Understanding GPS/GNSS

Principles and Applications

Third Edition

Elliott D. Kaplan
Christopher J. Hegarty

Editors

ARTECH
HOUSE

BOSTON | LONDON
artechhouse.com

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

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

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

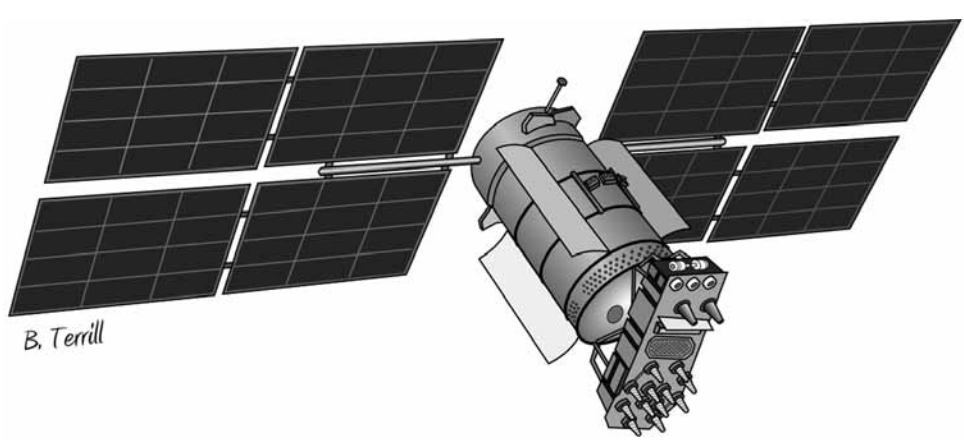


Figure 1.3 Glonass-M satellite. (Courtesy of Brian Terrill.)

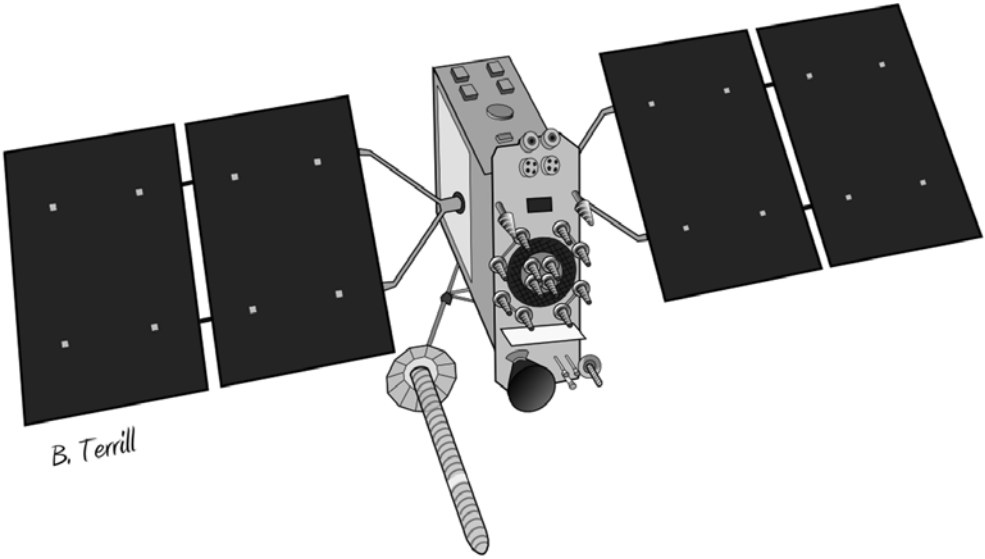


Figure 1.4 Glonass-K1 satellite. (Courtesy of Brian Terrill.)

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.

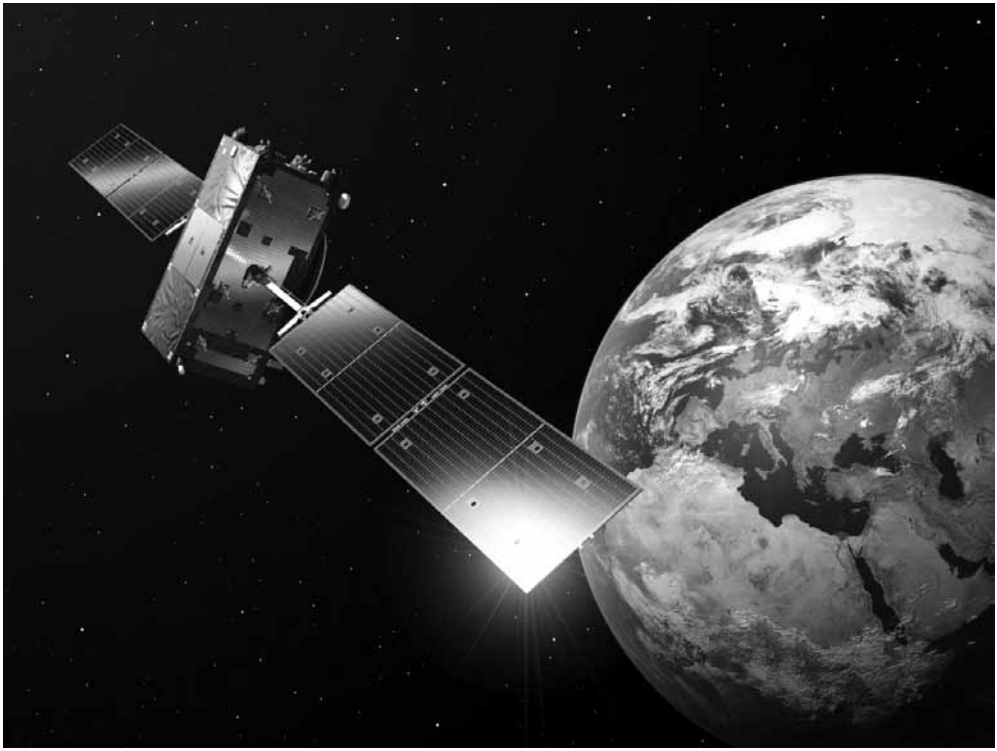


Figure 1.5 Galileo satellite. (©ESA-P. Carill.)

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

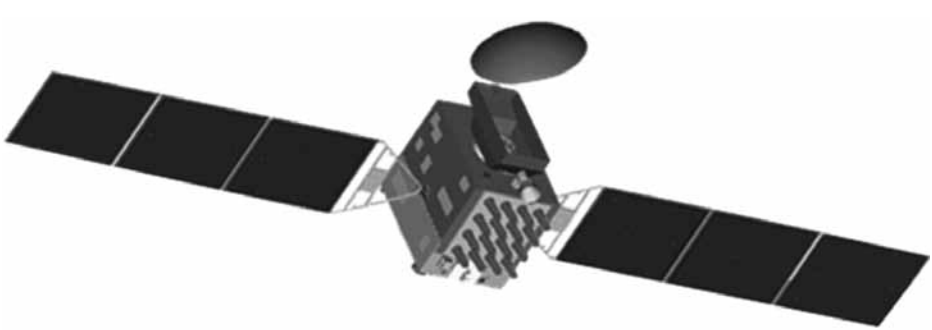


Figure 1.6 BDS GEO satellite [6].

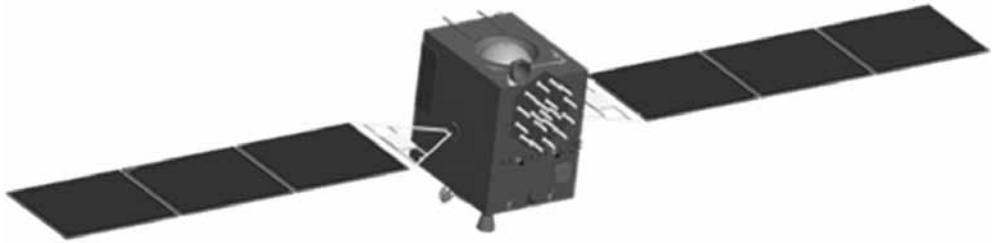


Figure 1.7 BDS IGSO/MEO satellite [6].

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

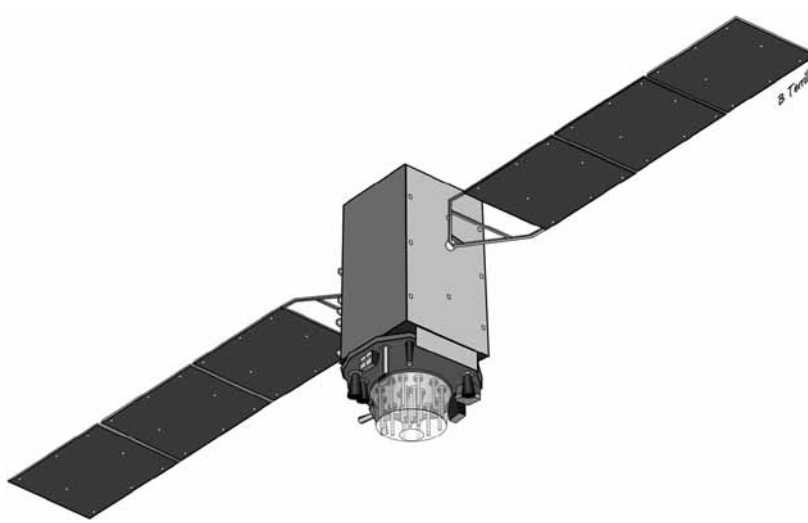


Figure 1.8 QZSS satellite. (Courtesy of Brian Terrill.)

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

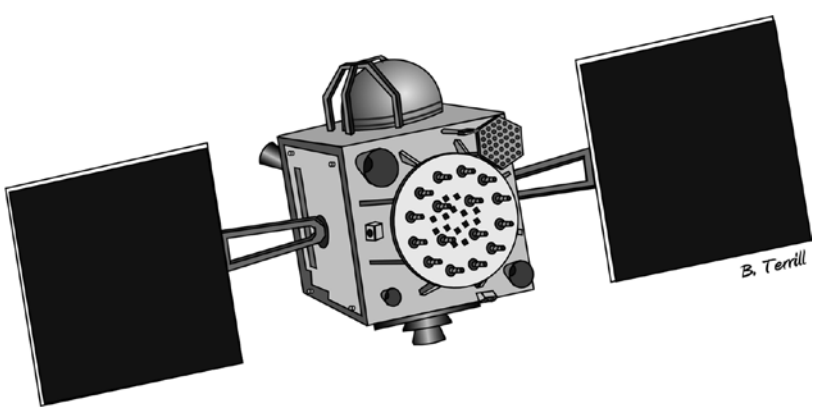


Figure 1.9 NavIC (IRNSS) satellite. (Courtesy of Brian Terrill.)

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.

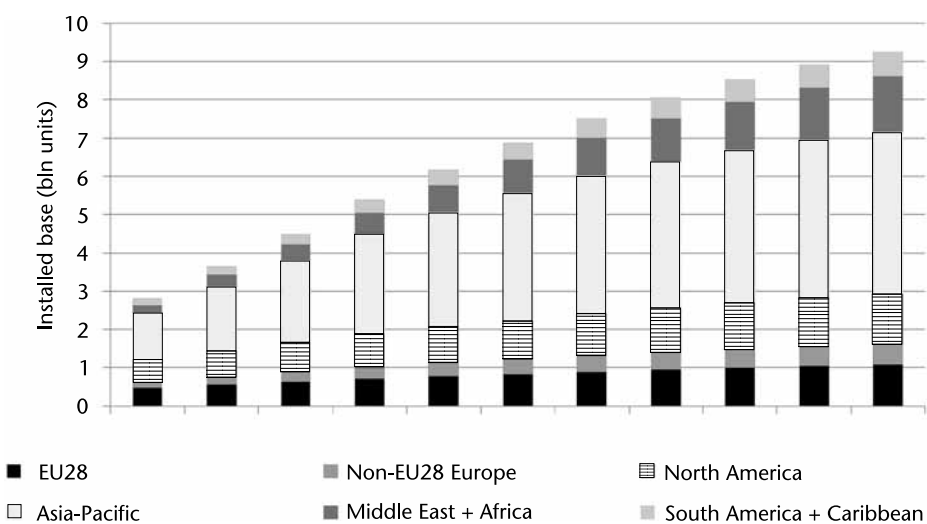


Figure 1.10 Installed base of GNSS devices by region. (Courtesy of GSA.)

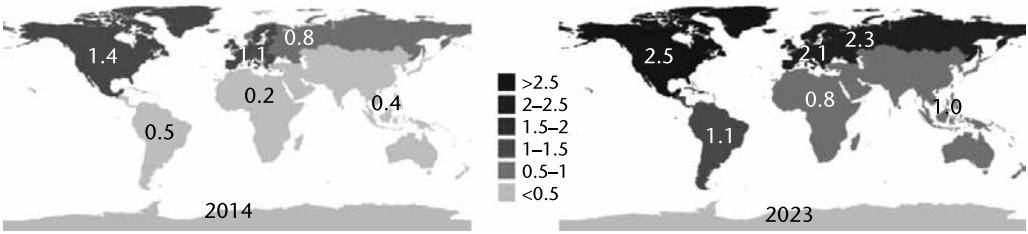


Figure 1.11 GNSS devices per capita: 2014 and 2023. (Courtesy of GSA.)

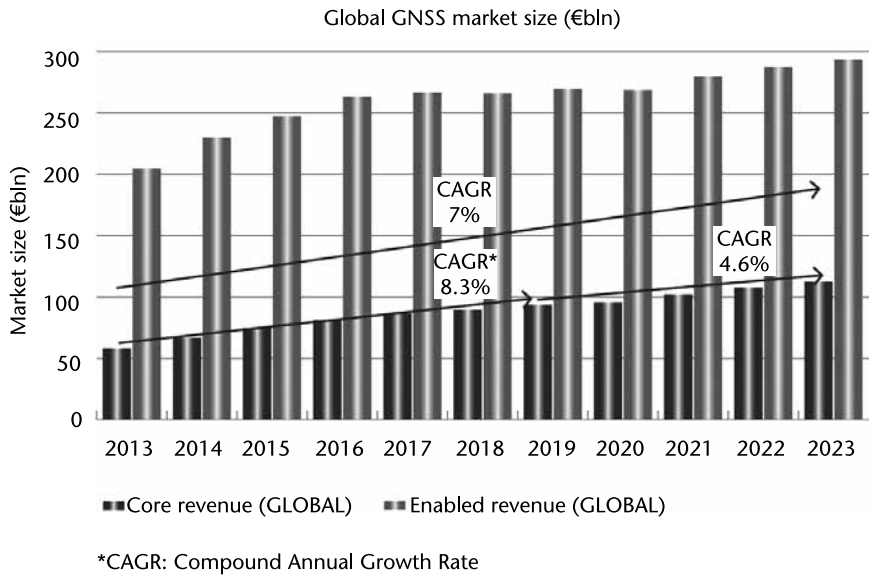


Figure 1.12 Global GNSS market size (billions of Euros). (Courtesy of GSA.)

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

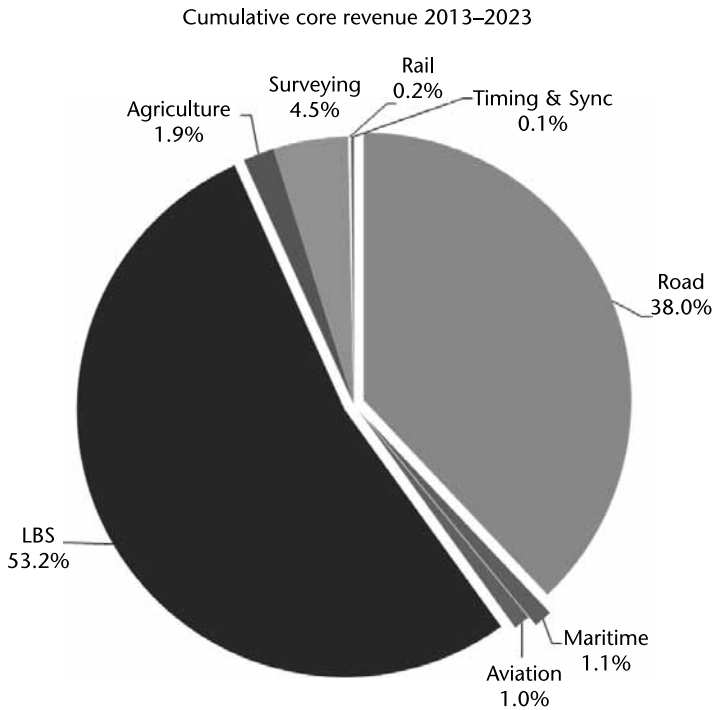


Figure 1.13 Cumulative core revenue 2013 to 2023 by market segment (billions of Euros). (Courtesy of GSA.)

(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.

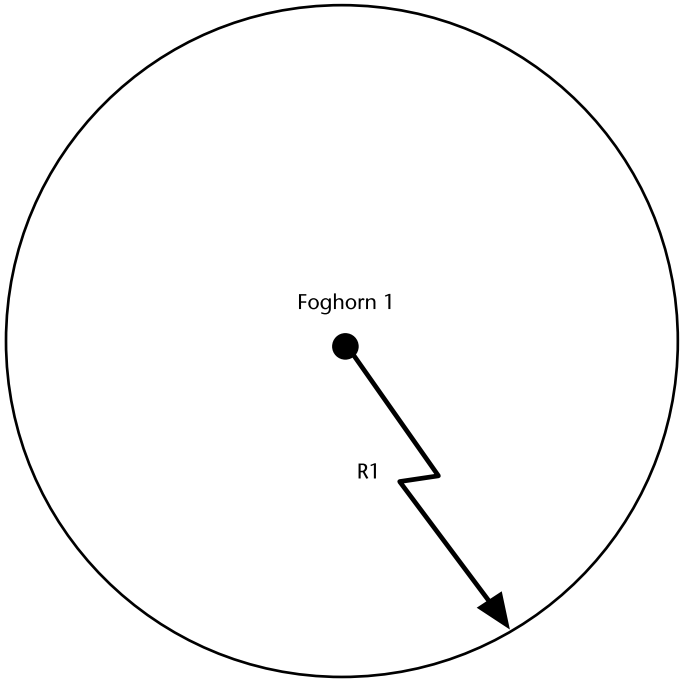


Figure 2.1 Range determination from a single source. (After: [1].)

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.

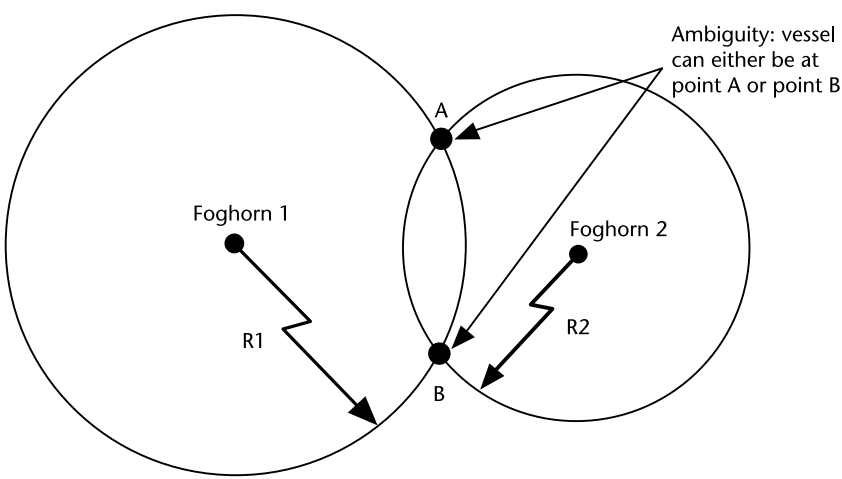


Figure 2.2 Ambiguity resulting from measurements to two sources. (After: [1].)

