c_0, c_1, \ldots, c_{M-1}]$, n) is used to denote a BCS modulation that uses the sequence ($[c_0, c_1, \ldots, c_{M-1}]$ for each symbol and a chipping rate of $R_c = n \times 1.023$ MHz $= 1/T_c$. As shown in [25], the autocorrelation function for a BCS($[c_0, c_1, \ldots, c_{K-1}]$, n) modulation with perfect spreading code is a piecewise linear function between the values:

$$R_{BCS}(n T_c / M) = \frac{1}{M} \sum_{m=0}^{M-1} c_m c_{m-n} \tag{2.37}$$

where n is an integer with magnitude less than or equal to M and where it is understood that $c_m = 0$ for $m \notin [0, M-1]$. The power spectral density is:

$$S_{BCS}(f) = T_c \left| \frac{1}{M} \sum_{m=0}^{M-1} c_m e^{-j2\pi m f T_c / M} \right|^2 \frac{\sin^2(\pi f T_c / M)}{(\pi f T_c / M)^2} \tag{2.38}$$

Given the success of BPSK-R modulations, why consider more advanced modulations like BOC or BCS? Compared to BPSK-R modulations, which only allow the signal designer to select carrier frequency and chip rate, BOC and BCS modulations provide additional design parameters for waveform designers to use. The resulting

modulation designs can provide enhanced performance when bandwidth is limited (due to implementation constraints at transmitter and receiver or due to spectrum allocations). Also, modulations can be designed to better share limited frequency bands available for use by multiple GNSS constellations. The spectra can be shaped in order to limit interference and otherwise spectrally separate different signals. To obtain adequate performance, such modulation design activities must carefully consider a variety of signal characteristics in the time and frequency domains and should not concentrate exclusively on spectrum shape.

2.5 Positioning Determination Using Ranging Codes

As mentioned in Section 2.4, GNSS satellite transmissions utilize DSSS modulation. DSSS provides the structure for the transmission of ranging codes and essential navigation data such as satellite ephemerides and satellite health. The ranging codes modulate the satellite carrier frequencies. These codes look like and have spectral properties similar to random binary sequences but are actually deterministic. A simple example of a short ranging code sequence is shown in Figure 2.24. These codes have a predictable pattern, which is periodic and can be replicated by a suitably equipped receiver.

2.5.1 Determining Satellite-to-User Range

Earlier, we examined the theoretical aspects of using satellite ranging codes and multiple spheres to solve for user position in three dimensions. That example was predicated on the assumption that the receiver clock was perfectly synchronized to system time. In actuality, this is generally not the case. Prior to solving for the three-dimensional user position, we will examine the fundamental concepts involving satellite-to-user range determination with nonsynchronized clocks and ranging codes. There are a number of error sources that affect range measurement accuracy (e.g., measurement noise, propagation delays); however, these can generally be considered negligible when compared to the errors experienced from nonsynchronized clocks. Therefore, in our development of basic concepts, errors other than clock offset are omitted. Extensive treatment of these error sources is provided in Section 10.2.

In Figure 2.25, we wish to determine vector **u**, which represents a user receiver's position with respect to the ECEF coordinate system origin. The user's position coordinates x_u, y_u, z_u are considered unknown. Vector **r** represents the vector offset from the user to the satellite. The satellite is located at coordinates x_s, y_s, z_s within the ECEF Cartesian coordinate system. Vector **s** represents the position of the satellite relative to the coordinate origin. Vector **s** is computed using ephemeris data broadcast by the satellite. The satellite-to-user vector **r** is

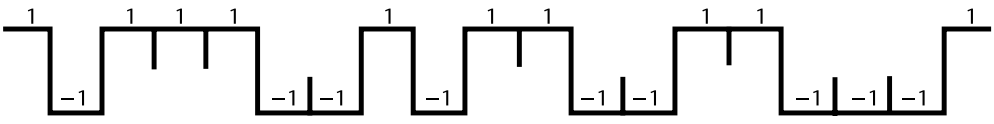

Figure 2.24 Ranging code.

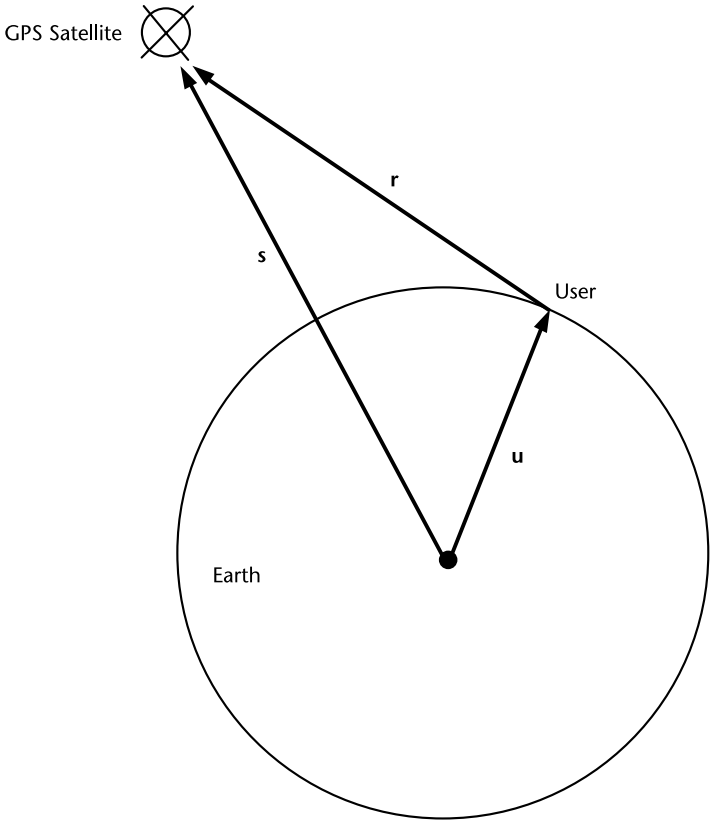

Figure 2.25 User position vector representation.

$$\mathbf{r} = \mathbf{s} - \mathbf{u} \tag{2.39}$$

The magnitude of vector **r** is

$$\|\mathbf{r}\| = \|\mathbf{s} - \mathbf{u}\| \tag{2.40}$$

Let r represent the magnitude of **r**

$$r = \|\mathbf{s} - \mathbf{u}\| \tag{2.41}$$

The distance r is computed by measuring the propagation time required for a satellite-generated ranging code to transit from the satellite to the user receiver antenna. The propagation time measurement process is illustrated in Figure 2.26. As an example, a specific code phase generated by the satellite at t_1 arrives at the receiver at t_2. The propagation time is represented by Δt. Within the receiver, an identical coded ranging code denoted as the replica code is generated at t, with respect to the receiver clock. This replica code is shifted in time until it achieves correlation with the satellite generated ranging code. If the satellite clock and the receiver clock were perfectly synchronized, the correlation process would yield the true propagation time. By multiplying this propagation time, Δt, by the speed of light, the true (i.e., geometric) satellite-to-user distance can be computed. We would

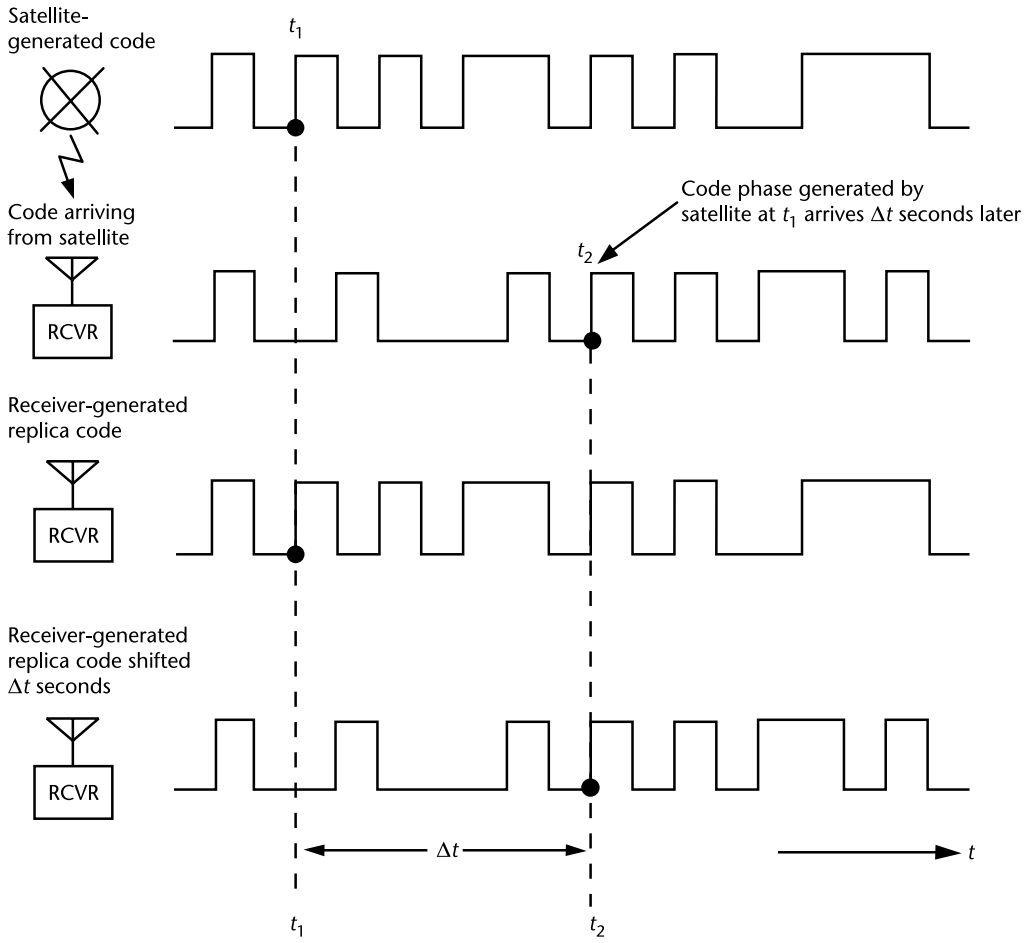

Figure 2.26 Use of replica code to determine satellite code transmission time.

then have the ideal case described in Section 2.1.2.1. However, the satellite and receiver clocks are generally not synchronized.

The receiver clock will generally have a bias error from system time. Further, the satellite timing system (usually referred to as the satellite clock) is based on a highly accurate free running atomic frequency standards (AFS) described in Section 2.7.1.5. Therefore, the satellite timing system is typically offset from system time. Thus, the range determined by the correlation process is denoted as the pseudorange ρ. The measurement is called pseudorange because it is the range determined by multiplying the signal propagation velocity, c, by the time difference between two nonsynchronized clocks (the satellite clock and the receiver clock). The measurement contains the geometric satellite-to-user range, an offset attributed to the difference between system time and the user clock, and an offset between system time and the satellite clock. The timing relationships are shown in Figure 2.27, where:

T_s = System time at which the signal left the satellite

T_u = System time at which the signal reached the user receiver

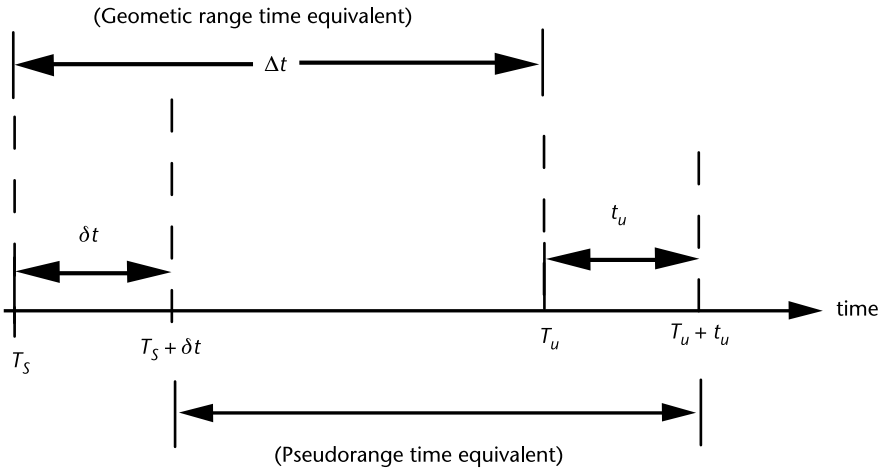

Figure 2.27 Range measurement timing relationships.

δt = Offset of the satellite clock from system time [advance is positive; retardation (delay) is negative]

t_u = Offset of the receiver clock from system time

$T_s + \delta t$ = Satellite clock reading at the time that the signal left the satellite

$T_u + t_u$ = User receiver clock reading at the time when the signal reached the user receiver

c = speed of light

$$\textit{Geometric range},\ r = c(T_u - T_s) = c\Delta t$$

$$\begin{aligned}\textit{Pseudorange},\ \rho &= c[(T_u + t_u) - (T_s + \delta t)] \\ &= c(T_u - T_s) + c(t_u - \delta t) \\ &= r + c(t_u - \delta t)\end{aligned}$$

Therefore, (2.39) can be rewritten as:

$$\rho - c(t_u - \delta t) = \|\mathbf{s} - \mathbf{u}\|$$

where t_u represents the advance of the receiver clock with respect to system time, δt represents the advance of the satellite clock with respect to system time, and c is the speed of light.

The satellite clock offset from system time, δt, is composed of bias and drift contributions. A SATNAV system ground monitoring network determines corrections for these offset contributions and transmits the corrections to the satellites for rebroadcast to the users in the navigation message. These corrections are applied within the user receiver to synchronize the transmission of each ranging code to system time. Therefore, we assume that this offset is compensated for and no

longer consider δt an unknown. (There is some residual offset, which is treated in Section 10.2.1, but in the context of this discussion we assume that this is negligible.) Hence, the preceding equation can be expressed as

$$\rho - ct_u = \|\mathbf{s} - \mathbf{u}\| \tag{2.42}$$

2.5.2 Calculation of User Position

In order to determine user position in three dimensions (x_u, y_u, z_u) and the offset t_u, pseudorange measurements are made to four satellites resulting in the system of equations

$$\rho_j = \|\mathbf{s}_j - \mathbf{u}\| + ct_u \tag{2.43}$$

where j ranges from 1 to 4 and references the satellites. Equation (2.43) can be expanded into the following set of equations in the unknowns x_u, y_u, z_u, and t_u:

$$\rho_1 = \sqrt{(x_1 - x_u)^2 + (y_1 - y_u)^2 + (z_1 - z_u)^2} + ct_u \tag{2.44}$$

$$\rho_2 = \sqrt{(x_2 - x_u)^2 + (y_2 - y_u)^2 + (z_2 - z_u)^2} + ct_u \tag{2.45}$$

$$\rho_3 = \sqrt{(x_3 - x_u)^2 + (y_3 - y_u)^2 + (z_3 - z_u)^2} + ct_u \tag{2.46}$$

$$\rho_4 = \sqrt{(x_4 - x_u)^2 + (y_4 - y_u)^2 + (z_4 - z_u)^2} + ct_u \tag{2.47}$$

where x_j, y_j, and z_j denote the jth satellite's position in three dimensions.

These nonlinear equations can be solved for the unknowns by employing closed form solutions [31–34], iterative techniques based on linearization, or Kalman filtering. (Kalman filtering provides a means for improving PVT estimates based on optimal processing of time sequence measurements and is described later. The following development regarding linearization is based on a similar development in [35].) If we know approximately where the receiver is, then we can denote the offset of the true position (x_u, y_u, z_u) from the approximate position $(\hat{x}_u, \hat{y}_u, \hat{z}_u)$ by a displacement (Δx_u, Δy_u, Δz_u). By expanding (2.44) to (2.47) in a Taylor series about the approximate position, we can obtain the position offset (Δx_u, Δy_u, Δz_u) as linear functions of the known coordinates and pseudorange measurements. This process is described later.

Let a single pseudorange be represented by

$$\begin{aligned}\rho_j &= \sqrt{\left(x_j - x_u\right)^2 + \left(y_j - y_u\right)^2 + \left(z_j - z_u\right)^2} + ct_u \\ &= f\left(x_u, y_u, z_u, t_u\right)\end{aligned} \tag{2.48}$$

Using the approximate position location $(\hat{x}_u, \hat{y}_u, \hat{z}_u)$ and time bias estimate $\hat{t}_u$, an approximate pseudorange can be calculated:

$$\begin{aligned}\hat{\rho}_j &= \sqrt{\left(x_j - \hat{x}_u\right)^2 + \left(y_j - \hat{y}_u\right)^2 + \left(z_j - \hat{z}_u\right)^2} + c\hat{t}_u \\ &= f\left(\hat{x}_u, \hat{y}_u, \hat{z}_u, \hat{t}_u\right)\end{aligned} \tag{2.49}$$

As stated above, the unknown user position and receiver clock offset is considered to consist of an approximate component and an incremental component:

$$\begin{aligned}x_u &= \hat{x}_u + \Delta x_u \\ y_u &= \hat{y}_u + \Delta y_u \\ z_u &= \hat{z}_u + \Delta z_u \\ t_u &= \hat{t}_u + \Delta t_u\end{aligned} \tag{2.50}$$