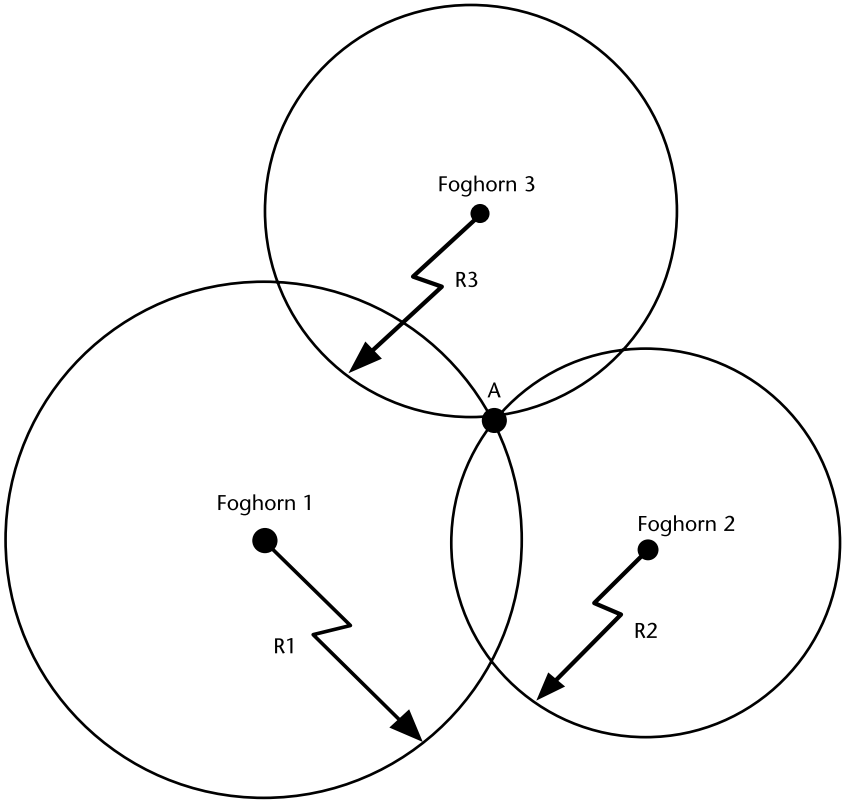


Figure 2.3 Position ambiguity removal by additional measurement. (After: [1].)

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

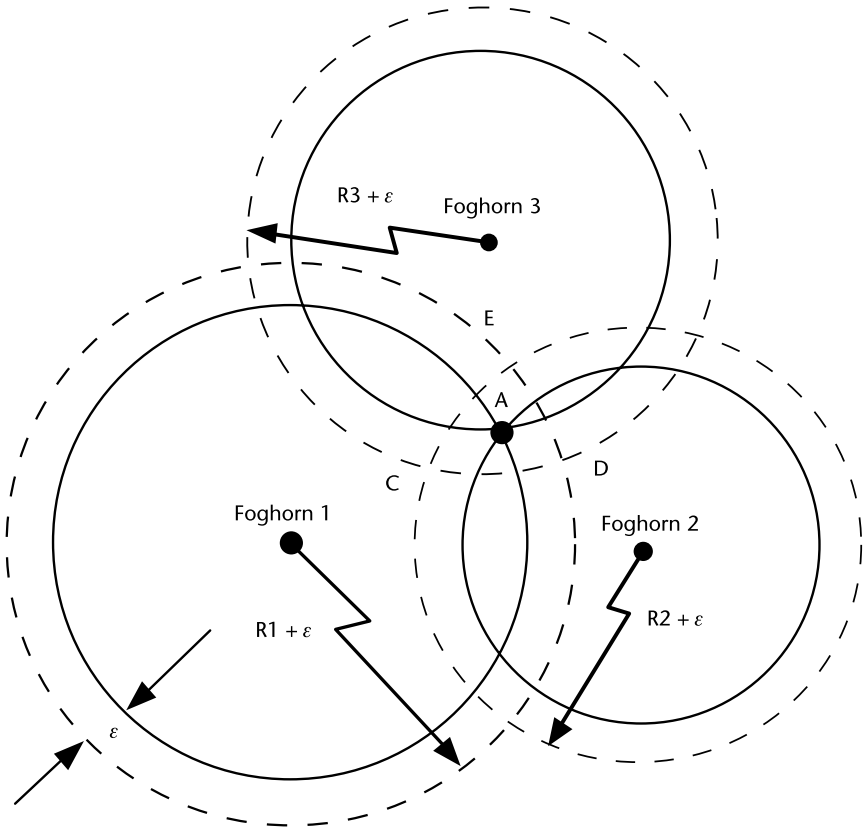


Figure 2.4 Effect of receiver clock offset on TOA measurements. (After: [1].)

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

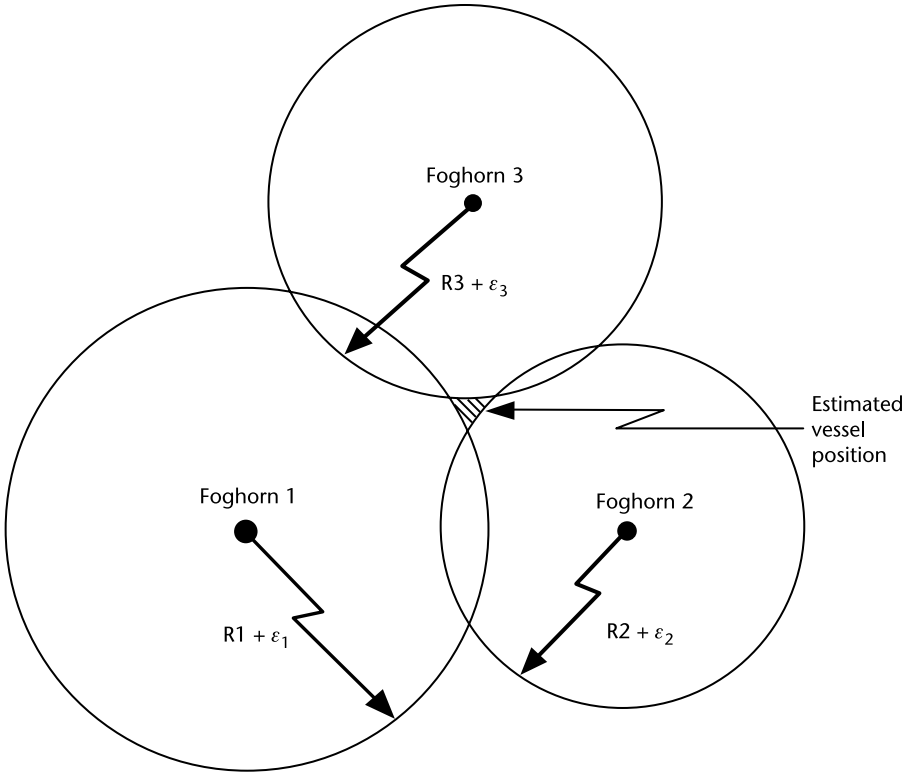


Figure 2.5 Effect of independent measurement errors on position certainty.

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

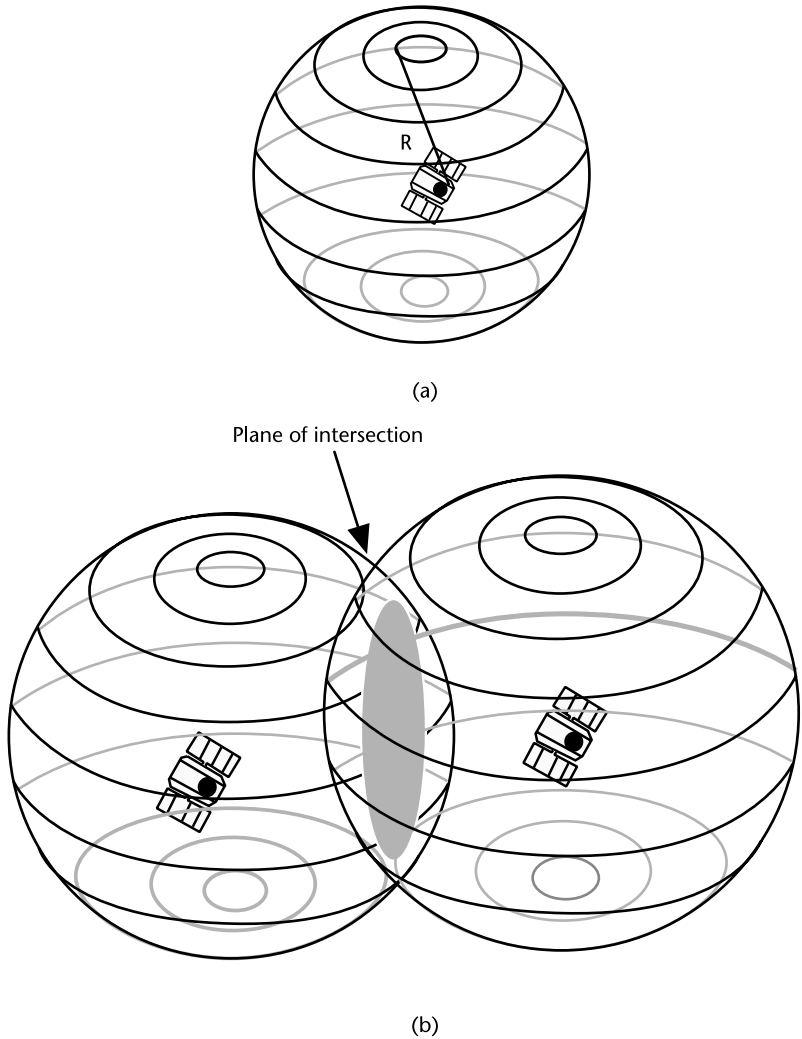


Figure 2.6 (a) User located on surface of sphere; (b) user located on perimeter of shaded circle (source: [2], reprinted with permission); (c) plane of intersection; (d) user located at one of two points on shaded circle (source: [2], reprinted with permission); and (e) user located at one of two points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

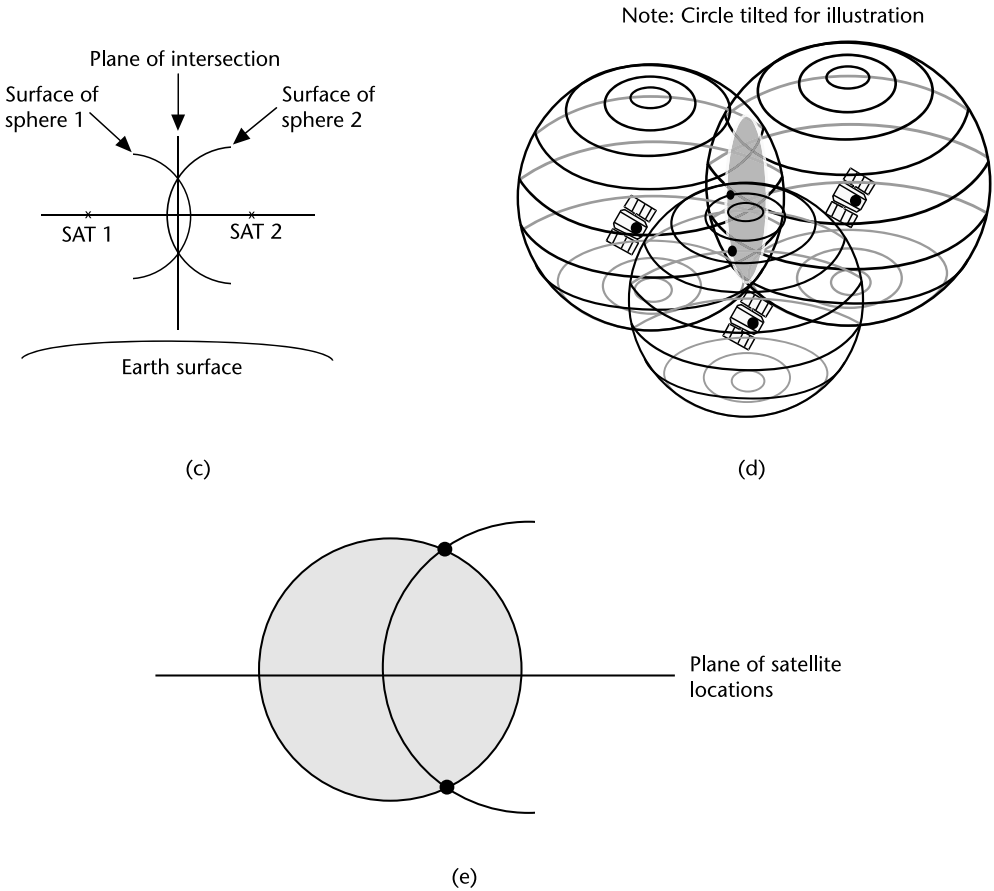


Figure 2.6 (continued)

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

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

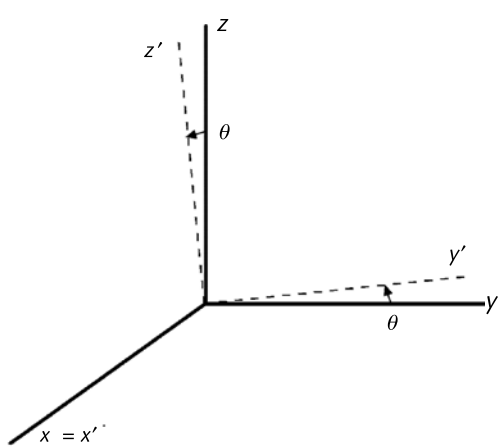


Figure 2.7 Example axial rotation, $\mathbf{R}_1(\theta)$ (x axis, positive θ).

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

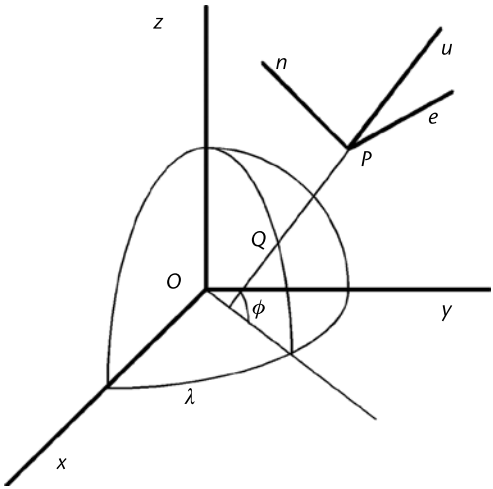


Figure 2.8 Relationships of ECEF and local tangent plane coordinate systems.

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

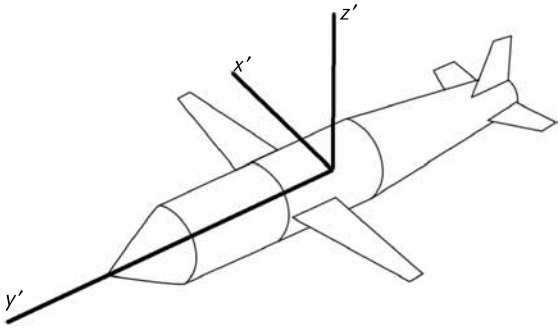


Figure 2.9 Example local body frame coordinate system.

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

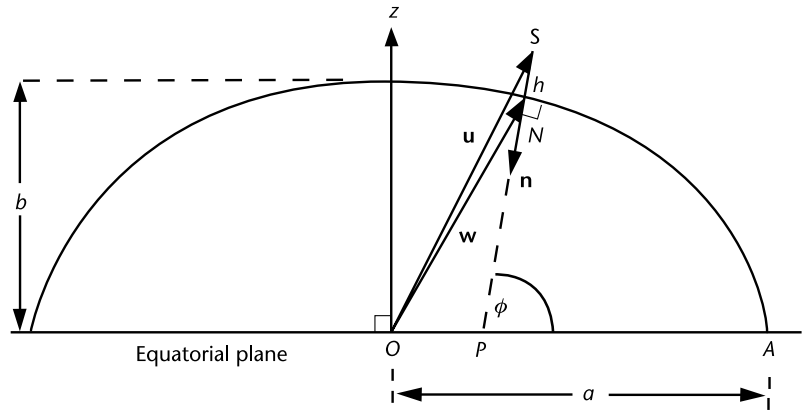


Figure 2.10 Ellipsoidal model of Earth (cross-section normal to equatorial plane).

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

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

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

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

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

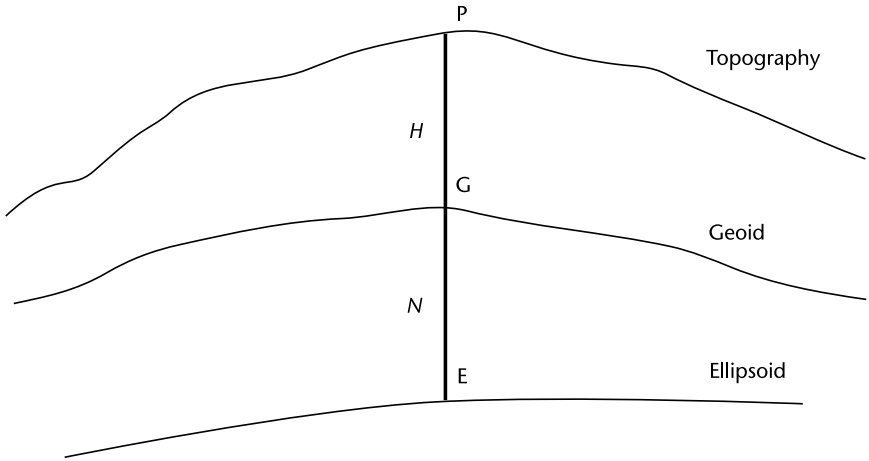


Figure 2.11 Relationships between topography, geoid, and ellipsoid.

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:

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

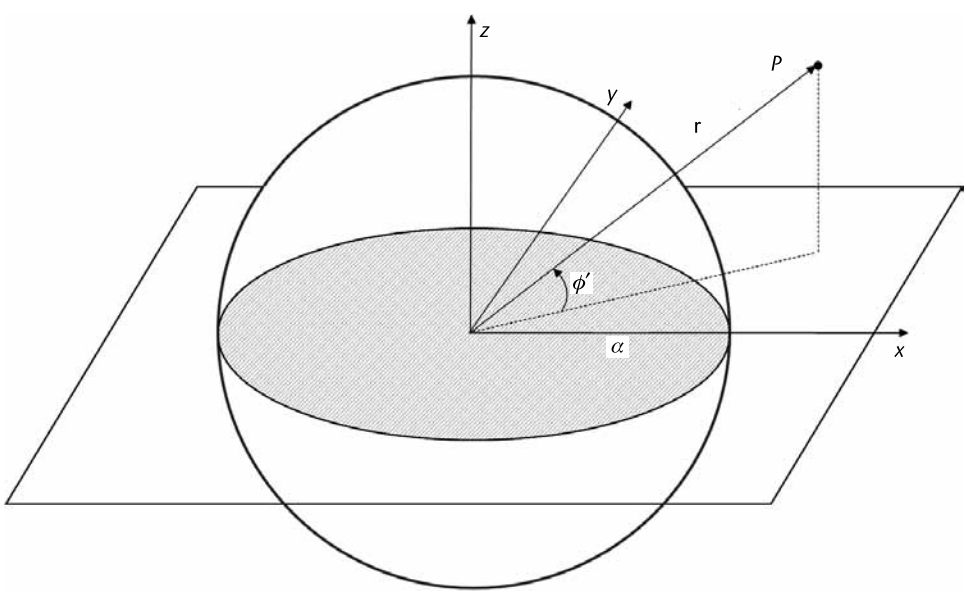


Figure 2.12 Illustration of spherical coordinate geometry

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

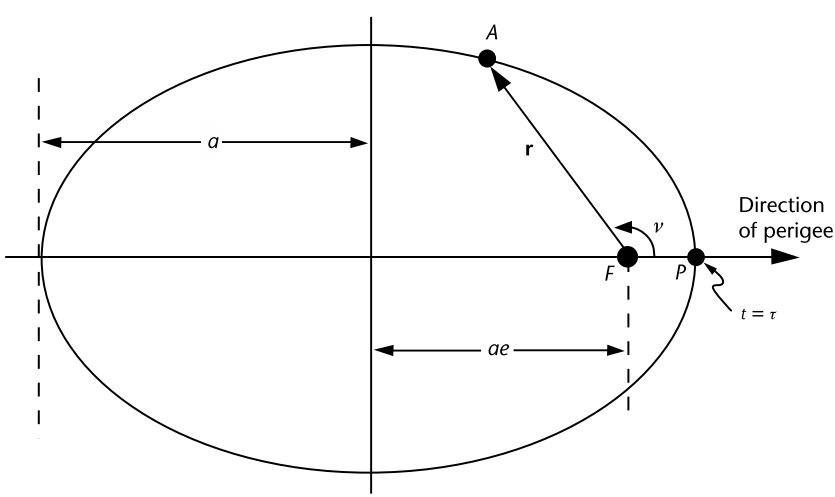


Figure 2.13 The three Keplerian orbital elements defining the shape of the satellite's orbit.