Therefore, we can write

$$f\ (x_u,\ y_u,\ z_u,\ t_u)\ =\ f(\hat{x}_u + \Delta x_u, \hat{y}_u + \Delta y_u, \hat{z}_u + \Delta z_u, \hat{t}_u + \Delta t_u)$$

This latter function can be expanded about the approximate point and associated predicted receiver clock offset $\left(\hat{x}_u, \hat{y}_u, \hat{z}_u, \hat{t}_u\right)$ using a Taylor series:

$$\begin{aligned}f(\hat{x}_u + \Delta x_u, \hat{y}_u + \Delta y_u, \hat{z}_u + \Delta z_u, \hat{t}_u + \Delta t_u) = f(\hat{x}_u, \hat{y}_u, \hat{z}_u, \hat{t}_u) + \frac{\partial f\left(\hat{x}_u, \hat{y}_u, \hat{z}_u, \hat{t}_u\right)}{\partial \hat{x}_u}\ \Delta x_u + \\ \frac{\partial f\left(\hat{x}_u, \hat{y}_u, \hat{z}_u, \hat{t}_u\right)}{\partial \hat{y}_u}\Delta y_u + \frac{\partial f\left(\hat{x}_u, \hat{y}_u, \hat{z}_u, \hat{t}_u\right)}{\partial \hat{z}_u}\Delta z_u + \frac{\partial f\left(\hat{x}_u, \hat{y}_u, \hat{z}_u, \hat{t}_u\right)}{\partial \hat{t}_u}\Delta t_u + \ldots\end{aligned} \tag{2.51}$$

The expansion has been truncated after the first-order partial derivatives to eliminate nonlinear terms. The partial derivatives evaluate as follows:

$$\begin{aligned}\frac{\partial f\left(\hat{x}_u, \hat{y}_u, \hat{z}_u, \hat{t}_u\right)}{\partial \hat{x}_u} &= -\frac{x_j - \hat{x}_u}{\hat{r}_j} \\ \frac{\partial f\left(\hat{x}_u, \hat{y}_u, \hat{z}_u, \hat{t}_u\right)}{\partial \hat{y}_u} &= -\frac{y_j - \hat{y}_u}{\hat{r}_j} \\ \frac{\partial f\left(\hat{x}_u, \hat{y}_u, \hat{z}_u, \hat{t}_u\right)}{\partial \hat{z}_u} &= -\frac{z_j - \hat{z}_u}{\hat{r}_j} \\ \frac{\partial f\left(\hat{x}_u, \hat{y}_u, \hat{z}_u, \hat{t}_u\right)}{\partial \hat{t}_u} &= c\end{aligned} \tag{2.52}$$

where

$$\hat{r}_j = \sqrt{\left(x_j - \hat{x}_u\right)^2 + \left(y_j - \hat{y}_u\right)^2 + \left(z_j - \hat{z}_u\right)^2}$$

Substituting (2.49) and (2.52) into (2.51) yields

$$\rho_j = \hat{\rho}_j - \frac{x_j - \hat{x}_u}{\hat{r}_j}\Delta x_u - \frac{y_j - \hat{y}_u}{\hat{r}_j}\Delta y_u - \frac{z_j - \hat{z}_u}{\hat{r}_j}\Delta z_u + ct_u \tag{2.53}$$

We have now completed the linearization of (2.48) with respect to the unknowns Δx_u, Δy_u, Δz_u, and Δt_u. (It is important to remember that we are neglecting secondary error sources such as Earth rotation compensation, measurement noise, propagation delays, and relativistic effects, which are treated in detail in Section 10.2.)

Rearranging the above expression with the known quantities on the left and unknowns on right yields

$$\hat{\rho}_j - \rho_j = \frac{x_j - \hat{x}_u}{\hat{r}_j}\Delta x_u + \frac{y_j - \hat{y}_u}{\hat{r}_j}\Delta y_u + \frac{z_j - \hat{z}_u}{\hat{r}_j}\Delta z_u - ct_u \tag{2.54}$$

For convenience, we will simplify the above equation by introducing new variables where

$$\begin{aligned}
\Delta\rho &= \hat{\rho}_j - \rho_j \\
a_{xj} &= \frac{x_j - \hat{x}_u}{\hat{r}_j} \\
a_{yj} &= \frac{y_j - \hat{y}_u}{\hat{r}_j} \\
a_{zj} &= \frac{z_j - \hat{z}_u}{\hat{r}_j}
\end{aligned} \tag{2.55}$$

The a_{xj}, a_{yj}, and a_{zj} terms in (2.55) denote the direction cosines of the unit vector pointing from the approximate user position to the jth satellite. For the jth satellite, this unit vector is defined as

$$\mathbf{a}_j = (a_{xj},\ a_{yj},\ a_{zj})$$

Equation (2.54) can be rewritten more simply as

$$\Delta\rho_j = a_{xj}\Delta x_u + a_{yj}\Delta y_u + a_{zj}\Delta z_u - c\Delta t_u$$

We now have four unknowns: Δx_u, Δy_u, Δz_u, and Δt_u, which can be solved for by making ranging measurements to four satellites. The unknown quantities can be determined by solving the set of linear equations next:

$$\begin{aligned}
\Delta\rho_1 &= a_{x1}\Delta x_u + a_{y1}\Delta y_u + a_{z1}\Delta z_u - c\Delta t_u \\
\Delta\rho_2 &= a_{x2}\Delta x_u + a_{y2}\Delta y_u + a_{z2}\Delta z_u - c\Delta t_u \\
\Delta\rho_3 &= a_{x3}\Delta x_u + a_{y3}\Delta y_u + a_{z3}\Delta z_u - c\Delta t_u \\
\Delta\rho_4 &= a_{x4}\Delta x_u + a_{y4}\Delta y_u + a_{z4}\Delta z_u - c\Delta t_u
\end{aligned} \tag{2.56}$$

These equations can be put in matrix form by making the definitions

$$\Delta\rho = \begin{bmatrix} \Delta\rho_1 \\ \Delta\rho_2 \\ \Delta\rho_3 \\ \Delta\rho_4 \end{bmatrix} \quad \mathbf{H} = \begin{bmatrix} a_{x1} & a_{y1} & a_{z1} & 1 \\ a_{x2} & a_{y2} & a_{z2} & 1 \\ a_{x3} & a_{y3} & a_{z3} & 1 \\ a_{x4} & a_{y4} & a_{z4} & 1 \end{bmatrix} \quad \Delta\mathbf{x} = \begin{bmatrix} \Delta x_u \\ \Delta y_u \\ \Delta z_u \\ -c\Delta t_u \end{bmatrix}$$

One obtains, finally,

$$\Delta\rho = \mathbf{H}\Delta\mathbf{x} \tag{2.57}$$

which has the solution

$$\Delta\mathbf{x} = \mathbf{H}^{-1}\Delta\rho \tag{2.58}$$

Once the unknowns are computed, the user's coordinates x_u, y_u, z_u and the receiver clock offset t_u are then calculated using (2.50). This linearization scheme will work well as long as the displacement (Δx_u, Δy_u, Δz_u) is within close proximity of the linearization point. The acceptable displacement is dictated by the user's accuracy requirements. If the displacement does exceed the acceptable value, the above process is reiterated with $\hat{\rho}$ being replaced by a new estimate of pseudorange based on the calculated point coordinates x_u, y_u, and z_u. In actuality, the true user-to-satellite measurements are corrupted by uncommon (i.e., independent) errors such as measurement noise, deviation of the satellite path from the reported ephemeris, and multipath. These errors translate to errors in the components of vector $\Delta\mathbf{x}$, as shown here:

$$\boldsymbol{\epsilon}_x = \mathbf{H}^{-1}\,\boldsymbol{\epsilon}_{meas} \tag{2.59}$$

where $\boldsymbol{\epsilon}_{meas}$ is the vector containing the pseudorange measurement errors and $\boldsymbol{\epsilon}_x$ is the vector representing errors in the user position and receiver clock offset.

The error contribution $\boldsymbol{\epsilon}_x$ can be minimized by making measurements to more than four satellites, which will result in an overdetermined solution set of equations similar to (2.57). Each of these redundant measurements will generally contain independent error contributions. Redundant measurements can be processed by least

squares estimation techniques that obtain improved estimates of the unknowns. Various versions of this technique exist and are usually employed in today's receivers, which generally employ more than four user to-satellite measurements to compute user position, velocity, and time (PVT). Appendix A provides an introduction to least squares techniques.

2.6 Obtaining User Velocity

GNSS provides the capability for determining three-dimensional user velocity, which is denoted $\dot{\mathbf{u}}$. Velocity can be estimated by forming an approximate derivative of the user position, as shown here:

$$\dot{\mathbf{u}} = \frac{d\mathbf{u}}{dt} = \frac{\mathbf{u}(t_2) - \mathbf{u}(t_1)}{t_2 - t_1}$$

This approach can be satisfactory provided the user's velocity is nearly constant over the selected time interval (i.e., not subjected to acceleration or jerk) and if the errors in the positions $\mathbf{u}(t_2)$ and $\mathbf{u}(t_1)$ are small relative to difference $\mathbf{u}(t_2) - \mathbf{u}(t_1)$.

In most GNSS receivers, velocity measurements are made by processing carrier-phase measurements, which enable precise estimation of the Doppler frequency of the received satellite signals. The Doppler shift is produced by the relative motion of the satellite with respect to the user. The satellite velocity vector $\mathbf{v}$ is computed using ephemeris information and an orbital model that resides within the receiver. Figure 2.28 is a curve of received Doppler frequency as a function of time measured by a user at rest on the surface of the Earth from a GNSS satellite. The received frequency increases as the satellite approaches the receiver and decreases as it recedes from the user. The reversal in the curve represents the time when the Doppler shift is zero and occurs when the satellite is at its closest position relative to the user. At

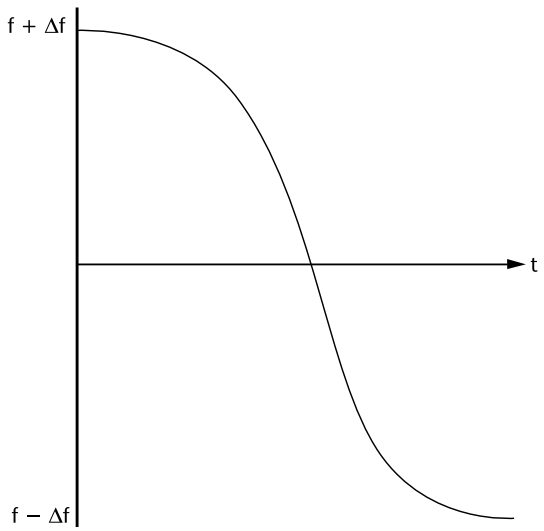

Figure 2.28 Received Doppler frequency by user at rest on Earth's surface.