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

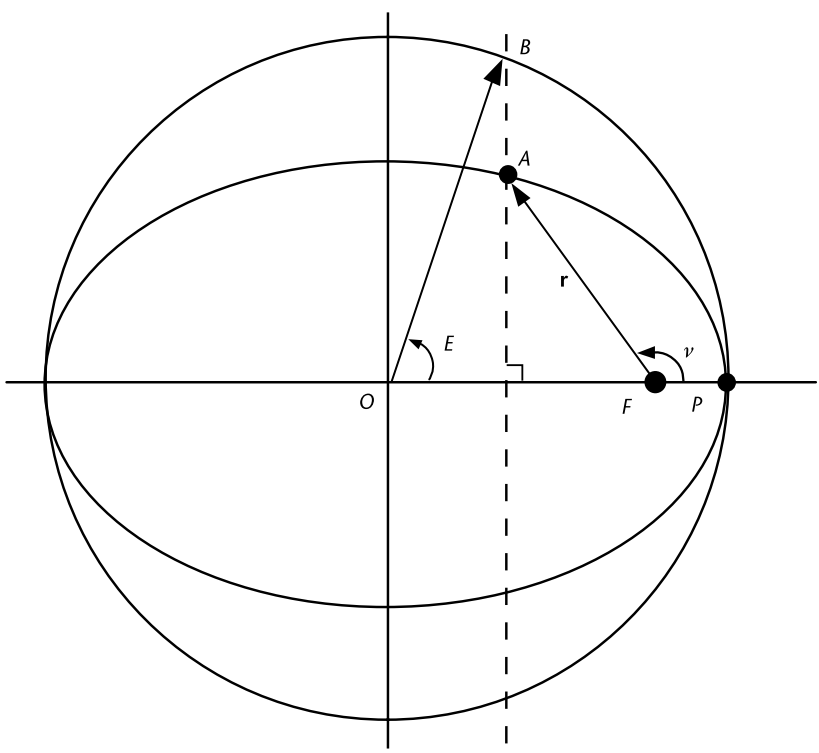


Figure 2.14 Relationship between eccentric anomaly and true anomaly.

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

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

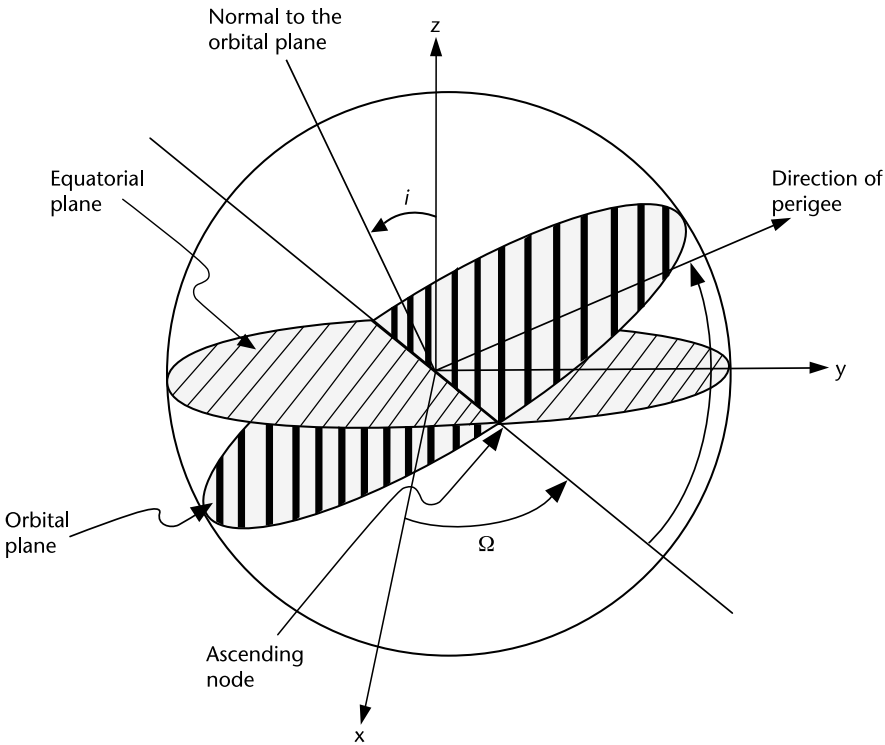


Figure 2.15 The three Keplerian orbital elements defining the orientation of the orbit.

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.

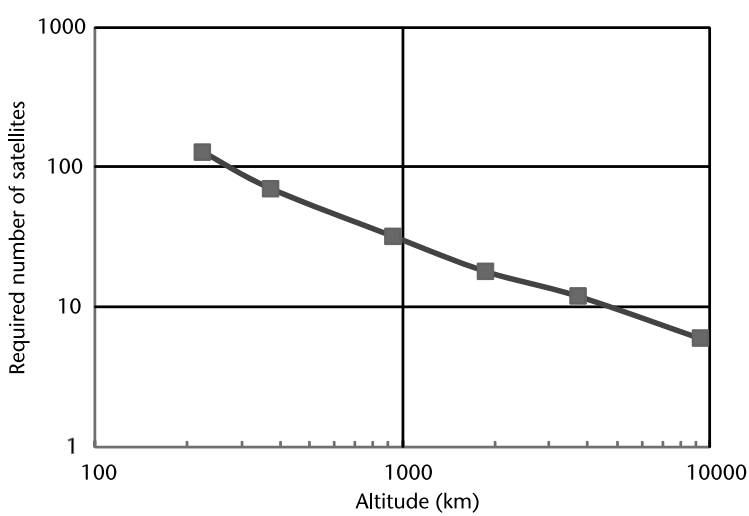


Figure 2.16 Number of satellites required to achieve at least one satellite in view worldwide at all times.

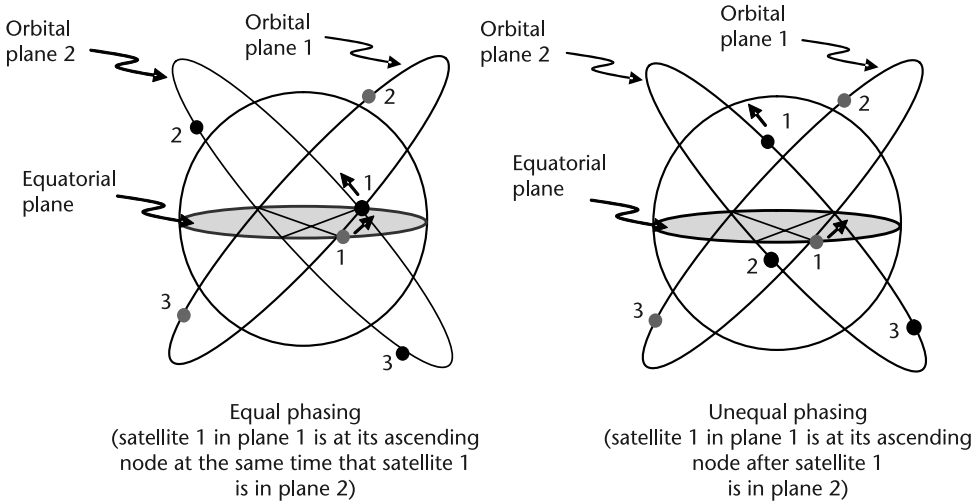


Figure 2.17 Equal versus unequal phasing between orbital planes.

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

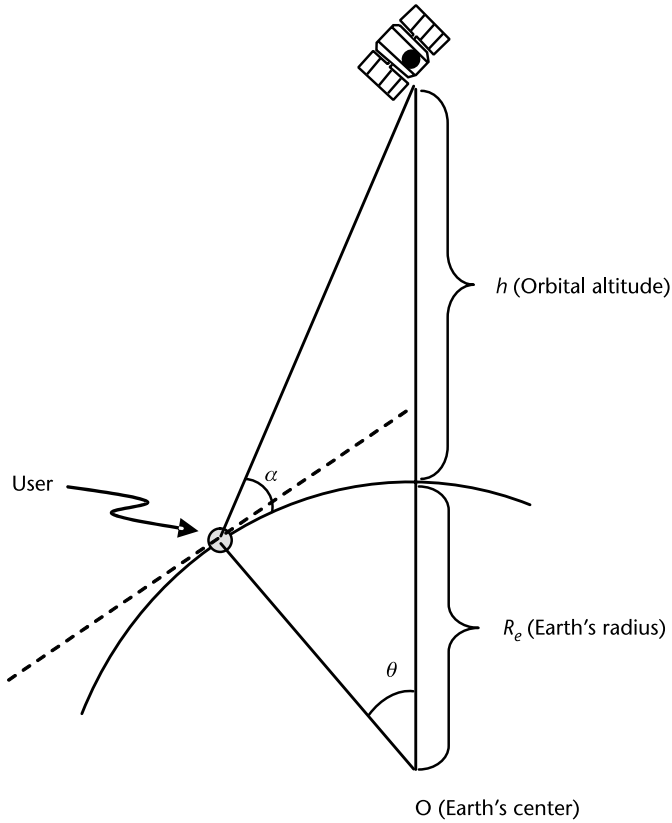


Figure 2.18 Relationship between elevation angle and Earth central angle (θ).

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

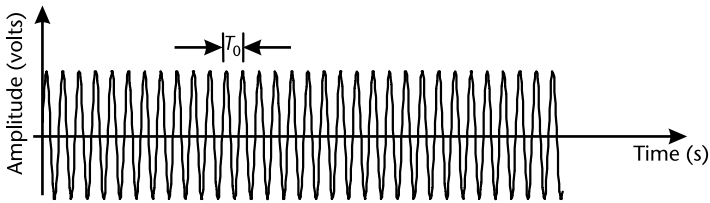


Figure 2.19 Radio frequency carrier.

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 $[0, 1] \rightarrow [-1, +1]$ or $[0, 1] \rightarrow [+1, -1]$. 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:

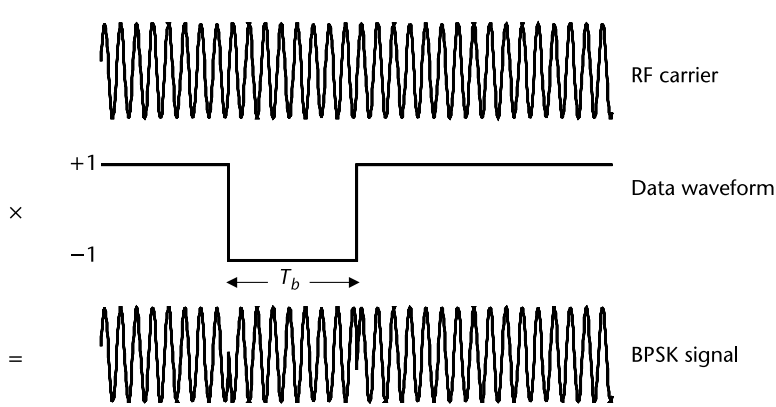


Figure 2.20 BPSK modulation.

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

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].

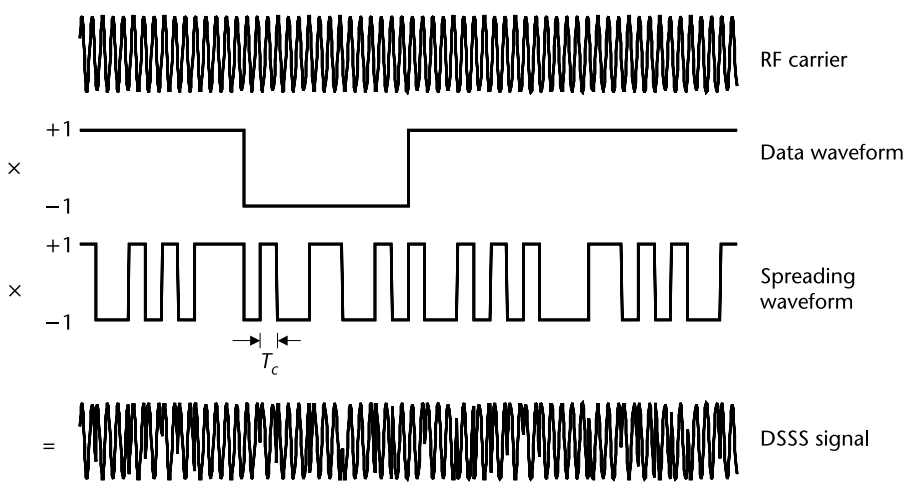


Figure 2.21 DSSS modulation.

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]:

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

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

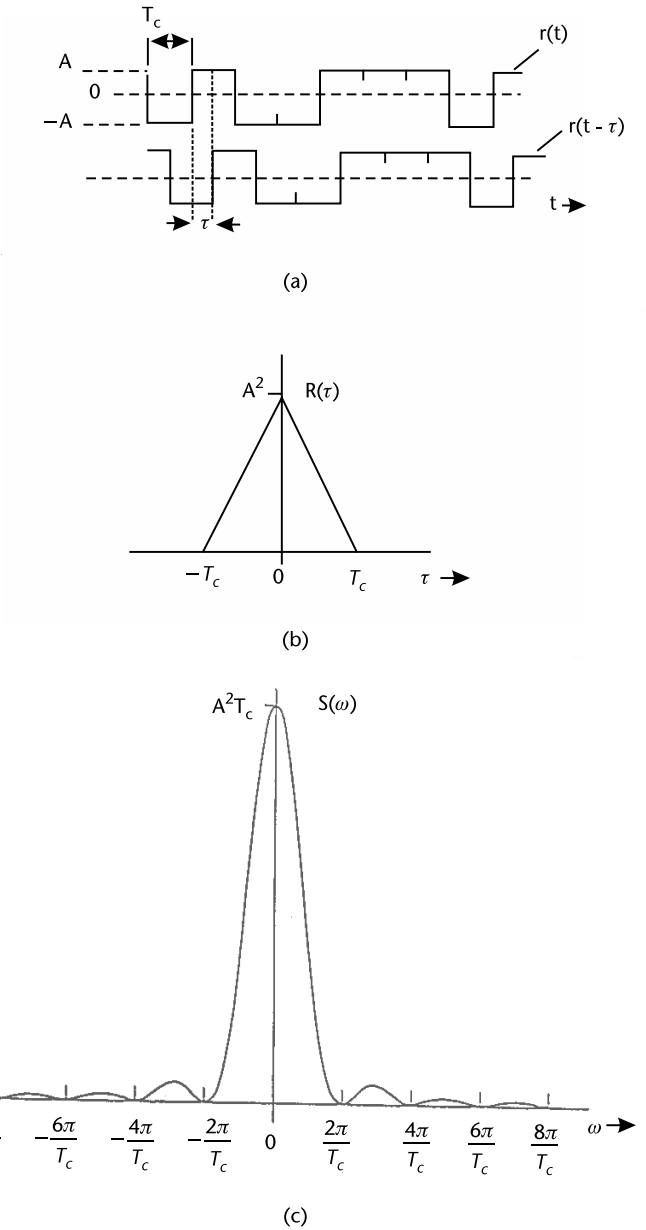


Figure 2.22 (a) A random binary code producing (b) the autocorrelation function and (c) the power spectrum of a DSSS signal.

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:

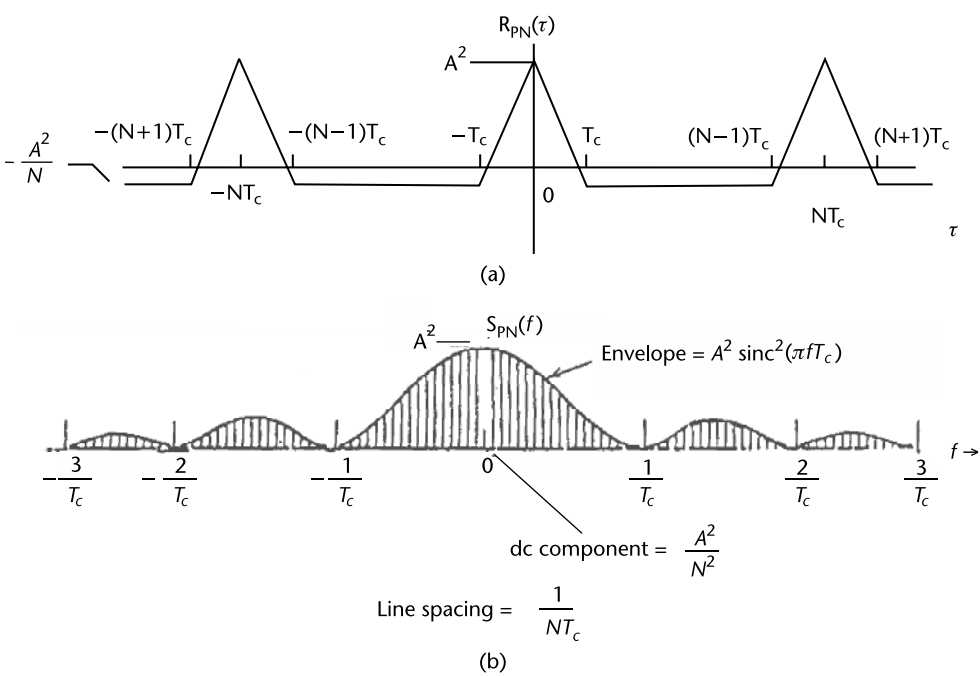


Figure 2.23 (a) The autocorrelation function of a DSSS signal generated from a maximum length pseudorandom sequence and (b) its line spectrum.

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

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:

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

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

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

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*
$\mathrm{BOC_s}(m, n)$	$2k - 1$	$\tau = jT_S/2$, $-2k + 2 \le j \le 2k - 2$	$(-1)^{j/2}(k-\lvert j/2\rvert)/k$	$(-1)^{(\lvert j\rvert-1)/2}/(2k)$
$\mathrm{BOC_c}(m, n)$	$2k + 1$	$\tau = jT_S/2$, $-2k + 1 \le j \le 2k - 1$	$(-1)^{j/2}(k-\lvert j/2\rvert)/k$	$(-1)^{(\lvert j\rvert+1)/2}/(2k)$

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

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

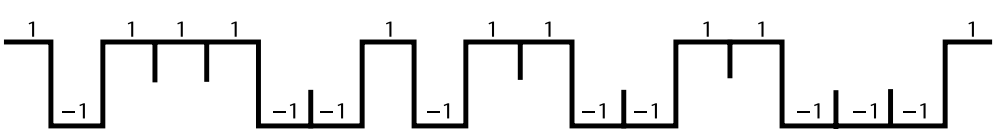


Figure 2.24 Ranging code.

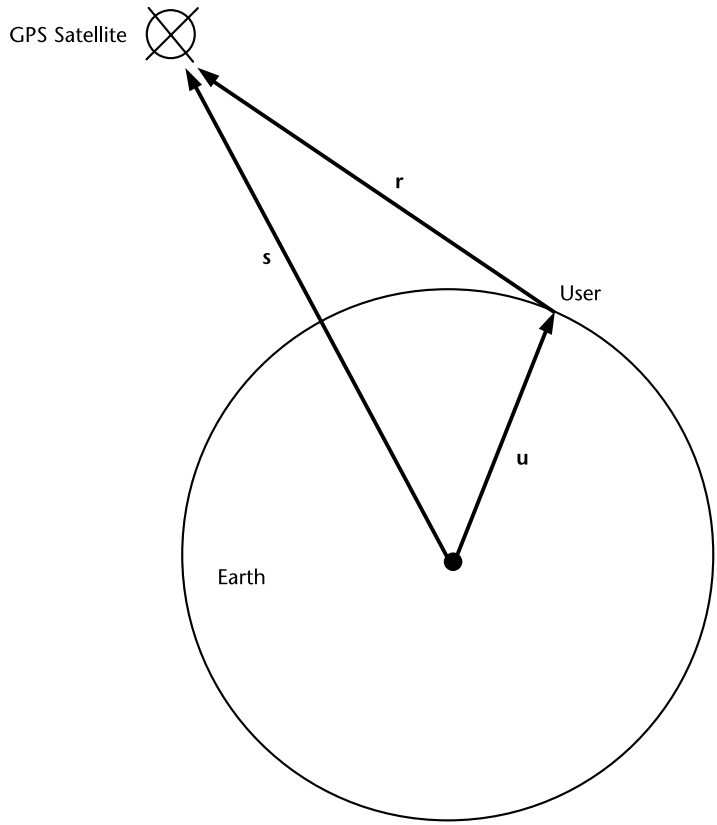


Figure 2.25 User position vector representation.