this point, the radial component of the velocity of the satellite relative to the user is zero. As the satellite passes through this point, the sign of Δf changes. At the receiver antenna, the received frequency, f_R, can be approximated by the classical Doppler equation as follows:

$$f_R = f_T\left(1-\frac{(\mathbf{v}_r\cdot\mathbf{a})}{c}\right) \tag{2.60}$$

where f_T is the transmitted satellite signal frequency, $\mathbf{v}_r$ is the satellite-to-user relative velocity vector, $\mathbf{a}$ is the unit vector pointing along the line of sight from the user to the satellite, and c is the speed of propagation. The dot product $\mathbf{v}_r \cdot \mathbf{a}$ represents the radial component of the relative velocity vector along the line of sight to the satellite. Vector $\mathbf{v}_r$ is given as the velocity difference

$$\mathbf{v}_r = \mathbf{v} - \dot{\mathbf{u}} \tag{2.61}$$

where $\mathbf{v}$ is the velocity of the satellite and $\dot{\mathbf{u}}$ is the velocity of the user, both referenced to a common ECEF frame. The Doppler offset due to the relative motion is obtained from these relations as

$$\Delta f = f_R - f_T = -f_T\frac{(\mathbf{v}-\dot{\mathbf{u}})\cdot\mathbf{a}}{c}$$

For example, at the GPS L1 frequency, 1,575.42 MHz, the maximum Doppler frequency for a stationary user on the Earth is approximately 4 kHz corresponding to a maximum line-of-sight velocity of approximately 800 m/s.

There are several approaches for obtaining user velocity from the received Doppler frequency. One technique is described herein. This technique assumes that the user position $\mathbf{u}$ has been determined and its displacement $(\Delta x_u, \Delta y_u, \Delta z_u)$ from the linearization point is within the user's requirements. In addition to computing the three-dimensional user velocity $\dot{\mathbf{u}} = (\dot{x}_u, \dot{y}_u, \dot{z}_u)$, this particular technique determines the receiver clock drift $\dot{t}_u$.

For the jth satellite, substituting (2.61) into (2.60) yields

$$f_{Rj} = f_{Tj}\left\{1-\frac{1}{c}\left[(\mathbf{v}_j-\dot{\mathbf{u}})\cdot\mathbf{a}_j\right]\right\} \tag{2.62}$$

The satellite transmitted frequency f_{Tj} is the actual transmitted satellite frequency.

As stated in Section 2.7.1.5, satellite frequency generation and timing is based on a highly accurate free-running AFS, which is typically offset from system time. Corrections are generated by the ground control/monitoring network periodically to correct for this offset. These corrections are available in the navigation message and are applied within the receiver to obtain the actual satellite transmitted frequency. Hence,

$$f_{Tj} = f_0 + \Delta f_{Tj} \tag{2.63}$$

where f_0 is the nominal transmitted satellite frequency (i.e., L1) and Δf_{Tj} is the correction determined from the navigation message update.

The measured estimate of the received signal frequency is denoted f_j for the signal from the jth satellite. These measured values are in error and differ from the f_{Rj} values by a frequency bias offset. This offset can be related to the drift rate $\dot{t}_u$ of the user clock relative to system time. The value $\dot{t}_u$ has the units seconds/second and essentially gives the rate at which the user's clock is running fast or slow relative to system time. The clock drift error, f_j, and f_{Rj} are related by the formula

$$f_{Rj} = f_j(1 + \dot{t}_u) \tag{2.64}$$

where $\dot{t}_u$ is considered positive if the user clock is running fast. Substitution of (2.64) into (2.62), after algebraic manipulation, yields

$$\frac{c(f_j - f_{Tj})}{f_{Tj}} + \mathbf{v}_j \cdot \mathbf{a}_j = \dot{\mathbf{u}} \cdot \mathbf{a}_j - \frac{cf_j\dot{t}_u}{f_{Tj}}$$

Expanding the dot products in terms of the vector components yields

$$\frac{c(f_j - f_{Tj})}{f_{Tj}} + v_{xj}a_{xj} + v_{yj}a_{yj} + v_{zj}a_{zj} = \dot{x}_u a_{xj} + \dot{y}_u a_{yj} + \dot{z}_u a_{zj} - \frac{cf_j\dot{t}_u}{f_{Tj}} \tag{2.65}$$

where $\mathbf{v}_j = (v_{xj}, v_{yj}, v_{zj})$, $\mathbf{a}_j = (a_{xj}, a_{yj}, a_{zj})$, and $\dot{\mathbf{u}} = (\dot{x}_u, \dot{y}_u, \dot{z}_u)$. All of the variables on the left side of (2.65) are either calculated or derived from measured values. The components of $\mathbf{a}_j$ are obtained during the solution for the user location (which is assumed to precede the velocity computation). The components of $\mathbf{v}_j$ are determined from the ephemeris data and the satellite orbital model. The f_{Tj} can be estimated using (2.63) and the frequency corrections derived from the navigation updates. (This correction, however, is usually negligible and f_{Tj} can normally be replaced by f_0.) The f_j can be expressed in terms of receiver measurements of delta range (see Chapter 8 for a more detailed description of receiver processing). To simplify the above equation, we introduce the new variable d_j, defined by

$$d_j = \frac{c(f_j - f_{Tj})}{f_{Tj}} + v_{xj}a_{xj} + v_{yj}a_{yj} + v_{zj}a_{zj} \tag{2.66}$$

The term f_j/f_{Tj} on the right side in (2.66) is numerically very close to 1, typically within several parts per million. Little error results by setting this ratio to 1. With these simplifications, (2.66) can be rewritten as

$$d_j = \dot{x}_u a_{xj} + \dot{y}_u a_{yj} + \dot{z}_u a_{zj} - c\dot{t}_u$$

We now have four unknowns: $\dot{\mathbf{u}} = \dot{x}_u, \dot{y}_u, \dot{z}_u, \dot{t}_u$ which can be solved by using measurements from four satellites. As before, we calculate the unknown quantities by solving the set of linear equations using matrix algebra. The matrix/vector scheme is

$$\mathbf{d} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix} \quad \mathbf{H} = \begin{bmatrix} a_{x1} & a_{y1} & a_{z1} & 1 \\ a_{x2} & a_{y2} & a_{z2} & 1 \\ a_{x3} & a_{y3} & a_{z3} & 1 \\ a_{x4} & a_{y4} & a_{z4} & 1 \end{bmatrix} \quad \mathbf{g} = \begin{bmatrix} \dot{x}_u \\ \dot{y}_u \\ \dot{z}_u \\ -c\dot{t}_u \end{bmatrix}$$