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

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

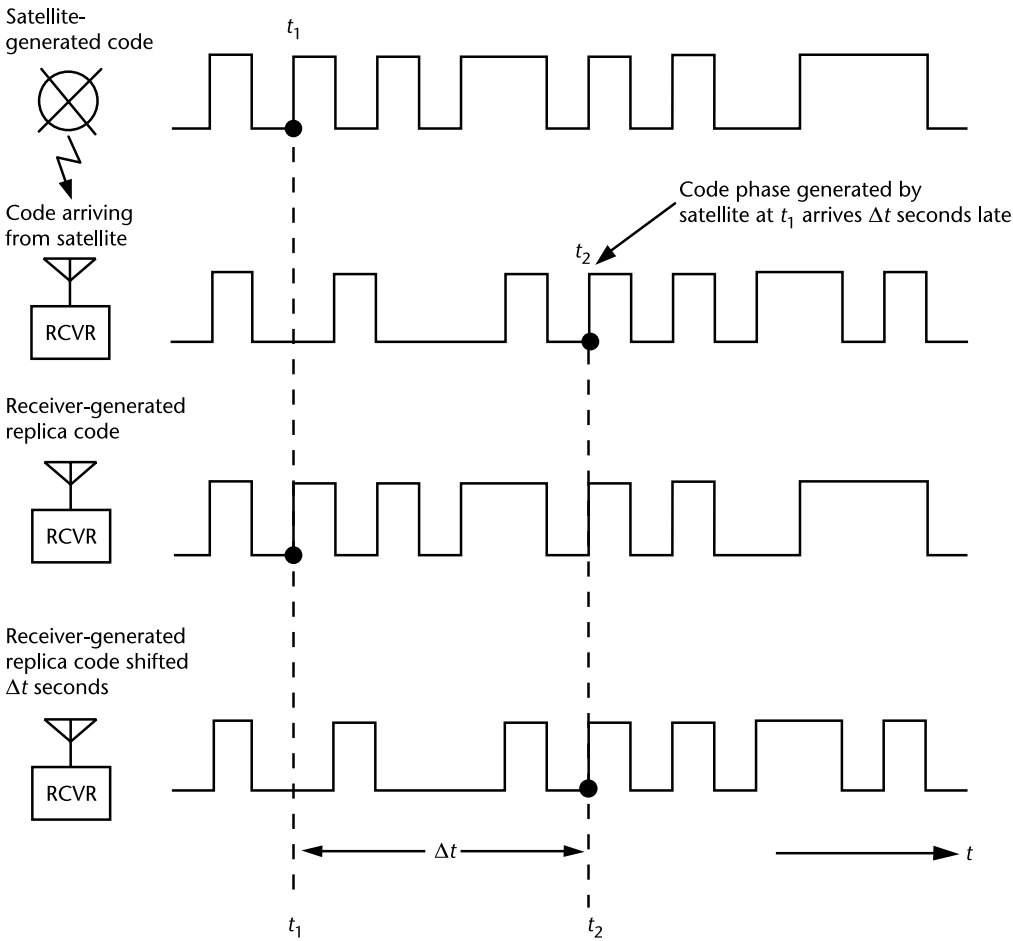


Figure 2.26 Use of replica code to determine satellite code transmission time.

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

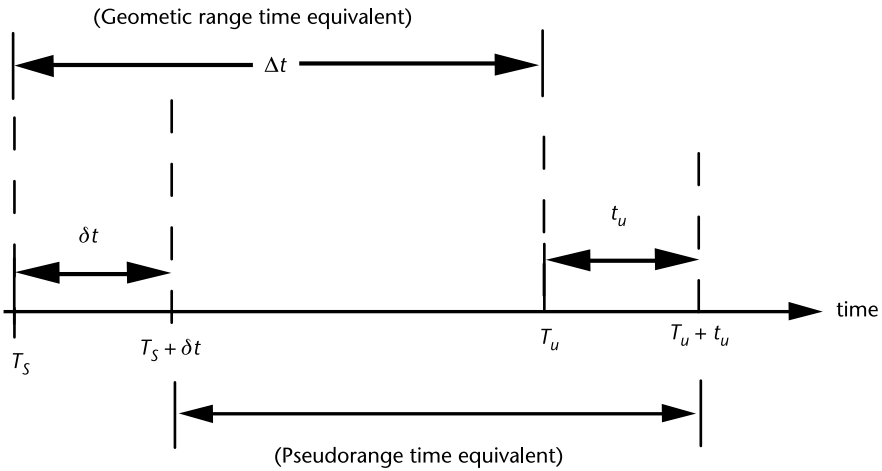


Figure 2.27 Range measurement timing relationships.

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

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

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

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

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

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

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:

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

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

The error contribution $\boldsymbol{\epsilon}_{\mathbf{x}}$ 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

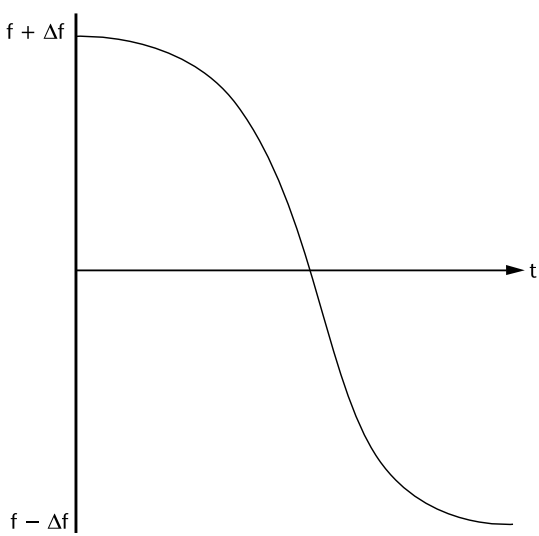


Figure 2.28 Received Doppler frequency by user at rest on Earth's surface.

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

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

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 **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

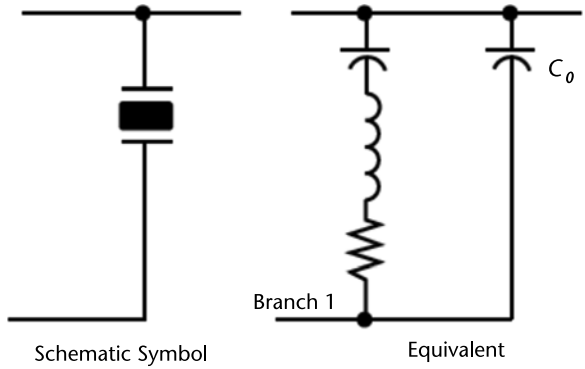


Figure 2.29 Crystal equivalent circuit. (From: [36].© Keysight Technologies, Inc. May 1997. Reproduced with permission, courtesy of Keysight Technologies.)

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

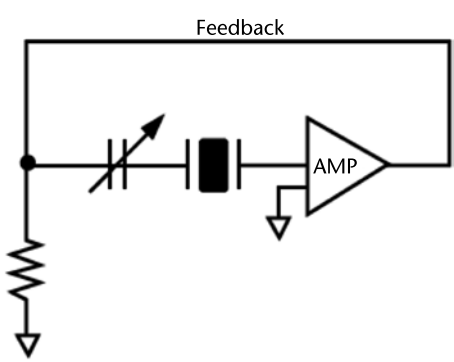


Figure 2.30 Simplified amplifier feedback (oscillator) circuit using a crystal resonator. (From: [36]. © Keysight Technologies, Inc. May 1997. Reproduced with permission, courtesy of Keysight Technologies.)

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,

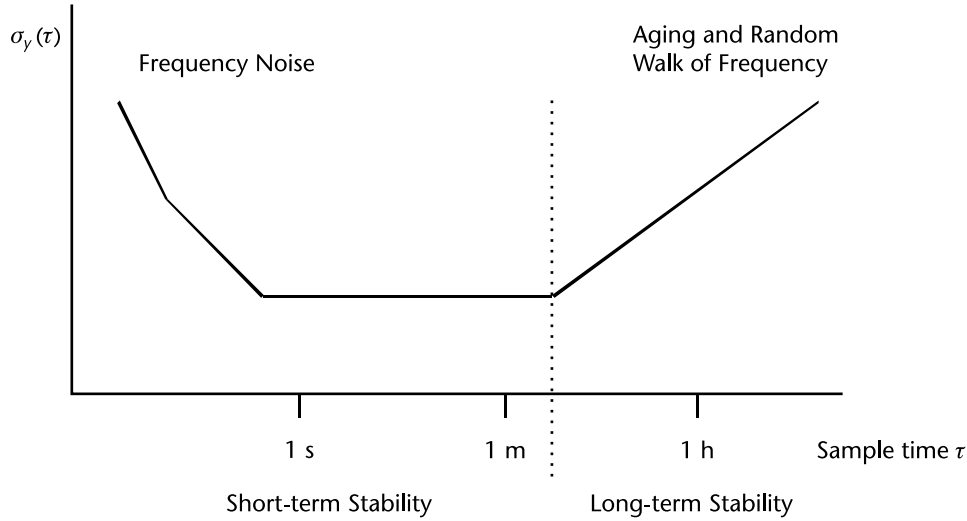


Figure 2.31 Time-domain stability.

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.

2.7.1.3 MCXO

The microcomputer controlled crystal oscillator (MCXO) exhibits aging and temperature stability that are ten times better than the TCXO. In the MCXO, a self-temperature sensing method is used that is much more sensitive than the external thermometer or thermistor that is used in TCXOs. Two modes of the crystal are excited simultaneously and heterodyned to generate a difference frequency that is a nearly linear function of temperature. The difference frequency is used to gate a reciprocal counter that uses the fundamental frequency as the time base. The counter's output is a number, N1, which varies with temperature, and is actually the period of the input signal in multiples of the master clock. The microcomputer compares N1 to stored calibration information and outputs a number, N2, to the correction circuit. In the active state, the power consumption is higher than that of the TCXO for non-CMOS outputs but is comparable to that of TCXOs for CMOS outputs. In standby mode, the power consumption is comparable to that of TCXOs. Another feature of the MCXO is that it has provisions to correct itself with a reference, GPS system time, for example.

2.7.1.4 OCXO

For airborne applications where low-power consumption and small size is not as critical, larger-size OCXOs with better performance are used. However, as mentioned earlier, the oven's high-power consumption precludes its use in handheld applications. In an OCXO, all temperature-sensitive components of the oscillator are maintained at a constant temperature in an oven. The oven temperature is set to coincide with the zero slope region of the crystal's frequency versus temperature characteristic. OCXOs require a few minutes to warm up and their power consumption is typically 1W or 2W at room temperature.

The characteristics of the different crystal oscillator types are summarized in Table 2.4.

2.7.1.5 Atomic Frequency Standard Description

A principal enabling technology for the deployment of GNSS is the atomic clock, or more precisely, atomic frequency standard (AFS), that each satellite uses to keep accurate time and frequency between ground updates. These atomic frequency standards utilized by GNSS were themselves the culmination of several Nobel Prizes in physics throughout the twentieth century. Despite many decades of scientific

Table 2.4 Summary of Different Crystal Oscillator Types

Oscillator Type	*TCXO*	*MCXO*	*OCXO*
Stability, $\sigma_y(\tau)$, τ = 1 second	10^{-9}	10^{-10}	10^{-12}
Aging/year	5×10^{-7}	5×10^{-8}	5×10^{-9}
Frequency offset after warm-up	10^{-6}	10^{-7} to 10^{-8}	10^{-8} to 10^{-10}
Warm-up time	10^{-6} to 10 seconds	10^{-8} to 10 seconds	10^{-8} to 5 minutes
Power	100 μW	200 μW	1–3W
Weight	50g	100g	200–500 g
Cost	$100	$1,000	$2,000

breakthroughs, these devices are still some of the most complicated and difficult technologies to produce reliably and of sufficient quality for the GNSS satellites. In this section, we will discuss the basics of how an AFS works and how Nobel Prize-winning breakthroughs could enable them to work better.

An AFS is built using two fundamental building blocks: one of them is a frequency source or local oscillator (LO), and the other is the atomic system. In GNSS and most other applications of frequency standards, the oscillator is a quartz crystal oscillator (XO) and usually oven controlled (OCXO). Quartz oscillators are found in many devices such as wristwatches, computers, radios, and radar. XOs are not stable enough to use for GNSS, so the LO is disciplined by the more accurate and stable frequency of the atomic system in a feedback arrangement as shown in Figures 2.32 and 2.33 for cesium (Cs) and rubidium (Rb) atomic systems, respectively.

2.7.1.6 AFS Principle of Operation

Within the atomic system, each atomic isotope (e.g., cesium or rubidium) is sensitive to particular frequencies determined by the unique arrangement of that isotope's electrons and nucleus as described with quantum mechanics. We do not need to

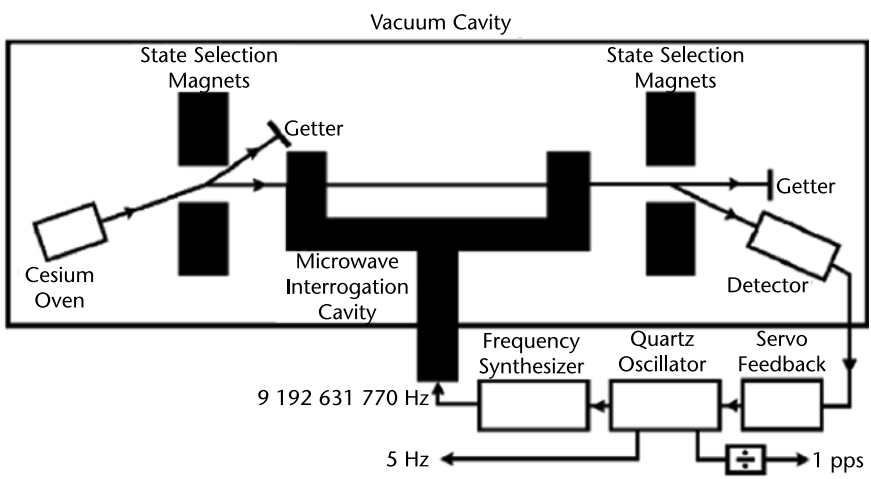


Figure 2.32 Cesium beam oscillator (From: [37].)

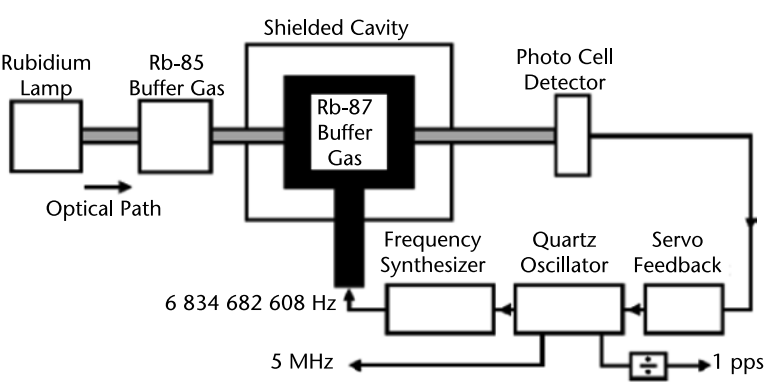


Figure 2.33 Rubidium oscillator. (From: [37].)

know quantum mechanics to understand the workings of an AFS within a SATNAV system, but we do need to understand the result. The key result is that an atom can only exist in a finite number of discrete states that dictate the discrete frequencies, and there are rules that govern how the atom can transition between those states. Atomic states are separated in energy and allowable transitions between states can be executed only when electromagnetic waves with a frequency proportional to the energy separation between the states interacts with the atom. The proportionality constant is Plank's constant, h. Therefore for two states, with energy E_1 and E_2 ($E_2 > E_1$), the transition can occur at frequency, f, if $f_{12}h = E_2 - E_1$. Figure 2.34 shows an example of what happens when the LO frequency is scanned while the atomic is monitored without the feedback turned on. When the feedback is on, the system would ideally hold the frequency of the LO to the center of the largest peak of the five transitions shown observed in the data. The states are typically labeled by an angular momentum nomenclature beyond the scope of this text, but here the strong transition is between a state with 3 units of angular momentum to another state with 5 units. The data in Figure 2.34 is an example taken from a frequency standard in development at the Air Force Research Laboratory [41].

In order for an atomic system to measure the frequency of the LO, a general sequence of steps must take place as illustrated in Figure 2.35. In Step 1, the AFS needs a gaseous sample of atoms because the atomic interactions of other phases of matter are too strong to make quality measurements. These atoms are in a random mixture of two states. In Step 2, the gas must be set in a known initial quantum state. In the case of cesium, atoms in one of the two states of interest are removed with a magnetic discriminator (see Figure 2.32) with a method invented by Otto Stern for which he won the Nobel Prize in 1973. In the case of rubidium, the atoms are forced into one common state with a mechanism called optical pumping (see Figure 2.33), invented by Alfred Kastler for which he won the Nobel Prize in 1966.

In Step 3, in Figure 2.35, the AFS needs a mechanism for illuminating the atoms with the electromagnetic waves from the LO. For both cesium and rubidium standards, a microwave frequency, synthesized from the LO, matches the atomic transition with the atoms in a microwave cavity.

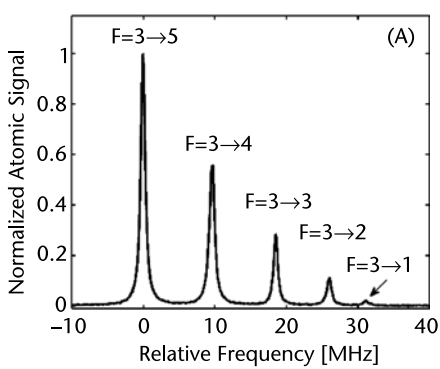


Figure 2.34 As the frequency of a local oscillator is scanned, a detector measures the atomic interaction. Here, transitions between several pairs of states (labeled by F numbers) occur.

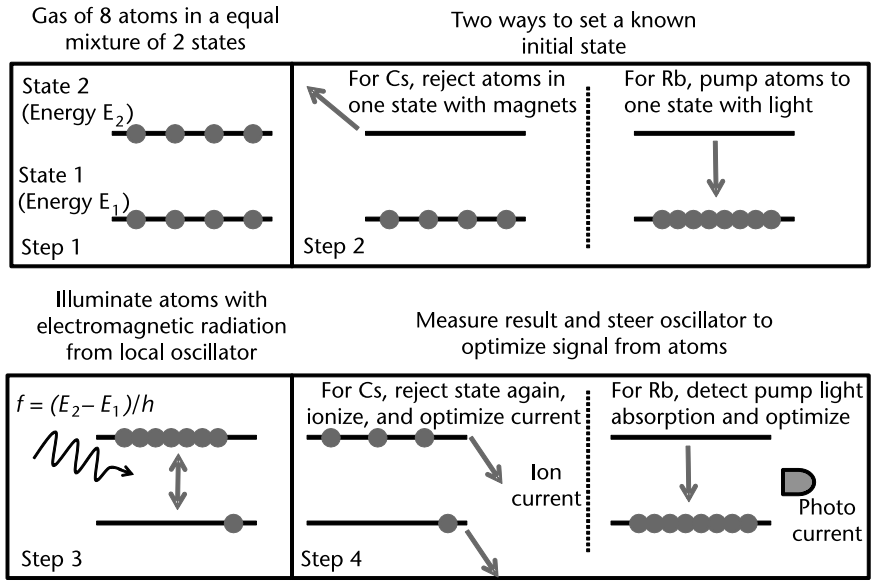


Figure 2.35 General sequence of steps needed to measure the LO frequency.

Finally, in Step 4, the AFS needs a mechanism for detecting the interaction of the waves from the LO with the atoms in the microwave cavity. For both cesium and rubidium, the same technique for state selection in Step 2 is used to detect the atomic interactions. For cesium, the magnetic discriminator selects the atoms that have interacted with the microwave field and directs them towards an ionization based detector. In the case of rubidium, the optical pump stops working when all of the atoms are pumped into the end state (state 1 in Figure 2.35). When this happens, the pump light passing through the rubidium gas in Figure 2.33 will not be absorbed by the atoms anymore. The microwave field causes transitioning of atoms back to the upper state where they can be pumped again. Therefore, a detector monitoring the light absorbed by the atoms will see more pump absorption when the LO is well matched to the atomic reference frequency.

The Cs AFS is a wonderful example of science enabling human advances, but it is not perfect. One problem is that the Cs AFS will eventually run out of Cs or fill up a disposal system because no one has discovered a way to recycle the atoms in this system. One could add more atoms, but this increases the size, weight, and cost of the system. The second and even more fundamental issue is that this system is very sensitive to magnetic fields and effectively couples magnetic field noise into frequency noise. To diminish the latter problem, several layers of passive magnetic shielding are required, which increases the size and weight of the system.

The rubidium AFS is more compact and longer-lived because the Rb atoms are stored in a heated glass cell and used over and over again. The Rb standard thus offers several key advantages: atoms are contained, thus increasing the AFS operating life without increasing the size. Also, the change from sorting magnets used in the Cs AFS to optical pumping makes the rubidium system more stable over time because it uses more of the atoms to make measurements and therefore produces a

stronger signal. The next generation of GPS (Block III) uses only Rb standards. In Figure 2.33, one might notice that there are two rubidium cells with different isotopes: 85 and 87. The 85 isotope of Rb is used as an optical filter for the lamp light such that the optical pumping works properly on the 87 isotope. This technique uniquely works for rubidium which is why the atom of choice switched away from cesium. See [42] for a comparison of Rb to Cs on GPS Block IIF.