Note that $\mathbf{H}$ is identical to the matrix used in Section 2.5.2 in the formulation for the user position determination. In matrix notation,

$$\mathbf{d} = \mathbf{Hg}$$

and the solution for the velocity and time drift are obtained as

$$\mathbf{g} = \mathbf{H}^{-1}\mathbf{d}$$

The phase measurements that lead to the frequency estimates used in the velocity formulation are corrupted by errors such as measurement noise and multipath. Furthermore, the computation of user velocity is dependent on user position accuracy and correct knowledge of satellite ephemeris and satellite velocity. The relationship between the errors contributed by these parameters in the computation of user velocity is similar to (2.57). If measurements are made to more than four satellites, least squares estimation techniques can be employed to obtain improved estimates of the unknowns.

2.7 Frequency Sources, Time, and GNSS

Various types of frequency sources are used within GNSS. These range from low-cost quartz crystal oscillators within user equipment to highly accurate atomic frequency standards (AFSs) onboard the satellites as well as at various ground control segment components. Each individual SATNAV system time is based on an ensemble of some or all of these AFSs that are contained within that particular system. When combined with a time scale based on astronomical observations, a version of UTC is formed. Most civil and military applications use a version of UTC for their timekeeping needs.

2.7.1 Frequency Sources

2.7.1.1 Quartz Crystal Oscillators

The fundamental concept of a quartz crystal oscillator is that the crystal behaves like a tuned circuit due to its physical characteristics. This is depicted in Figure 2.29 where Branch 1 represents the crystal and C_0 represents the capacitance in

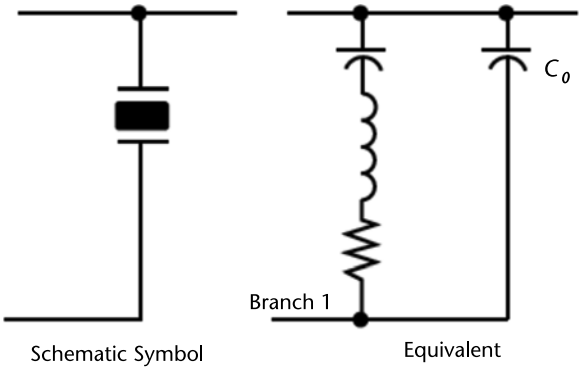

Figure 2.29 Crystal equivalent circuit. (From: [36].© Keysight Technologies, Inc. May 1997. Reproduced with permission, courtesy of Keysight Technologies.)

the wire leads and the crystal holder [36]. From [37], "a quartz crystal has piezoelectric characteristics. That is, the crystal strains (expands or contracts) when a voltage is applied. When the voltage is removed or reversed in polarity, the strain is reversed." When placed into a circuit shown in Figure 2.30 [36], the voltage from the crystal is amplified and then fed back to the crystal thus creating an oscillating circuit (i.e., oscillator). The oscillator resonance frequency is determined by the rate of crystal expansion and contraction. This resonance frequency is a function of the crystal physical characteristics. Note that the oscillator output frequency can be the fundamental crystal resonance frequency or at or near a harmonic of the fundamental frequency denoted as an overtone [36]. As stated in [36], the vibration setup in the quartz crystal may produce both harmonic and nonharmonic signals and overtones. The harmonic overtones are desirable since they allow the production of higher-frequency crystal resonators using essentially the same crystal cut. However, nonharmonic overtones are undesirable as they may lead to the generation of unwanted signals at frequencies spaced close to the one desired [36]. Most high-stability oscillators use either the third or fifth overtone frequency to achieve a high Q. (It is sometime difficult to tune the circuit with overtones higher than five.) The ratio of the resonance frequency to the bandwidth of which the circuit will oscillate is denoted as the quality factor, Q. A typical quartz oscillator Q ranges from 10^4 to

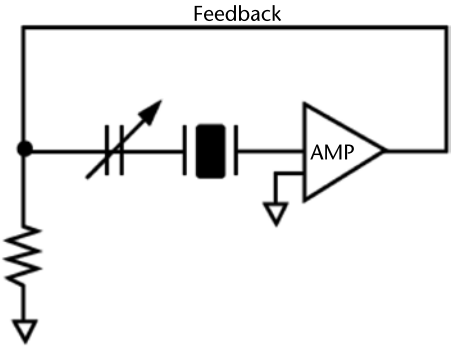

Figure 2.30 Simplified amplifier feedback (oscillator) circuit using a crystal resonator. (From: [36]. © Keysight Technologies, Inc. May 1997. Reproduced with permission, courtesy of Keysight Technologies.)

10^6, whereas for highly stable oscillators, the maximum $Q = 1.6 \times 10^7/f$, where f is the resonance frequency in megahertz [37].

All crystal oscillators undergo aging, which is a gradual change in frequency over many days or months. At a constant temperature, aging has an approximately logarithmic dependence on time. The aging rate is highest when it is first turned on. When the temperature is changed, a new aging cycle starts. The primary causes of aging are stress relief in the crystal's mounting structure, mass transfer to or from the crystal's surface due to adsorption or desorption of contamination, changes in the oscillator circuitry, and impurities and strains in the quartz material. Most manufacturers pre-age their crystals by placing their crystals in a high temperature oil bath for a number of days.

The frequency of a crystal is inversely proportional to its thickness. A typical 5-MHz crystal is on the order of 1 million atomic layers thick. The adsorption or desorption of contamination equivalent to the mass of one atomic layer of quartz changes the frequency by about 1 part per million (ppm). In order to achieve low aging, crystals must be hermetically sealed in an ultra-clean, high-vacuum environment. The aging rates of typical commercially available crystal oscillators range from 5 ppm to 10 ppm per year for an inexpensive XO (crystal oscillator) to 0.5 ppm per year for a temperature compensated crystal oscillator (TCXO) and to 0.05 ppm per year for an oven controlled crystal oscillator (OCXO). The highest precision OCXOs can age a few parts in 10^{12} per day or less than 0.01 ppm per year.

Causes of short-term instabilities include temperature fluctuations, Johnson noise in the crystal, random vibration, noise in the oscillator circuitry, and fluctuations at various resonator interfaces. Long-term performance is limited primarily by temperature sensitivity and aging. In a properly designed oscillator, the resonator is the primary noise source close to the carrier and the oscillator circuitry is the primary source far from the carrier. The noise close to the carrier has a strong inverse relationship to the resonator Q. Optimum low noise performance is only achievable in a vibration-free laboratory environment [38, 39].

The Allan variance, $\sigma_y(\tau)$, is the standard method for describing short-term stability of oscillators in the time domain. It is a measurement of the frequency jitter over short periods of time, normally from 1 microsecond to 1,000 seconds. Stability specifications for time periods greater than 1,000 seconds are usually considered long-term stability measurements. For the Allan variance method, fractional frequencies, $y = \Delta f/f$, are measured over a time interval, τ. The differences between successive pairs of measurements of y, $(y_{k+1} - y_k)$, are squared and one-half of the time average of their sum is calculated over the sampling period.

$$\sigma_y^2(\tau) = \frac{1}{2m}\sum_{j=1}^{m}\left(y_{k+1} - y_k\right)^2$$

The classical variance diverges for some commonly observed noise processes such as the random walk where the variance increases with an increasing number of data points. However, the Allan variance converges for all noise processes observed in precision oscillators. Figure 2.31 displays time-domain stability for a typical precision oscillator. For $\sigma_y(\tau)$ to properly measure random frequency fluctuations,

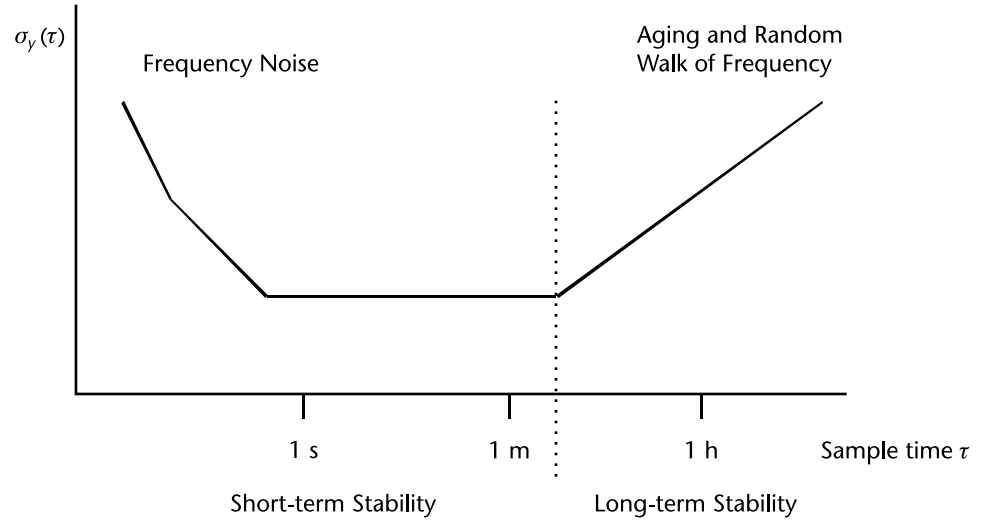

Figure 2.31 Time-domain stability.

aging must be subtracted from the data for long sample times. Appendix B provides additional details on the Allan variance and other measures of frequency stability.

The frequency versus temperature characteristics of crystal oscillators do not repeat exactly upon temperature cycling. For a TCXO, this thermal hysteresis is the difference between the frequency versus temperature characteristics for increasing temperatures and decreasing temperatures. Hysteresis is the major factor limiting the stability of TCXOs. Typical values range from 0.1 ppm to 1 ppm when the temperature cycling ranges are 0°C to 60°C and –55°C to 85°C. For an OCXO, the lack of repeatability is called "retrace" and is defined as the nonrepeatability of the frequency versus temperature characteristic at the oven temperature when it is cycled on and off. Retrace limits the achievable accuracy in applications where the OCXO is on/off cycled. Typical specifications, after a 24-hour off-period at 25°C, range from 1×10^{-9} to 2×10^{-8}. Low-temperature storage during the off-period and extending the off-period usually make the retrace worse.

2.7.1.2 TCXO

In a TCXO, a control network, composed of a temperature sensor (thermistor) and a varactor, is used to counteract the temperature-induced frequency change of the crystal. In contrast to the OCXO, the power consumption is very low (several milliwatts), which makes the TCXO attractive for handheld receivers, while the stability is relatively high. Furthermore, TCXOs are preferred to OCXOs in applications where a warm-up period is unacceptable. For a TCXO, the only warm-up time is the time required for the components to reach thermal equilibrium. As stated previously, TCXOs exhibit thermal hysteresis causing the frequency to jump when first started up. Keeping the TCXO biased would eliminate this effect. TCXOs provide a 20 times improvement in the crystal's frequency variation versus temperature in comparison to noncompensated oscillators [40]. TCXOs have improved in recent years to the point where they have comparable performance to oven-stabilized oscillators at lower cost and in smaller packages.