2.7.1.7 Advanced Atomic Frequency Standards

Despite the frequency stability offered by the atomic frequency standards on GNSS, the standards are still one of the limiters for overall accuracy of the navigation signal. The position equivalent time error is roughly equal to the ephemeris error (see Section 10.2.2) [43]. Beyond improving system accuracy in ideal conditions, improved frequency standards can also decrease system maintenance and improve reliability because a lower rate of accumulated time error requires less intervention from the ground to maintain GNSS operability. The rubidium and cesium atomic frequency standards have several fundamental limitations. The cesium atomic frequency standard performance is limited by the temperature of the atoms. This is because the beam of atoms expands in the transverse direction until too few atoms pass through the whole system. They are also moving so fast (Cs atoms at 50°C have a most probable speed over 200 m/s) that there is not enough time to interact with the microwave fields for best results. Also, there is a large spread of possible velocities, each of which have a different frequency shift associated with the Doppler effect, meaning that a larger than optimal spread of frequencies will interact with the frequency standard. Cold, but still gaseous atoms would be better.

In 1997, the Nobel Prize was given for laser cooling of atomic gases to William Phillips, Steven Chu, and Claude Cohen-Tannoudji. The result of laser cooling is to remove nearly all of the thermal energy from a cloud of atoms and then by adjusting those lasers frequencies, the atoms can be launched at a known, slow velocity. The result is that the long microwave-atom interaction times can be used and large stability and accuracy improvements can be made. Note that these improvements are only realizable with commensurate improvements to many other parts of the system such as magnetic field and stray light controls as well as stringent alignment tolerances. The civilian and military time standards now use this approach operationally at the National Institute of Standards and Technology (NIST) [44] and the U.S. Naval Observatory (USNO).

Today, the state of the art does not use microwaves. Scientists are using optical transitions at hundreds of terahertz, rather than microwave transitions of the order of 10 GHz. Here, a laser acts as the local oscillator. The benefits of optical transitions are that the frequency stability is improved by the ratio of the frequency, roughly 500 THz to 10 GHz, or about 50,000. Also, many systematic errors reduce in a similar way, especially errors due to magnetic fields. However, a frequency standard at several hundred terahertz is not particularly useful because the electronic systems on GNSS would have no way of counting or generating signals from such a high frequency. In 2005, the Nobel Prize was awarded to John Hall and Theodore Hänsch for inventing a way to solve this problem. Their device, called the frequency comb, effectively divides optical frequencies between 100 and

500 THz by a factor of roughly 1 million. The frequency comb device is flexible meaning that it can be designed to a specific input and output frequency amongst a broad range. Today, there are several examples of clocks accurate to a few parts in 10^{18} [45–47] that utilize both laser cooling and frequency combs; that is 1,000 times better than clocks on GNSS. Thus, the GNSS of the future does not need to be limited by the atomic frequency standard technology that enabled its creation.

2.7.2 Time and GNSS

A SATNAV system disseminates a realization of UTC that provides the capability for time synchronization of users either worldwide or within its coverage region. Applications range from *time-tagging* of banking transactions to communications system packet switching synchronization. Worldwide time dissemination is an especially useful feature in military frequency-hopping communications systems where time synchronization enables all users to change frequencies simultaneously. In many countries, UTC is used as the definition of time in legal matters [48].

2.7.2.1 UTC Generation

UTC is a composite time scale. That is, UTC is comprised of inputs from a time scale derived from atomic standards and information regarding the Earth's rotation rate. The time scale based on atomic standards is called International Atomic Time (TAI). TAI is a uniform time scale based on the atomic second, which is defined as the fundamental unit of time in the International System of Units [49]. The atomic second is defined as "the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom." The Bureau International des Poids et Mesures (BIPM) is the international body responsible for computing TAI. TAI is derived from more than 400 atomic standards located at laboratories in various countries [48]. The BIPM statistically processes these inputs to calculate definitive TAI. TAI is referred to as a paper time scale because it is not kept by a physical clock [50].

The other time scale used to form UTC is called Universal Time 1 (UT1). UT1 is a measure of the Earth's rotation angle with respect to the Sun. It is one component of the Earth orientation parameters that define the actual orientation of the ECEF coordinate system with respect to space and celestial bodies and is treated as a time scale in celestial navigation [50]. UT1 remains a nonuniform time scale due to variations in the Earth's rotation. Also, UT1 drifts with respect to atomic time. This is on the order of several milliseconds per day and can accumulate to 1 second in a 1-year period. The International Earth Rotation and Reference System Service (IERS) is responsible for definitively determining UT1. Civil and military timekeeping applications require knowledge of the Earth's orientation as well as a uniform time scale. UTC is a time scale with these characteristics. The IERS determines when to add or subtract leap seconds to UTC such that the difference between UTC and UT1 does not exceed 0.9 second. Thus, UTC is synchronized with solar time at the level of approximately one second [50]. Each SATNAV system provider maintains an ensemble of AFSs and forms its own version of UTC. These versions are usually kept within nanoseconds of the international standard UTC, provided by the BIPM approximately one month in arrears.

2.7.2.2 SATNAV System Time

As stated earlier in this chapter, a fundamental GNSS principle of operation is the need for synchronization of each satellite AFS (i.e., satellite clock) to its corresponding SATNAV system time (e.g., Galileo satellite clocks must be synchronized to Galileo system time). System time is an internal time scale within a SATNAV system. Based on an ensemble of atomic frequency standards (AFSs), system time provides the exact timing needed by a SATNAV system's users to make precise PVT measurements. When combining measurements from multiple GNSS constellations, the difference between SATNAV system times (e.g., BeiDou-GPS) must be accounted for (see Chapter 11).

For most SATNAV systems, system time is based on a continuous time scale. That is, it is not modified to reflect variations in the Earth's rotation (i.e., not adjusted for leap seconds). SATNAV system time is typically steered to a local realization of UTC, modulo 1 s [48], enabling interoperability with other SATNAV systems. The exception to the above is GLONASS system time, which adds or subtracts leap seconds to follow UTC. Descriptions of system time are provided for each SATNAV system discussed in the following chapters.

References

[1] NAVSTAR GPS Joint Program Office (JPO), *GPS NAVSTAR User's Overview*, YEE-82-009D, GPS JPO, March 1991.

[2] Langley, R., "The Mathematics of GPS," *GPS World Magazine*, Advanstar Communications, July/ August 1991, pp. 45–50.

[3] GPS Directorate, *Navstar GPS Space Segment/Navigation User Interfaces, IS-GPS-200*, Revision H, September 24, 2013.

[4] Long, A. C., et al., (eds.), *Goddard Trajectory Determination System (GTDS) Mathematical Theory*, Revision 1, FDD/552-89/001, Goddard Space Flight Center, Greenbelt, MD, July 1989.

[5] Xu, G., *GPS: Theory, Algorithms and Applications*, New York: Springer-Verlag, 2003.

[6] Noureldin, A., T. B. Karamat and J. Gregory, *Fundamentals of Inertial Navigation, Satellite-Based Positioning and Their Integration*, New York: Springer-Verlag, 2013.

[7] Bowring, B. R., "Transformation from Spatial to Geographical Coordinates," *Survey Review*, Vol. XXIII, No. 181, July 1976, pp. 323–327.

[8] Pavlis, N. K., et al., "The Development and Evaluation of the Earth Gravitational Model 2008 (EGM2008)," *Journal of Geophysical Research: Solid Earth*, Vol. 117, No. B4, April 2012.

[9] Rapp, R. H., "Separation Between Reference Surfaces of Selected Vertical Datums," *Bulletin Geodesique*, Vol. 69, No. 1, 1995, pp. 26–31.

[10] Milbert, D. G., "Computing GPS-Derived Orthometric Heights with the GEOID90 Geoid Height Model," *Technical Papers of the 1991 ACSM-ASPRS Fall Convention*, Atlanta, GA, October 28–November 1, 1991, pp. A46–55.

[11] Parker, B., et al., "A National Vertical Datum Transformation Tool," *Sea Technology*, Vol. 44, No. 9, September 2003, pp. 10–15.

[12] Petit, G., and B. Luzum (eds.), "IERS Conventions (2010)," *IERS Technical Note 36*, Verlag des Bundesamts fur Kartographie und Geodasie, Frankfurt am Main, 2010.

[13] Drewes, H., et al. (eds.), "The Geodesist's Handbook," *Journal of Geodesy*, 2012.

[14] http:// www.iers.org/IERS/EN/Home/home_node.html.

[15] http://itrf.ign.fr/ITRF_solutions/2014.

[16] http://igs.org.

[17] Battin, R. H., *An Introduction to the Mathematics and Methods of Astrodynamics*, New York: AIAA, 1987.

[18] Walker, J. G., "Satellite Constellations," *Journal of the British Interplanetary Society*, Vol. 37, 1984, pp. 559–572.

[19] Rider, L., "Analytical Design of Satellite Constellations for Zonal Earth Coverage Using Inclined Circular Orbits," *The Journal of the Astronautical Sciences*, Vol. 34, No. 1, January–March 1986, pp. 31–64.

[20] Adams, W. S., and L. Rider, "Circular Polar Constellations Providing Continuous Single or Multiple Coverage Above a Specified Latitude," *The Journal of the Astronautical Sciences*, Vol. 35, No. 2, April–June 1987, pp. 155–192.

[21] Jorgensen, P. S., "NAVSTAR/Global Positioning System 18-Satellite Constellations," *NAVIGATION, Journal of The Institute of Navigation*, Vol. 27, No. 2, Summer 1980, pp. 89–100.

[22] Proakis, J., *Digital Communications*, 4th ed., New York: McGraw-Hill, 2000.

[23] Simon, M., et al., *Spread Spectrum Communications Handbook*, McGraw-Hill, New York, 1994.

[24] Betz, J., "Binary Offset Carrier Modulations for Radionavigation," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 48, No. 4, Winter 2001–2002.

[25] Hegarty, C., J. Betz, and A. Saidi, "Binary Coded Symbol Modulations for GNSS," *Proceedings of The Institute of Navigation Annual Meeting*, Dayton, OH, June 2004.

[26] Butman, S., and U. Timor, "Interplex - An Efficient Multichannel PSK/PM Telemetry System," *IEEE Transactions on Communication Technology*, Vol. COM-20, No. 3, June 1972.

[27] Spilker, J. J., Jr., *Digital Communications by Satellite*, Englewood Cliffs, NJ: Prentice-Hall, 1977.

[28] Cangiani, G., R. Orr, and C. Nguyen, *Methods and Apparatus for Generating a Constant-Envelope Composite Transmission Signal*, U.S. Patent Application Publication, Pub. No. US 2002/0075907 A1, June 20, 2002.

[29] Forssell, B., *Radionavigation Systems*, Upper Saddle River, NJ: Prentice Hall, 1991, pp. 250–271.

[30] Holmes, J. K., *Coherent Spread Spectrum Systems*, Malabar, FL: Krieger Publishing Company, 1990, pp. 344–394.

[31] Leva, J., "An Alternative Closed Form Solution to the GPS Pseudorange Equations," *Proceedings of The Institute of Navigation (ION) National Technical Meeting*, Anaheim, CA, January 1995.

[32] Bancroft, S., "An Algebraic Solution of the GPS Equations," *IEEE Trans. on Aerospace and Electronic Systems*, Vol. AES-21, No. 7, January 1985, pp. 56–59.

[33] Chaffee, J. W., and J. S. Abel, "Bifurcation of Pseudorange Equations," *Proceedings of The Institute of Navigation National Technical Meeting*, San Francisco, CA, January 1993, pp. 203–211.

[34] Fang, B. T., "Trilateration and Extension to Global Positioning System Navigation," *Journal of Guidance, Control, and Dynamics*, Vol. 9, No. 6, November/December 1986, pp. 715–717.

[35] Hofmann-Wellenhof, B., et al., *GPS Theory and Practice*, 2nd ed., New York: Springer-Verlag, 1993.

[36] Hewlett-Packard, "Fundamentals of Quartz Oscillators," Application Note 200-2, May 1997.

[37] Lombardi, M.; *Fundamentals of Time and Frequency,* NIST, Boca Raton, FL: CRC Press, 2002.

[38] Vig, John R., "Quartz Crystal Resonators and Oscillators for Frequency Control and Timing Applications—A Tutorial," U.S. Army Communications – Electronics Command, SLCET-TR-88-1 (Rev. 8.2.7b), Fort Monmouth, NJ, April 1999.

[39] Vig, John R., and A. Ballato, "Frequency Control Devices," *Reference for Modern Instrumentation, Techniques, and Technology: Ultrasonic Instruments and Devices II*, editied by Emmanuel P. Papadakis, Vol. 24, Academic Press, 1998, pp. 637–701, 25 vols. Physical Acoustics.

[40] Cantor, S. R., A. Stern, and B. Levy, "Clock Technology," *Institute of Navigation 55th Annual Meeting*, Cambridge, MA, June 28–30, 1999.

[41] Zameroski, N. D. et al., "Pressure Broadening and Frequency Shift of the $5S_{1/2} \rightarrow 5D_{5/2}$ and $5S_{1/2} \rightarrow 7S_{1/2}$ Two Photon Transitions in ^{85}Rb by the Noble Gases and N_2," *Journal of Physics B*, Vol. 47, No. 22, 2014, p. 225205.

[42] Emmer, W., "Atomic Frequency Standards for the GPS IIF Satellites," *Proceedings from the 29th Annual Precise Time and Time Interval (PTTI) Meeting*, 1997, p. 201.

[43] Taylor, J., "AEP Goes Operational: GPS Control Segment Upgrade Details," *GPS World*, Vol. 19, No. 6, 2008, p. 27.

[44] Heavner, T. P., "First Accuracy Evaluation of NIST-F2," *Metrologia*, Vol. 51, No. 3, 2014.

[45] Bloom, B. J., "An Optical Lattice Clock with Accuracy and Stability the 10^{-18} Level," *Nature*, Vol. 506, 2014, p. 71.

[46] Hinkley, N., "An Atomic Clock with 10^{-18} Instability," *Science*, Vol. 22, 2013.

[47] Chou, C. W., "Frequency Comparison of Two High-Accuracy Al+ Optical Clocks," *Physical Review Letters*, Vol. 104, 2010, p. 070802.

[48] Lewandowski, W., and E. Arias, "GNSS Times and UTC," *Metrologia*, Vol. 48, July 20, 2011, pp. S219–S224.

[49] Bureau International des Poids et Mesures, *The International System of Units (SI)*, 8th ed., Sèvres, France: Organisation Intergouvernementale de la Convention du Mètre, 2006.

[50] Kaplan, E., and Hegarty, C., *Understanding GPS: Principles and Applications*, 2nd ed., Norwood, MA: Artech House, 2006.

CHAPTER 3

Global Positioning System

Arthur J. Dorsey, Willard A. Marquis, John W. Betz, Christopher J. Hegarty, Elliott D. Kaplan, Phillip W. Ward, Michael S. Pavloff, Peter M. Fyfe, Dennis Milbert, and Lawrence F. Wiederholt

3.1 Overview

GPS is comprised of three segments: satellite constellation, ground control/monitoring network, and user receiving equipment. The formal United States Air Force (USAF) GPS Directorate programmatic terms for these components are space, control, and user equipment segments, respectively. The satellite constellation is the set of satellites in orbit that provide the ranging signals and data messages to the user equipment. The control segment (CS) tracks and maintains the satellites in space. The CS also monitors satellite health and signal integrity and maintains the orbital configuration of the satellites. Furthermore, the CS updates the satellite clock corrections and ephemerides as well as numerous other parameters essential to determining user position, velocity, and time (PVT). The user receiver equipment (i.e., user segment) performs the navigation, timing, or other related functions (e.g., surveying). An overview of each system segment is provided next followed by further elaboration on each segment starting in Section 3.2.

3.1.1 Space Segment Overview

The space segment is the constellation of satellites from which users make ranging measurements. The space vehicles (SVs) (i.e., satellites) transmit pseudorandom noise (PRN)-coded signals from which the ranging measurements are made. This concept makes Global Positioning System (GPS) a passive system for the user with signals only being transmitted and the user passively receiving the signals. Thus, an unlimited number of users can simultaneously use GPS. A satellite's transmitted ranging signal is modulated with data that includes information that defines the position of the satellite. An SV includes payloads and vehicle control subsystems. The primary payload is the navigation payload used to support the GPS PVT mission, and the secondary payload is the nuclear detonation (NUDET) detection system, which supports detection and reporting of Earth-based radiation phenomena.

The vehicle control subsystems perform such functions as maintaining the satellite pointing to Earth and the solar panels pointing to the Sun.

3.1.2 Control Segment Overview

The CS has responsibility for maintaining the satellites and their proper functioning. This includes maintaining the satellites in their proper orbital positions (called stationkeeping) and monitoring satellite subsystem health and status. The CS also monitors the satellite solar arrays, battery power levels, and propellant levels used for maneuvers. Furthermore, the CS activates spare satellites (if available) to maintain system availability. The CS updates each satellite's clock, ephemeris, and almanac and other indicators in the navigation message at least once per day. Updates are more frequently scheduled when improved navigation accuracies are required. (Frequent clock and ephemeris updates result in reducing the space and control contributions to range measurement error. Further elaboration on the effects of frequent clock and ephemeris updates is provided in Section 3.3.2. Several analyses and studies have shown that users benefit from reduced navigation errors with more frequent uploads, thus reducing the upload age of data and accompanying broadcast navigation message errors [72, 73].)

The ephemeris parameters are a quasi-Keplerian representation of the GPS satellite orbits and are valid only for a time interval of 3 or 4 hours with the once-per-day normal upload schedule. Navigation message data can be stored for at least a 60-day duration with time validity intervals that grow progressively longer but with decreased accuracy in the event that an upload cannot be provided for an extended period. Initially, Block IIR SVs had the requirement of storing 180 + 30 days of navigation data. This requirement has been reduced to 60 days.

The almanac is a reduced precision subset of the ephemeris parameters. Almanac data is used to predict the approximate satellite position and aid in satellite signal acquisition. Furthermore, the CS resolves satellite anomalies, and collects pseudorange and carrier phase measurements at the remote monitor stations to determine satellite clock corrections, almanac, and ephemeris. To accomplish the above functions, the CS is comprised of three different physical components: the master control station (MCS), monitor stations, and the ground antennas, each of which is described in more detail in Section 3.3.

3.1.3 User Segment Overview

The user receiving equipment comprises the user segment. Each set of equipment is typically referred to as a GPS receiver, which processes the L-band signals transmitted from the satellites to determine user PVT. While PVT determination is the most common use, receivers are designed for other applications such as computing user platform attitude (i.e., heading, pitch, and roll) or as a timing source. Section 3.4 provides further discussion on the user segment.

3.2 Space Segment Description

The space segment has two principal aspects. One aspect is the constellation of satellites in terms of the orbits and positioning within the orbits. The other aspect is the features of the satellites that occupy each orbital slot. Each aspect is described next.

3.2.1 GPS Satellite Constellation Description

The nominal GPS constellation consists officially of 24 satellites in 6 MEO orbital planes, known as the baseline 24-slot constellation. For many years, the U.S. Air Force has been operating the constellation with more than the baseline number of satellites. In June 2011, the U.S. Air Force formalized this by introducing the concept of an expandable 24-slot constellation, in which 3 of the 24 baseline orbital slots are expanded to contain two satellites. That is, in each of 3 expanded orbital slots, 2 satellites are inserted, yielding an expanded GPS constellation size of up to 27 satellites. This reconfiguration resulted in improved coverage and geometric properties in most parts of the world [1]. (Section 11.2.1 discusses geometric properties.) Additional satellites (beyond 27) are typically located next to satellites that are expected to need replacement in the near future.

Within the baseline 24-slot GPS constellation, the satellites are positioned in six Earth-centered orbital planes with four satellites in each plane. The nominal orbital period of a GPS satellite is one-half of a sidereal day or 11 hours 58 minutes [2], yielding an orbital radius (i.e., nominal distance from the center of mass of the Earth to the satellite) of approximately 26,600 km. The orbits are nearly circular with a nominal inclination relative to the equatorial plane of 55°, and the orbital planes are equally spaced about the equator at a 60° separation. This satellite constellation provides a continuous global user navigation and time determination capability.

Figure 3.1 depicts a view from space of the baseline 24-slot GPS constellation, while Figure 3.2 shows the satellite orbits in a planar projection referenced to the epoch time of 0000 h 1 July 1993 UTC(USNO). Thinking of an orbit as a ring, Figure 3.2 opens each orbit and lays it flat on a plane. Similarly, for the Earth's equator, it is like a ring that has been opened and laid on a flat surface. The slope of each orbit represents its inclination with respect to the Earth's equatorial plane, which is nominally 55°. Also depicted in Figure 3.2 are the three orbital slots that form the basis of the expandable constellation. Note that two satellites in expanded slots (shown in white in Figure 3.2) replace the original single baseline slot.

The orbital plane locations with respect to the Earth are defined by the right ascension of the ascending node (RAAN), while the location of the satellite within the orbital plane is defined by the argument of latitude. The RAAN is the point of intersection of each satellite orbit when the satellite is traveling northward with the equatorial plane in inertial space.

The orbital slot assignments of the baseline and expandable 24-slot GPS constellations are contained in [3] and are provided in Tables 3.1 and 3.2. Tables 3.1 and 3.3 define the nominal, properly geometrically spaced, baseline 24-slot constellation for GPS. Slots for the expandable constellation are noted with an asterisk

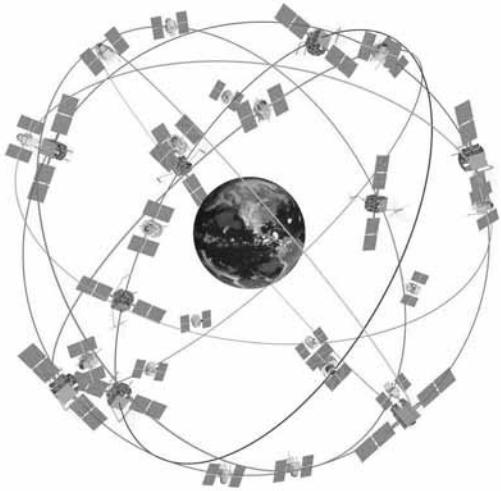


Figure 3.1 Nominal GPS satellite constellation. (Source: Lockheed Martin Corp. Reprinted with permission.)

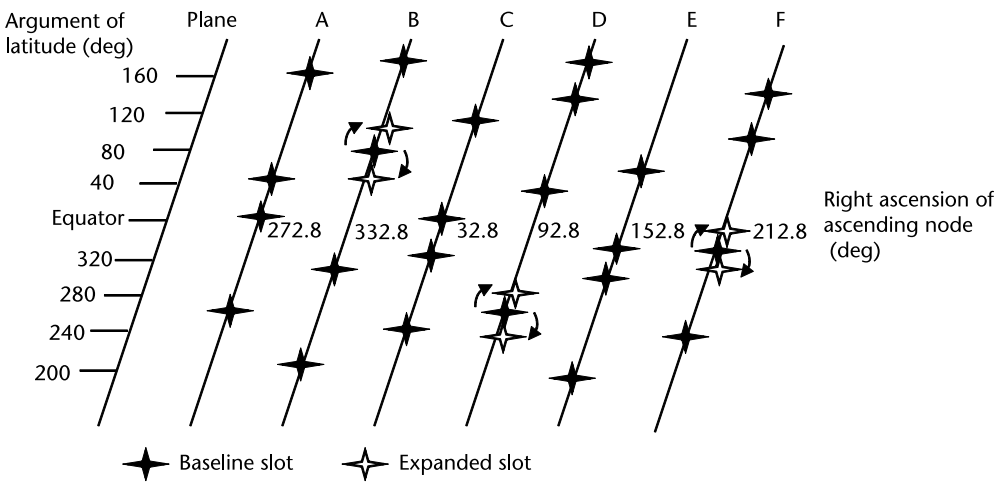


Figure 3.2 GPS constellation planar projection.

in the table, and at the bottom, the parameters in the expanded configuration are shown.

Several different notations are used to refer to the satellites in their orbits. One nomenclature assigns a letter to each orbital plane (i.e., A, B, C, D, E, and F) with each satellite within a plane assigned a number from 1 to 4. Thus, a satellite referenced as B3 refers to satellite number 3 in orbital plane B. As shown in Table 3.2, the B1, D2, and F2 slots are expandable. When the slots are expanded, an "F" or "A" is appended to the letter/number designator to denote whether the satellite is in the fore or aft expanded slot (e.g., B1F is ahead of B1A when the B1 slot is expanded). A second notation used is a NAVSTAR satellite number assigned by the U.S. Air Force. This notation is in the form of space vehicle number (SVN) 60 to refer to NAVSTAR satellite 60. The third notation represents the configuration of the PRN code generators onboard the satellite. These PRN code generators are

Table 3.1 Baseline 24-Slot Constellation Slot Assignments as of the Defined Epoch [3]

Slot	*RAAN (deg)*	*Argument of Latitude (deg)*	*Slot*	*RAAN (deg)*	*Argument of Latitude (deg)*
A1	272.847	268.126	D1	92.847	135.226
A2	272.847	161.786	D2*	92.847	265.446
A3	272.847	11.676	D3	92.847	35.136
A4	272.847	41.806	D4	92.847	167.356
B1*	332.847	80.956	E1	152.847	197.046
B2	332.847	173.336	E2	152.847	302.596
B3	332.847	309.976	E3	152.847	66.066
B4	332.847	204.376	E4	152.847	333.686
C1	32.847	111.876	F1	212.847	238.886
C2	32.847	11.796	F2*	212.847	345.226
C3	32.847	339.666	F3	212.847	105.206
C4	32.847	241.556	F4	212.847	135.346

Source: [3].

Table 3.2 Expandable 24-Slot Constellation Slot Assignments as of the Defined Epoch [3]

Expandable Slot	*Expanded Slot*	*RAAN*	*Argument of Latitude*
B1 expands to:	B1F	332.847	94.916
	B1A	332.847	66.356
D2 expands to:	D2F	92.847	282.676
	D2A	92.847	257.976
F2 expands to:	F2F	212.847	0.456
	F2A	212.847	334.016

Source: [3]. RAAN = right ascension of the ascending node, argument of latitude = geodetic latitude at the given epoch, coordinate system reference = Fundamental Katalog (FK)5/J2000.00, and epoch: 0000Z, 1 July 1993; Greenwich Hour Angle: 18 hours 36 minutes 14.4 seconds.

Table 3.3 Reference Orbit Parameters

Reference Orbit Parameter	*Nominal Value*	*Operational Range*	*Required Tolerance*
Semimajor axis (km)	26,559.7	Note 1	Note 2
Eccentricity	0.0	0.0 to 0.02	0.0 to 0.03
Inclination (°)	55.0	±3	N/A
RAAN (°)	Note 3	±180	N/A
Argument of perigee (°)	0.0	±180	N/A
Argument of latitude at epoch (°)	Note 3	±180	Note 1

Source: [3]. Note 1: The semi-major axis and orbital period will be adjusted to maintain the relative spacing of the satellite mean arguments of latitude to within 4° of the epoch values, with one year or more between orbit adjustments. Note 2: The nominal value shown provides stationary ground tracks.

configured uniquely on each satellite, thereby producing unique versions of its navigation broadcast signals. Thus, a satellite can be identified by the PRN codes that it generates. Occasionally, the PRN assignment for a given SVN can change during the satellite's mission duration. (GPS satellite signals are described in Section 3.7.)

3.2.2 Constellation Design Guidelines

This section provides a basic overview of the constraints and considerations leading to the selection of the nominal GPS constellation, known as the baseline 24-slot constellation. Section 3.2.2.1 details the main considerations leading to the original 24-slot constellation, and then Section 3.2.2.2 presents the ability to add up to three satellites to the baseline constellation in the configuration known as the expandable 24-slot constellation.

As discussed in Section 2.3.2, several considerations are involved in the design of the GPS constellation. One primary optimization parameter is the geometric contribution to navigation accuracy; the constellation must be designed to ensure the satellite geometry is sufficiently diverse to provide good observability to users throughout the world. This geometry is measured by a parameter called dilution of precision (DOP) and is described in more detail in Section 11.2.1. Studies have been ongoing for decades concerning trade-offs on different possible satellite configurations. Some studies have investigated the use of 30 satellites in 3 orbital planes as well as the utility of geostationary satellites. Most of this work is done with a nominal constellation assuming that all satellites are healthy and operational, but more sophisticated studies consider satellite failures. Single or multiple satellite failures provide a new dimension around which to optimize performance from a geometry consideration. Another overall design consideration is line-of-sight observability of the satellites by the ground stations to maintain the ephemeris of the satellites and the uploading of this data.

3.2.2.1 Baseline GPS Constellation

This section presents the main trade-offs leading to the selection of the baseline 24-slot GPS constellation. We refer to the seven constellation design considerations presented in Section 2.3.2.3 in this discussion of the trade-offs leading to the baseline GPS constellation.

The need for global coverage and the need for good and changing geometric diversity worldwide eliminate the use of geostationary satellites for navigation, although a constellation of geosynchronous satellites with enough inclination could theoretically be used to provide global coverage including the poles. One factor weighing against the use of an inclined GEO constellation to provide global coverage for navigation includes consideration regarding increased satellite power (and thus payload weight) required from GEO to provide the necessary power flux density at the surface of the Earth relative to satellites at lower altitudes. Another factor weighing against the use of inclined GEOs for satellite navigation is the regulatory coordination issue associated with GEO orbits. Thus, the constraint of global coverage, geometric diversity, and practical considerations drove the GPS satellite navigation constellation to inclined LEO or MEO orbits.

Constraint for minimum sixfold coverage, plus the need to minimize the size of the constellation for cost reasons, drove the desired GPS constellation to higher altitude. With satellite costs in excess of $100 million [4], even for small satellites like GPS, the differences in constellation size drive the desired altitude to MEO. To the first approximation, an order of magnitude more satellites would be required to provide the necessary sixfold coverage from LEO versus MEO, which, when launch costs are factored in, drives the overall cost differential between LEO and MEO to be billions of dollars. Moreover, constellations of LEO satellites tend to have worse geometric properties from a dilution of precision perspective than MEOs. With LEO and GEO altitudes shown to be undesirable, MEO altitudes were determined to be preferable for GPS. Ultimately, inclined orbits were selected for GPS with approximately 12-hour periods. This was seen as the best compromise between coverage, DOP characteristics, and cost. The exact nominal orbital altitude selected was 20,182 km (orbital radius of 26,560 km), which results in an orbital period of one half the sidereal day. Some desirable characteristics of this orbital altitude include daily repeating ground tracks, a relatively high altitude, which, in turn, produces good DOP properties, and a relatively low number of satellites required to provide the redundancy of coverage required for navigation. It is true that stationkeeping is more frequent at the GPS 12-hour orbital altitude than other potential altitudes in the 20,000–25,000-km range due to the resonance issue discussed is Section 2.3.2.1, and so other satellite navigation architectures, such as that for Galileo, address consideration for constellation design and make slight modifications to the exact orbital altitude of the MEO constellation. (Galileo is discussed in Chapter 5.)

The robustness considerations drove the desire for multiple satellites per orbital plane versus a more generalized Walker-type constellation that could provide the same level of coverage with fewer satellites but in separate orbital planes (see the discussion at the end of Section 2.3.2.2). Ultimately, a 6-plane configuration was selected with 4 satellites per plane for GPS. The orbital planes are inclined by 55°, in accordance with Walker's results. The planes are equally spaced by 60° in right ascension of the ascending node around the Equator. Satellites are not equally spaced within the planes, and there are phase offsets between planes to achieve improved DOP characteristics of the constellation when probable failures are considered. Some design choices were also influenced by historical constraints that are no longer relevant. For instance, the space shuttle was originally planned to be used to deploy the GPS constellation until the *Challenger* disaster in 1986. Hence, the GPS constellation can be considered a tailored Walker constellation.

3.2.2.2 Expandable GPS Constellation

Even though the GPS baseline constellation consists of 24 orbital slots, the U.S. Air Force maintains more than 24 satellites on orbit today. This was formalized in 2011 by the definition of the expandable 24-slot constellation shown in Tables 3.1 and 3.2 [1, 3]. The additional 3 slots were added to three alternating orbital planes (B, D, and F). One of the four baseline slots in those three planes is expandable to two slots, one fore and one aft of the baseline slot. The idea is that the U.S. Air Force could add one, two, or three satellites to the baseline constellation simply by relo-

cating a baseline satellite within that plane to a nearby slot and adding a satellite in that plane. The expansion slots are depicted graphically in Figure 3.2.

Today, the U.S. Air Force flies even more than the 27 satellites in the expandable constellation configuration. At the time of this writing, there are 31 GPS satellites operational on-orbit. The greater number of satellites on orbit provides greater accuracy and robustness of the constellation. This has been enabled by two facts. The GPS satellites have lived much longer than their design life, and the U.S. government has worked hard to maintain its stringent commitments to the world regarding the size of the GPS constellation. In particular, the U.S. government is committed to maintaining 24 satellites with 95% confidence, and 21 specific orbital slots with 98% confidence, and both commitments are formally documented in the Standard Positioning Service Performance Standard and the Precise Positioning Service Performance Standard [3, 5]. In order to keep these commitments to the world, the U.S. government maintains a high assurance approach when scheduling satellite acquisitions and launches.

3.2.3 Space Segment Phased Development

The development of the control and space segments has been phased in over many years, starting in the mid-1970s, and is continuing. This development started with a concept validation phase and has progressed to several production phases. The satellites associated with each phase of development are called a block of satellites. Characteristics of each phase and block are presented in the following sections.

3.2.3.1 Satellite Block Development

Seven satellite blocks have been developed to date. The initial concept validation satellites were called Block I. The last remaining prototype Block I satellite was disposed of in the fall of 1995. Block II satellites were the initial production satellites while Block IIA refers to upgraded production satellites. With the exception of one Block I launch failure, all Block I, II, and IIA satellites were launched and decommissioned. Block IIR satellites, denoted as the replenishment satellites, have been deployed. Modernized Block IIR versions denoted as Block IIR-M have also been launched. Block IIF satellites, referred to as the follow-on or sustainment satellites are also on orbit. At the time of this writing, the first GPS III satellite is planned for launch in the 2018 timeframe [6]. Since satellites are launched only as replacements for satellite failures, their scheduling is difficult to predict, especially when most satellites have far out lived their design lifetime. Also at the time of this writing, the constellation consisted of 31 operational satellites [7]. Table 3.4 describes the configuration of the current satellite constellation. Thus, the current optimized constellation has up to 7 orbital slots unevenly spaced around each plane with some satellites in relatively close proximity to provide redundant coverage for near-term predicted failures (i.e., expanded residual/test, or auxiliary orbital slots) [8]:

> The term "residual/test" status is defined as a satellite that is partially functional but the signals are not part of the PNT solution. A satellite in an auxiliary slot broadcasts GPS signals but not in one of the 24 primary slots. Most of the auxiliary satellites are available for users full-time but are "paired" with other satellites with

Table 3.4 Satellite Constellation Configuration as of August 2016

SVN	*PRN*	*Launch Date*	*Usable Date*	*Orbital Slot*
Type: Block IIR				
43	13	July 23, 1997	January 31, 1998	F2F
46	11	October 7, 1999	January 3, 2000	D2F
51	20	May 11, 2000	June 1, 2000	E7
44	28	July 16, 2000	August 17, 2000	B3
41	14	November 10, 2000	December 10, 2000	F1
54	18	January 30, 2001	February 15, 2001	E4
56	16	January 29, 2003	February 18, 2003	B1A
45	21	March 31, 2003	April 12, 2003	D3
47	22	December 21, 2003	January 12, 2004	E6
59	19	March 20, 2004	April 5, 2004	C5
60	23	June 23, 2004	July 9, 2004	F4
61	02	November 6, 2004	November 22, 2004	D1
Type: Block IIR-M				
53	17	September 26, 2005	December 16, 2005	C4
52	31	September 25, 2006	October 12, 2006	A2
58	12	November 17, 2006	December 13, 2006	B4
55	15	October 17, 2007	October 31, 2007	F2A
57	29	December 20, 2007	January 2, 2008	C1
48	07	March 15, 2008	March 24, 2008	A4
50	05	August 17, 2009	August 27, 2009	E3
Type: Block IIF				
62	25	May 28, 2010	August 27, 2010	B2
63	01	July 16, 2011	October 14, 2011	D2A
65	24	October 4, 2012	November 14, 2012	A1
66	27	May 15, 2013	June 21, 2013	C2
64	30	February 21, 2014	May 30, 2014	A3
67	06	May 17, 2014	June 10, 2014	D4
68	09	August 2, 2014	September 17, 2014	F3
69	03	October 29, 2014	December 12, 2014	E1
71	26	March 25, 2015	April 20, 2015	B1F
72	08	July 15, 2015	August 12, 2015	C3
73	10	October 31, 2015	December 9, 2015	E2
70	32	February 5, 2016	March 9, 2016	F5

Source: [9].

known failed components to minimize the user impact if one satellite in the pair fails unexpectedly. A satellite is maintained as an on-orbit auxiliary only if it can still provide users the expected level of performance, does not degrade the overall constellation, and is not in danger of failing in such a manner as to prevent proper disposal.

Since the state of the constellation varies, the Internet is the best source for current status information. One such Web site is operated and maintained by the U.S. Coast Guard Navigation Center [7].

3.2.3.2 Navigation Payload Overview

The navigation payload is responsible for the generation and transmission of ranging codes and navigation data on the L1, L2, and (starting with Block IIF) L5 carrier frequencies to the control and user segments. Control of the navigation payload is via reception of the data from the CS via the tracking, telemetry and control (TT&C) links. The navigation payload is only one part of the spacecraft with other systems being responsible for such functions as attitude control and solar panel pointing. Figure 3.3 is a generic block diagram of a navigation payload. Atomic frequency standards (AFSs) are used as the basis for generating the extremely stable ranging codes and carrier frequencies transmitted by the payload. Each satellite contains multiple AFSs to meet the mission reliability requirements, with only one AFS operating at any time. Since the AFSs operate at their natural frequencies, a frequency synthesizer, phase-locked to the AFS, generates the basic 10.23-MHz reference that serves as the timing reference within the payload for ranging signal and transmit-frequency generation. (Note that the actual generated reference frequency is adjusted to compensate for relativistic effects. This is discussed in Section 10.2.3.) The navigation data unit (NDU) in the Block IIF, known as the mission data unit (MDU) in the Block IIR, IIR-M and GPS III designs, contains the ranging code generators that generate the signals listed in Table 3.5. (Details of each ranging code and navigation message are provided in Section 3.7.) The NDU/MDU also contains a processor that stores the uploads received from the CS containing multiple days of navigation message data, and assures that the current issue of navigation message data is provided. The combined baseband ranging codes are then sent to the L-band subsystem where they are modulated onto the L-band carrier frequencies and amplified for transmission to the user. The L-band subsystem contains numerous components including the L1, L2, and L5 (Block IIF and GPS III only) transmitters

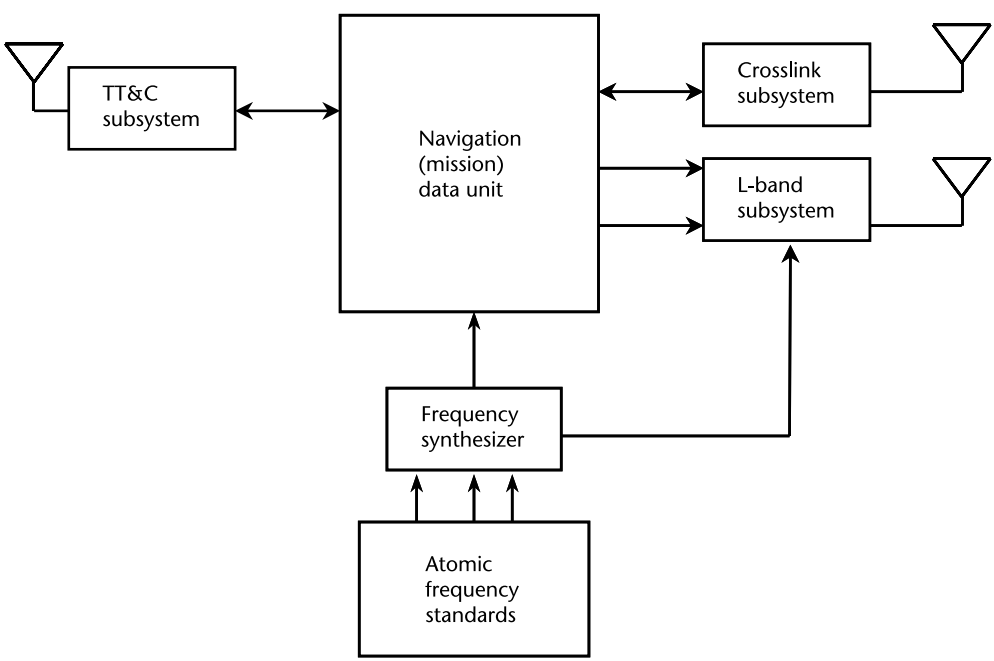


Figure 3.3 Satellite navigation payload.

Table 3.5 Satellite Block Ranging Signals and Associated Navigation Data Type

Satellite Type	*Ranging Signals*	*Navigation Data Type*
Block IIR	L1: C/A, P(Y); L2: P(Y)	Legacy
Block IIR-M	L1: C/A, P(Y), M; L2: L2C, P(Y), M	Legacy: C/A, P(Y); MNAV: M; CNAV: L2C
Block IIF	L1: C/A, P(Y), M; L2: L2C, P(Y), M; L5	Legacy: C/A, P(Y); MNAV: M; CNAV: L2C, L5
GPS III	L1: C/A, P(Y), M, L1C; L2: L2C, P(Y), M; L5	Legacy: C/A, P(Y); MNAV: M; CNAV: L2C, L5; CNAV-2: L1C

and associated antennas. The NDU/MDU processor also interfaces to the crosslink receiver/transmitter for intersatellite communication as well as ranging on Block IIR, IIR-M, and IIF SVs. This crosslink receiver/transmitter uses a separate antenna and feed system. (It should be noted that the intersatellite ranging is functional on these SVs (this capability is denoted as AutoNav); however, the U.S. government has chosen not to add this capability to the CS.) As stated above, the primary and secondary SV payloads are navigation and NUDET, respectively. Occasionally, two of the Block IIA satellites had carried additional payloads such as laser reflectors for satellite laser ranging (i.e., validation of predicted ephemeris), and free electron measurement experiments. The U.S. Air Force is planning to add both laser reflectors and a Distress Alerting Satellite System (DASS) payload to GPS III SVs 11+ [10].

3.2.3.3 Block I Initial Concept Validation Satellites

Block I satellites were developmental prototypes to validate the initial GPS concept. These satellites demonstrated that navigation was possible from moving pseudorange transmitters, atomic clocks could operate in space, momentum management could be done with magnets, thermal control could be accomplished by yaw steering with reaction wheels, and satellites can fly in the harsh radiation environment (i.e., Van Allen belts) of MEO. Ten were launched in all. The Block I satellites, built by Rockwell International, were launched between 1978 and 1985 from Vandenberg Air Force Base, California. The onboard storage capability supported about 3.5 days of navigation messages. The navigation message data was transmitted for a 1-hour period and was valid for an additional 3 hours. All of the Block I satellites carried three rubidium atomic frequency standards (AFSs) and from the fourth SV each also carried one cesium AFS. These satellites were designed for a mean mission duration (MMD) of 4.5 years, a design life of 5 years and inventory expendable (e.g. fuel, battery life, and solar panel power capacity) of 7 years. AFS failures were common on the first satellites requiring that a second source vendor be developed. This also caused the mindset that two different AFS technologies needed to be flown to ensure mission success. Another discovery was that the onboard memory in which the navigation message data were stored was very susceptible to single event upsets caused by ionized radiation particle strikes. At that time, all operational crews were trained to recognize this condition and quickly correct it by uploading a new navigation data to the satellite. Some Block I satellites operated for more than double their design life. A picture of a Block I satellite is presented in Figure 3.4.

Figure 3.4 Block I satellite.

3.2.3.4 Block II-Initial Production Satellites

On-orbit operation of the Block I satellites provided valuable experience that led to several significant capability enhancements in subsystem design for the Block II operational satellites (see Figure 3.5). These improvements included radiation hardening to prevent random memory upset from events such as cosmic rays to improve reliability and survivability. Besides these enhancements, several other refinements were incorporated to support the fully operational GPS system requirements. While most of the changes affected only the CS/space interface, some also affected the user signal interface. The significant changes are identified as the following. To provide security, selective availability (SA) and antispoofing (AS) capabilities were added. (SA was discontinued on May 1, 2000. The United States has no intention to use SA again [1].). System integrity was improved by the addition of automatic

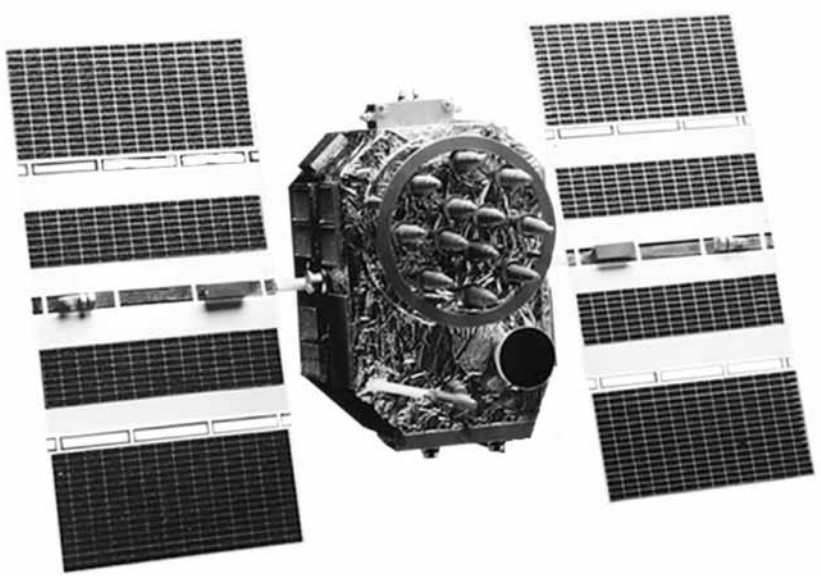


Figure 3.5 Block II satellite.

error detection for certain error conditions. After detection of these error conditions, there is a changeover to the transmission of a nonstandard PRN code to prevent the usage of a corrupted signal or data. Nine Block II satellites were built by Rockwell International, and the first was launched in February 1989 from Cape Canaveral Air Force Station, Florida. The onboard navigation message storage capacity was sized for a 14-day mission. Autonomous onboard momentum control was implemented in the satellite within the attitude and velocity control system, thus eliminating the need for ground contact to perform momentum dumping. Two cesium and two rubidium AFSs were onboard. These satellites were designed for a mean mission duration (MMD) of 6 years, a design life of 7.5 years, and inventory expendables (e.g., fuel, battery life, and solar panel power capacity) of 10 years. The last Block II satellite, SVN 15, was removed from service in August 2006 after 15.84 years of operational service. The Block II average life was 11.92 years.

3.2.3.5 Block IIA-Upgraded Production Satellites

The Block IIA satellites were very similar to the Block II satellites, but with a number of system enhancements to allow an extended operation period out to 180 days (see Figure 3.6). The onboard navigation data storage capability was tested to assure retention for the 180-day period. For approximately the first day on-orbit, the navigation message data was broadcast for a 2-hour period and was valid over a 4-hour interval. For the remainder of the first 14 days, the navigation message was broadcast for a 4-hour period with a validity period of 6 hours (two additional hours). Following this initial 14-day period, the navigation message data broadcast periods had gradually extended from 6 hours to 144 hours. With this additional onboard storage retention capability, the satellites could function continuously for a

Figure 3.6 Block IIA satellite.

period of six months without ground contact. However, the accuracy of the Operational Control System (OCS) ephemeris and clock predictions and thus the accuracy of the navigation message data would have gracefully degraded over time such that the position error would have grown to 10,000-m spherical error probable (SEP) at 180 days. With no general onboard processing capability, no updates to stored reference ephemeris data were possible. So, as a result, full system accuracy was only available when the OCS was functioning properly and navigation messages were uploaded on a daily basis. Block IIA electronics were radiation hardened. Nineteen Block IIA satellites were built by Rockwell International with the first launched in November 1990 from Cape Canaveral Air Force Station, Florida, and the last launched in November 1997. The life expectancy of the Block IIA was the same as that of the Block II. The last operational Block IIA satellite, SVN 23, was decommissioned on January 25, 2016, after an operational life of over 25 years. The URE performance for the GPS II/IIA averaged approximately 1.1m or better for several years easily surpassing the requirement of 6m. The average life of the Block IIA satellites was 17.3 years.

3.2.3.6 Block IIR—Replenishment Satellites

The GPS Block IIR (replenishment) and the GPS Block IIR-M (modernized replenishment) satellites (Figure 3.7) currently perform as the backbone of the GPS constellation. All 21 IIR SVs have been launched since 1997 (the first Block IIR satellite was lost in a booster failure early that year). Lockheed Martin built and is now supporting the operation of these satellites.

The Block IIR began development following contract award in 1989 as a totally compatible upgrade and replacement to the Block II and Block IIA SVs. All the basic GPS features are supported: L-band broadcast signal with C/A and P(Y) code

Figure 3.7 Artist's concept of the GPS Block IIR satellite. (Source: Lockheed Martin Corp. Reprinted with permission.)

on L1, and P(Y) on L2, ultrahigh frequency (UHF) crosslink capability, attitude determination system to stabilize the SV bus platform, reaction control system to maintain the on-orbit location in the constellation, and sufficient electrical power capacity for the life of the vehicle.

There are two versions of the Block IIR SV. The classic IIR and its AFSs, autonomy, reprogrammability, and improved antenna panel will be described first. The features of the modernized IIR (the IIR-M) will be covered later in this section.

Classic IIR

The baseline (nonmodernized) GPS Block IIR is sometimes called the classic IIR.

The Block IIR satellites are designed for a MMD of 6 years, a design life of 7.5 years, and inventory expendables (e.g., fuel, battery life, and solar panel power capacity) of 10 years. As of early 2017, there were 12 IIR and 7 IIR-M SVs in the operational 31-SV constellation, with a twentieth SV in residual status. The oldest IIR SV (SVN 43) was over 19 years old at time of this writing, exceeding its design and expendables life requirements by several factors. This SV continues to perform among the best in the constellation in terms of availability, accuracy, and lifetime [11–13]. See Figure 3.8 for main IIR SV components.

Next-Generation Atomic Frequency Standards

All IIR SVs contain three next-generation rubidium AFSs (RAFS). The IIR design has a significantly enhanced physics package that improves stability and reliability [14].

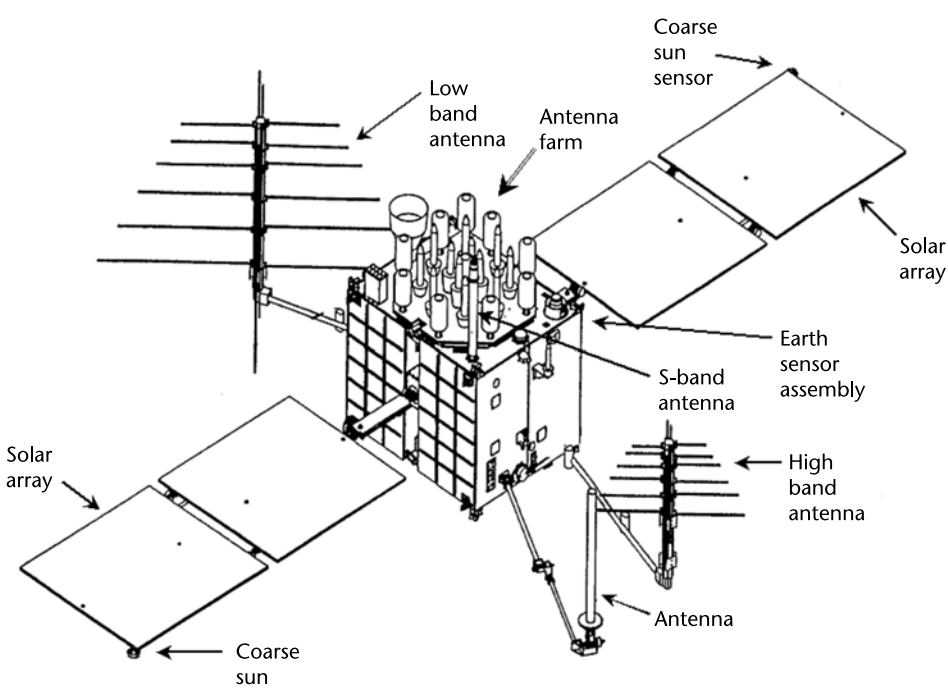


Figure 3.8 Major Block IIR satellite subcomponents. (Source: Lockheed Martin Corp. Reprinted with permission.)

The RAFS has a MMD of 7.5-year life. It is coupled with a redundant voltage controlled VCXO and software functionality into what is called the time-keeping system (TKS). The TKS loop provides a timing tuning capability to stabilize and control clock performance. To date, only two RAFS have experienced issues requiring activation of a redundant RAFS, although these two units are still available for future use. Thus, 40 spare RAFS are available.

IIR Accuracy

An accurate onboard AFS provides the key to good GPS PVT accuracy [11]. The IIR specification requires that the total IIR user range error (URE) value should be less than 2.2m when operating with a RAFS (URE is the contribution of the pseudorange error from the CS and space segment). The URE performance for GPS IIR has averaged approximately 0.8m or better for several years [15]. Thus, the required specification is easily surpassed.

There is also a significantly improved solar pressure model (by an order of magnitude compared to the original II/IIA model) used in the MCS when computing the orbit of the IIR [16, 17]. This increases the accuracy of the ephemeris modeling on the ground.

Enhanced Autonomy

The advanced capabilities of the Block IIR SV include a redundancy management system called REDMAN, which monitors bus subcomponent functionality and provides for warning and component switching to maintain SV health.

The Block IIR uses nickel hydrogen (NiH2) batteries, which require no reconditioning and accompanying operator burden.

When in Earth eclipse, automatic pointing of the solar array panels is accomplished via an onboard orbit propagation algorithm to enable quiescent reacquisition of the Sun following eclipse exit. This provides a more stable and predictive SV bus platform and orientation for the L-band signal.

Block IIR has an expanded nonstandard code (NSC) capability to protect the user from spurious signals. It is enabled automatically in response to the detection of the most harmful on-orbit RAFS and voltage-controlled crystal oscillator (VCXO) discontinuities.

Reprogrammability

There are several reprogrammable computers on board: the redundant SV bus spacecraft processor unit (SPU) and the redundant navigation system mission data unit (MDU). Reprogrammability allows the CS operator to upload flight software changes to on-orbit SVs. This feature has been employed on-orbit in several instances.

The SPU was provided with new rolling buffers to save high-speed telemetry data for SV functions even when not in contact with the CS. On-board software macros can also be triggered by specified telemetry behavior to handle additional autonomous reconfigurations beyond the original SV design.

The MDU was provided with diagnostic buffers to give detailed insight into the behavior of the TKS. It was also given a jumpstart capability allowing current TKS parameters to be saved to a special area of memory and reused following the

load of a new program. This feature reduces by about 4 hours, the time required to recover from a new program load. The MDU software was also upgraded to support IIR modernization, the addition of the selective availability/antispoofing module (SAASM) capability, the addition of on-orbit storage buffers, and numerous other changes.

Improved Antenna Panel

Lockheed Martin, under an internal research and development effort, developed new L-band and UHF antenna element designs [18]. Figure 3.9 presents the average directivity of the new antenna panel broadcast pattern for the L1 signal for all 12 improved antenna panels (SV-specific directivity plots, as well as SV-specific directivity and phase data, are available online at http://www.lockheedmartin.com/us/products/gps/gps-publications.html). The new L1 power received on the ground (Earth service; space service is the signal beyond the edge of Earth) is at least –154.5 dBW (edge-of-Earth, as compared to the current typical IIR performance of –155.5 dBW) and the new L2 power received on the ground is at least –159.5 dBW (edge-of-Earth, as compared to the current typical IIR performance of –161.5 dBW). This provides greater signal power to the user. The last 4 of the 12 classic IIRs and all of the modernized IIRs have the improved antenna panel.

Block IIR-M-Modernized Replenishment Satellites

Beginning in 2005, the GPS IIR-M initiated new services to military and civilian users [19, 20]. The IIR-M was the result of an effort to bring this modernized

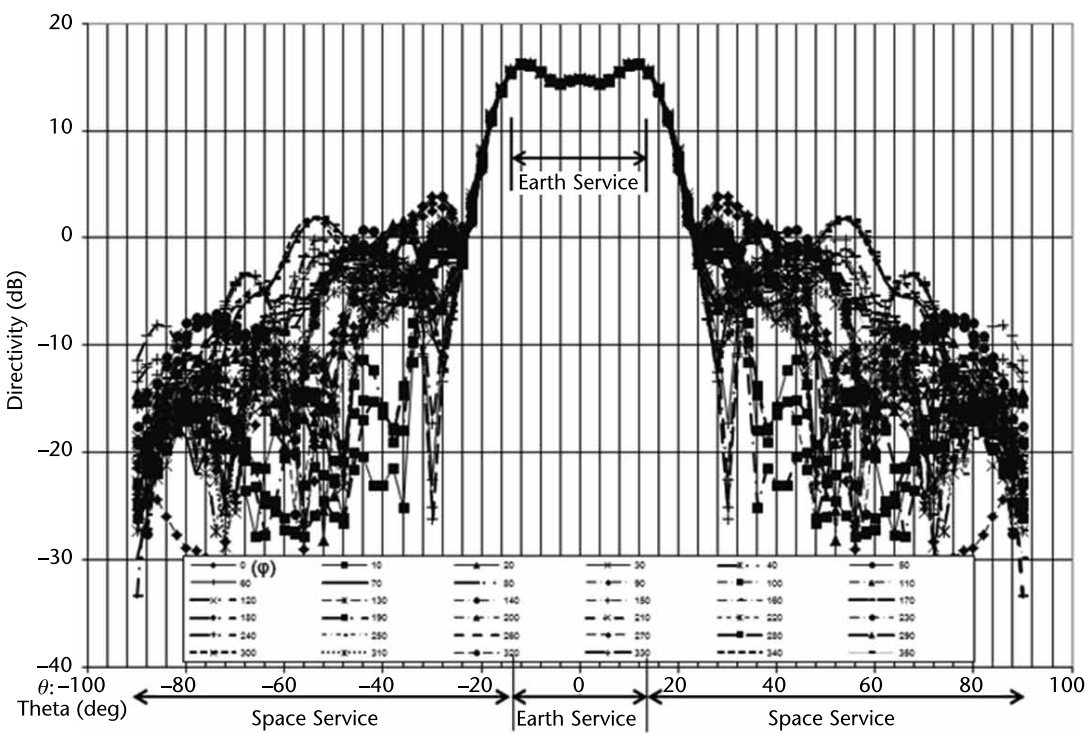


Figure 3.9 Average improved antenna pattern, L1 signal. (Source: Lockheed Martin Corp. Reprinted with permission.)

functionality to IIR SVs that were built years earlier and placed into storage until they were needed for launch. This modernization program was accomplished within existing solar array capability, available on-board processor margins, and available vehicle structural capabilities.

Eight Block IIR SVs were modernized. These SVs were launched between late 2005 and late 2009. Following the launch, the first IIR-M SV began broadcasting the first modernized signals. The first test of the default Civil Navigation (CNAV) message on the new L2C signal was in September 2009 [21, 22]. As the CS has been gradually upgraded, additional testing of signals and data broadcast have been tested. Daily CNAV uploads began in late 2014.

Modernized Signals

The new additional L-band signals and increased L-band power significantly improved navigation performance for users worldwide. Three new signals were provided: two new military codes (denoted as M-code) on L1 and on L2, and a new civilian code on L2. The new L2 civil signal denoted as L2C is an improved signal sequence over L1 C/A, enabling ionospheric error correction to be done by civilian users. The new signal structure is totally backward-compatible with existing L1 C/A and P(Y) and L2 P(Y). M code provides the authorized user with more signal security. (Refer to Section 3.7.2.3 for further details.)

Modernized Hardware

The new navigation panel boxes consist of a redesigned L1 transmitter, a redesigned L2 transmitter, and the new waveform generator/modulator/intermediate power amplifier/dc-dc converter (WGMIC) (Figure 3.10). The WGMIC is a new box coupling the brand-new waveform generator with the functionality of the L1 signal modulator/intermediate power amplifier (IPA), the L2 signal modulator/IPA, and the dc-to-dc converter. The waveform generator provides much of the new modernized signal structure and controls the power settings on the new transmitters. To manage the thermal environment of these higher-power boxes, heat pipes were incorporated into the fabrication of the structural panel. Lockheed Martin has successfully used similar heat pipes on other satellites it has built for this thermal control.

The improved IIR antenna panel discussed earlier in this section was also installed on all 8 IIR-M SVs (refer to Table 3.6 for the various IIR SV and antenna panel versions). This provides greater signal power to the user. The antenna redesign effort was begun prior to the modernization decision, but significantly enhances the new IIR-M features. L-band power was increased on both L1 and L2 frequencies. Received L1 was increased at least double the power, and received L2 was increased by at least quadruple power at lower elevation angles [18].

3.2.3.7 Block IIF-Follow-On Sustainment Satellites

In 1995, the Air Force GPS JPO released a request for proposal (RFP) for a set of satellites to sustain the GPS constellation, designated as Block II Follow-on, or IIF. The RFP also requested the offeror to include the modifications to the GPS control segment necessary to operate the IIF SV. While necessary for service sustainment,

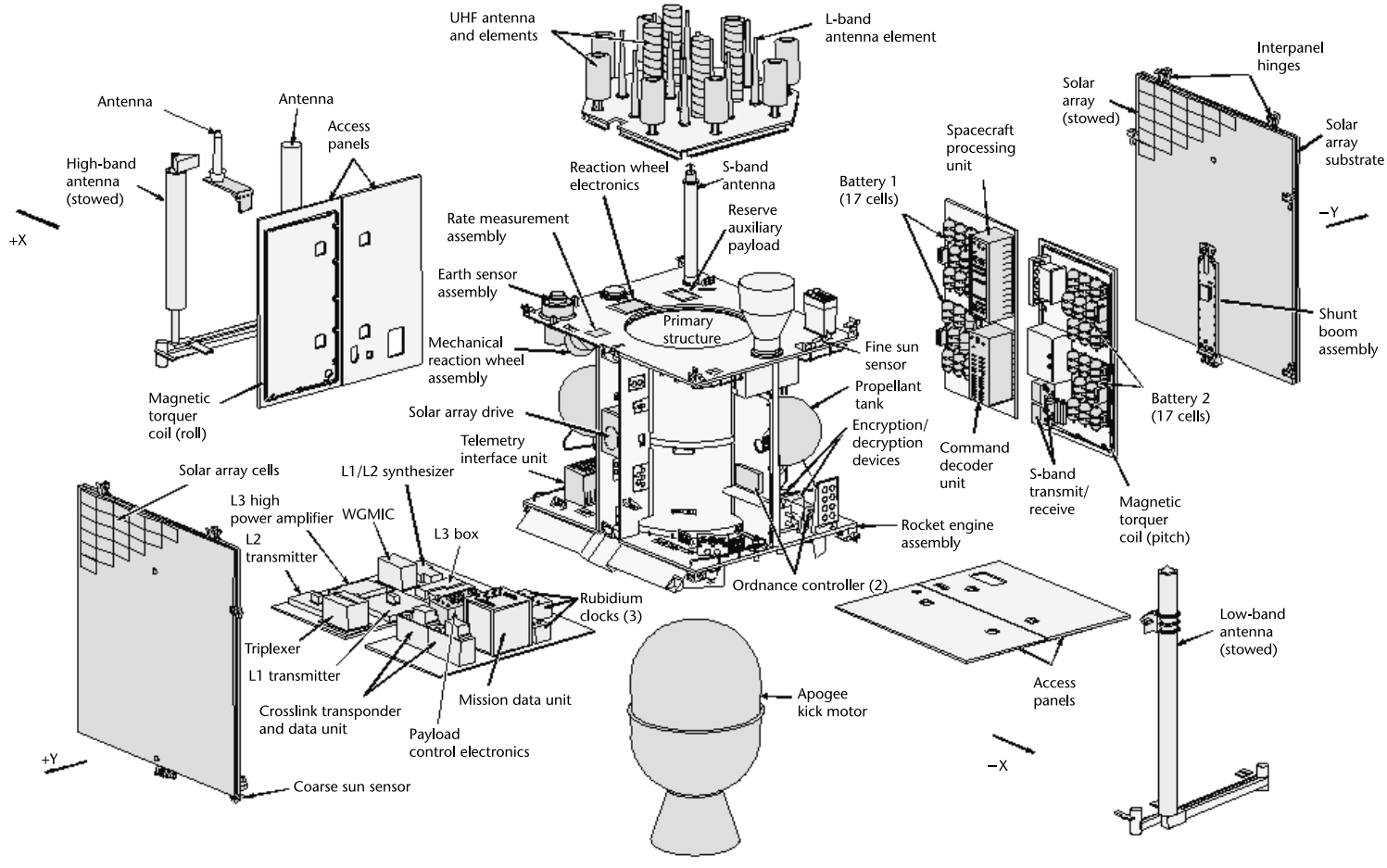


Figure 3.10 Block IIR-M expanded view. (Source: Lockheed Martin Corp. Reprinted with permission.)

Table 3.6 IIR/IIR-M Antenna Versions

	SV Type		*Antenna Panel Type*	
SVN (Launch Order)	*Classic IIR SV*	*IIR-M SV*	*Legacy Antenna Panel*	*Improved Antenna Panel*
43	X	—	X	—
46	X	—	X	—
51	X	—	X	—
44	X	—	X	—
41	X	—	X	—
54	X	—	X	—
56	X	—	X	—
45	X	—	X	—
47	X	—	—	X
59	X	—	—	X
60	X	—	—	X
61	X	—	—	X
53	—	X	—	X
52	—	X	—	X
58	—	X	—	X
55	—	X	—	X
57	—	X	—	X
48	—	X	—	X
49	—	X	—	X
50	—	X	—	X

the IIF SV procurement afforded the Air Force the opportunity to start adding new signals and additional flexibility to the system beyond the capabilities and improvements of the IIR SV. A new military acquisition code on L2 was required, as well as an option for a new civil L5 signal at a frequency within 102.3 MHz of the existing L2 frequency of 1227.6 MHz. The L5 frequency that was eventually settled upon was 1,176.45 MHz, placing it in a frequency band that is protected for Aeronautical Radionavigation Services. (L5 signal characteristics are described in Section 3.7.2.2.)

The RFP also allowed the offeror to provide additional best-value features that could be considered during the proposal evaluation. Boeing (then Rockwell) included several best value features in its proposal and was awarded the IIF contract in April 1996. Several of these features were to improve service performance, including a URE 3m or less in AutoNav mode, an age of data for the URE of less than 3 hours using the UHF crosslink to update the navigation message, and design goals for AFS Allan variance performance better than specification. Other features supported the addition of auxiliary payloads on the IIF SV and reduction of operational complexity for the operators via greater use of the UHF crosslink communication system.

The original planned launch date for the first IIF SV was April 2001. However, due to the longevity of the Block II and IIA SVs and projected service life of the IIR SVs, the need date for a IIF launch was extended sufficiently to allow the Air Force to direct modifications to the IIF SV that resulted in the present design. The first

modification was enabled when the Delta II launch vehicle (LV) was deselected for IIF, leaving the larger Evolved Expendable Launch Vehicle (EELV) as the primary LV. The larger fairing of the EELV enabled the "Big Bird" modification to the IIF SV, which expanded the spacecraft volume, nadir surface area, power generation and thermal dissipation capability. Around the same time, extensive studies were performed by the GPS Modernization Signal Development Team (GMSDT) to evaluate new capabilities needed from GPS, primarily to add new military and civil ranging signals. The GMSDT was formed as a Government/FFRDC/Industry team to evaluate the deficiencies of the existing signal structure and recommend a new signal structure that would address the key areas of modulation and signal acquisition, security, data message structure, and system implementation. Today's M-code signal structure is the result of those studies. The complete list of ranging signals provided by the IIF SV is shown in Table 3.7. (Details of all GPS signals are provided in Section 3.7.) It should be noted that the new ranging signals also carry improved versions of the SV clock offset and ephemeris data in their respective navigation messages to eliminate some of the resolution limitations the original navigation message had imposed as the URE has continued to improve.

The original flexibility and expandability features of the IIF SV in both the spacecraft and navigation payload designs allowed the addition of these new signals without major revisions to the IIF design. An expanded view of the Block IIF SV is depicted in Figure 3.11. The figure shows all the components of the spacecraft subsystems such as: attitude determination and control subsystem which keeps the antennas pointing at the Earth and the solar panels at the Sun; the electrical power subsystem that generates, regulates, stores and shunts the dc power for the satellite; and the TT&C subsystem, which allows the MCS operators to communicate with and control the satellite on-orbit. To support the increase in dc power requirement due to the increased transmit power, the solar arrays were switched from silicon technology to higher efficiency triple-junction gallium arsenide. Additionally, the thermal design had to be revised to accommodate the additional transmitter thermal loads. Other than some realignment to maintain weight and thermal balance, no other modifications were required for the spacecraft.

The navigation payload on the Block IIF SV includes two rubidium AFSs and one cesium AFS per the contract requirement for dual technology. These AFSs provide the tight frequency stability necessary to generate high accuracy ranging signals. The Navigation Data Unit (NDU) generates all the baseband forms of the ranging signals. The original NDU design included a spare slot which allowed the addition of M-code and L5 codes within the same envelope. The original NDU computer was designed with 300% expansion memory margin and 300% computational reserve (throughput margin), so that there was sufficient reserve to support the generation of the new navigation messages for M-code and L5 plus other modernization requirements. The computer program is reprogrammable on-orbit and is loaded from onboard EEPROM memory when power is applied, avoiding

Table 3.7 Block IIF Ranging Signal Set

Link (Frequency)	L1 (1,575.42 MHz)	L2 (1,227.6 MHz)	L5 (1,176.45 MHz)
Open (civil) signals	C/A-code	CM-code, CL-code	I5-code, Q5-code
Military (restricted) signals	P(Y)-code, M-code	P(Y)-code, M-code	—

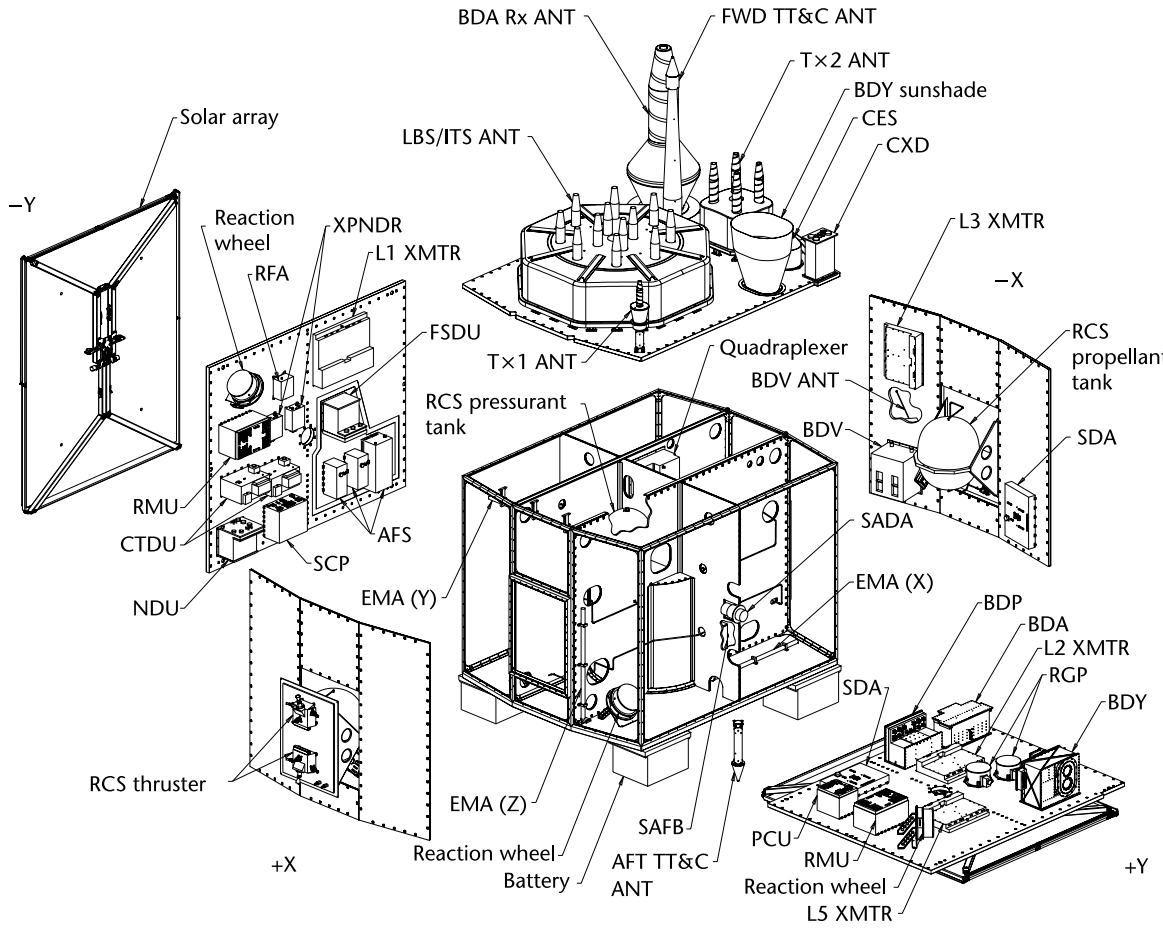


Figure 3.11 Expanded view of the Block IIF. (Source: The Boeing Company, Reprinted with permission.)

the need for large blocks of contact time with the ground antennas. The L-band subsystem generates about 350W of RF power for transmitting the three sets of signals in Table 3.7.

The Block IIF SV is designed for a 12-year design life with a MMD of 9.9 years. The Block IIF SV transmits all the same signals as the IIR-M, plus L5. An on-orbit depiction of the Block IIF SV is shown in Figure 3.12. The nadir facing side contains a set of UHF and L-band antennas and other components that are very reminiscent of all the previous GPS satellites.

The original IIF contract was for a basic buy of six SVs and two options of 15 and 12 SVs in groups of three for a possible total of 33 SVs. The Air Force exercised its option to buy an additional six SVs for a total of 12. The first Block IIF SV was launched in May 2010 and the twelfth was launched in February 2016. The URE performance for GPS IIF has ranged from 0.25m to 0.5m, once again easily surpassing the required performance of 3m.

3.2.3.8 GPS III Satellites

The GPS III SV (Figure 3.13) has been designed and is now being built to bring new capabilities to both military and civil PVT users throughout the globe. GPS

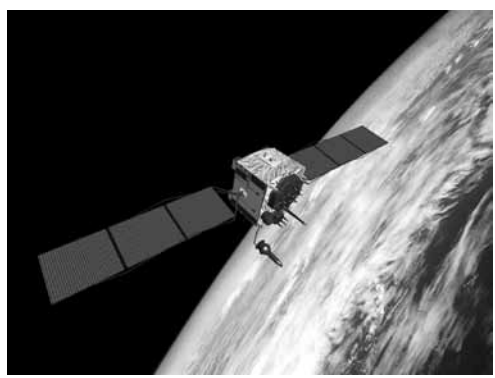


Figure 3.12 Block IIF satellite. (Source: The Boeing Company, Reprinted with permission.)

Figure 3.13 GPS III satellite. (Source: Lockheed Martin Corp. Reprinted with permission.)

III is fully backward compatible with existing GPS capabilities, but with important improvements as well as an expansion capability for the future of GPS.

GPS III contract award was announced in May 2008. The SV Critical Design Review (CDR) was successfully completed in August 2010 [23], marking the completion of the GPS III design phase. At the time of this writing, the production phase was well underway for several SVs, including integration and test (Figure 3.14), and the first GPS III SV is available for launch. These SVs are being built by Lockheed Martin, its navigation payload subcontractor, Harris (formerly ITT), communications payload subcontractor, General Dynamics, and numerous other subcontractors.

The new expandable GPS III design is based on the Lockheed Martin A2100 bus design and its long heritage. Important elements have also been pulled from the successful GPS Block IIR and IIR-M SVs and their heritage of over 250 years of accumulated on-orbit performance. The GPS III SV design itself has the capacity to accommodate new advanced capabilities in the active production line, as soon as they are mature and determined ready to add to the SV allowing it to readily adapt to new or changing requirements as it serves the user in the future. This includes the addition of a new civil signal (L1C) designed to be interoperable with similar signals broadcast by other GNSS constellations. (L1C characteristics are described in Section 3.7.2.4.)

Performance Requirements

On orbit, the GPS III SV will provide longer SV life, improved accuracy, and improved availability compared to all previous GPS generations. Table 3.8 provides a summary of the critical performance requirements for the program and parallel requirements for all earlier GPS versions, where available [24, 25]. Based on factory test results, GPS III will meet or exceed all of these critical requirements.

The GPS III SV requirements include a 12-year MMD and a 15-year design life [24]. The GPS III L-band signals will consist of the heritage L1 C/A, L1 P(Y), and L2 P(Y); the modernized L1M, L2C, and L2M; and full support for the new L5 and L1C civilian signals. The GPS III M-code received signal power at Earth will

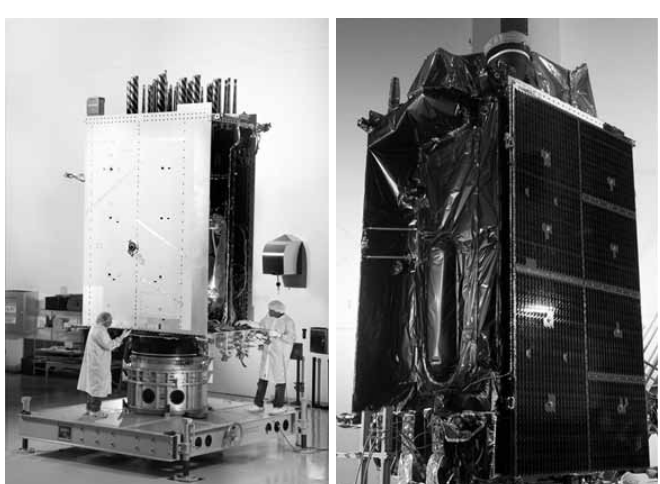


Figure 3.14 GPS III satellite in integration and testing.

Table 3.8 Comparison of GPS Satellite Specifications

Specifications		*I*	*II*	*IIA*	*IIR*	*IIR-M*	*IIF*	*III*
Accuracy (meters)	*User Range Error at 1 day*	—	7.6	7.6	2.2	2.2	3.0	1.0
	MMD Req.	4.5	6	6	6	6	9.9	12
	Design Life	5	7.5	7.5	7.5	7.5	12	15
SV Lifetime (years)	*Expendables*	7	10	10	10	10	—	—
	L1C/A	–160.0	–160.0	–158.5	–158.5	–158.5	–158.5	–158.5
	L1P	–163.0	–163.0	–161.5	–161.5	–161.5	–161.5	–161.5
	L1M	—	—	—	—	–158.0	–158.0	–158.0
	L1C	—	—	—	—	—	—	–157.0
	L2C	—	—	—	—	–160.0	–160.0	–158.5
	L2P	–166.0	–166.0	–164.5	–164.5	–161.5	–161.5	–161.5
Signal Power (dBW)	*L2M*	—	—	—	—	–161.5	–161.0	–158.0
	L5	—	—	—	—	—	–157.9	–157.0
SV Availability	%	Avail. levied on full constellation at 98% with SV goal of 95%					98.08%	99.45%

The L1C value is the total received power from both data and pilot channels, whereas the L5 values represent only the in-phase signal component power levels. The L5 total power levels (sum of in-phase and quadrature components) is 3 dB greater

be boosted to at least –153 dBW at 5° elevation as compaired to –158 dBW for IIR and IIF. This will provide for significantly improved service for military users in stressed conditions.

Navigation accuracy is one of the primary concerns for users. The GPS Block II and IIA SVs were required to meet a daily requirement of 7.6-m URE. The IIR is required to meet 2.2m at 24 hours when operating with a rubidium RAFS. IIF is required to meet 3-m URE at 24 hours. GPS III will be required to meet a 1.0-m URE requirement at 24 hours, a twofold to threefold improvement over current operational satellites.

GPS III Design Overview

The basic GPS III SV design can be highlighted by examining its various elements and subsystems: the navigation payload element (NPE), the network communications element (NCE), the hosted payload element (HPE), the antenna subsystem element, and the vehicle bus element with its subsystems. Figure 3.15 is an expanded view of the GPS III SV [24] showing its basic structure and notional component location. A brief description of each element and subsystem follows.

The NPE includes the payload computer [the mission data unit (MDU)], the L-band transmitters (L1, L2, L3, L5), the atomic frequency standards (AFSs), signal combiners, and signal filters. The MDU incorporates the waveform generator functionality first introduced in the modernized IIR-M SV [19, 20, 26]. The GPS III MDU has other significant advanced capabilities, including on-board ephemeris propagation, which uses a very small daily navigation upload to generate the broadcast navigation message for all legacy and modernized signals.

Each GPS III SV has three enhanced rubidium AFS units ("clocks") which build upon the strong heritage from the GPS IIR/IIR-M SVs [14]. The GPS III SV also includes a fourth slot for an enhanced new or experimental frequency standard

AFS	Atomic Frequency Standard
ECTS	Enhanced Crosslink Transponder System
ES	Earth Sensor Assembly
FBA	Fuse Box Assembly
IMU	Inertial Measurement Unit
LLED	Low Level Event Detector
MDU	Mission Data Unit
NPE	Navigation Payload Element
OBC	On-Board Computer
RAFS	Rubidium Atomic Frequency Standard
RIU	Remote Interface Unit
RWA	Reaction Wheel Assembly
SADA	Solar Array Drive Assembly
SGLS	Space-Ground Link System
SPRU	Scalable Power Regulation Unit
SUT	SGLS-USB Transponder
TFU	Transient Filter Unit
TT&C	Telemetry, Tracking and Command
UDU	Uplink/Downlink Unit
USB	Universal S-Band

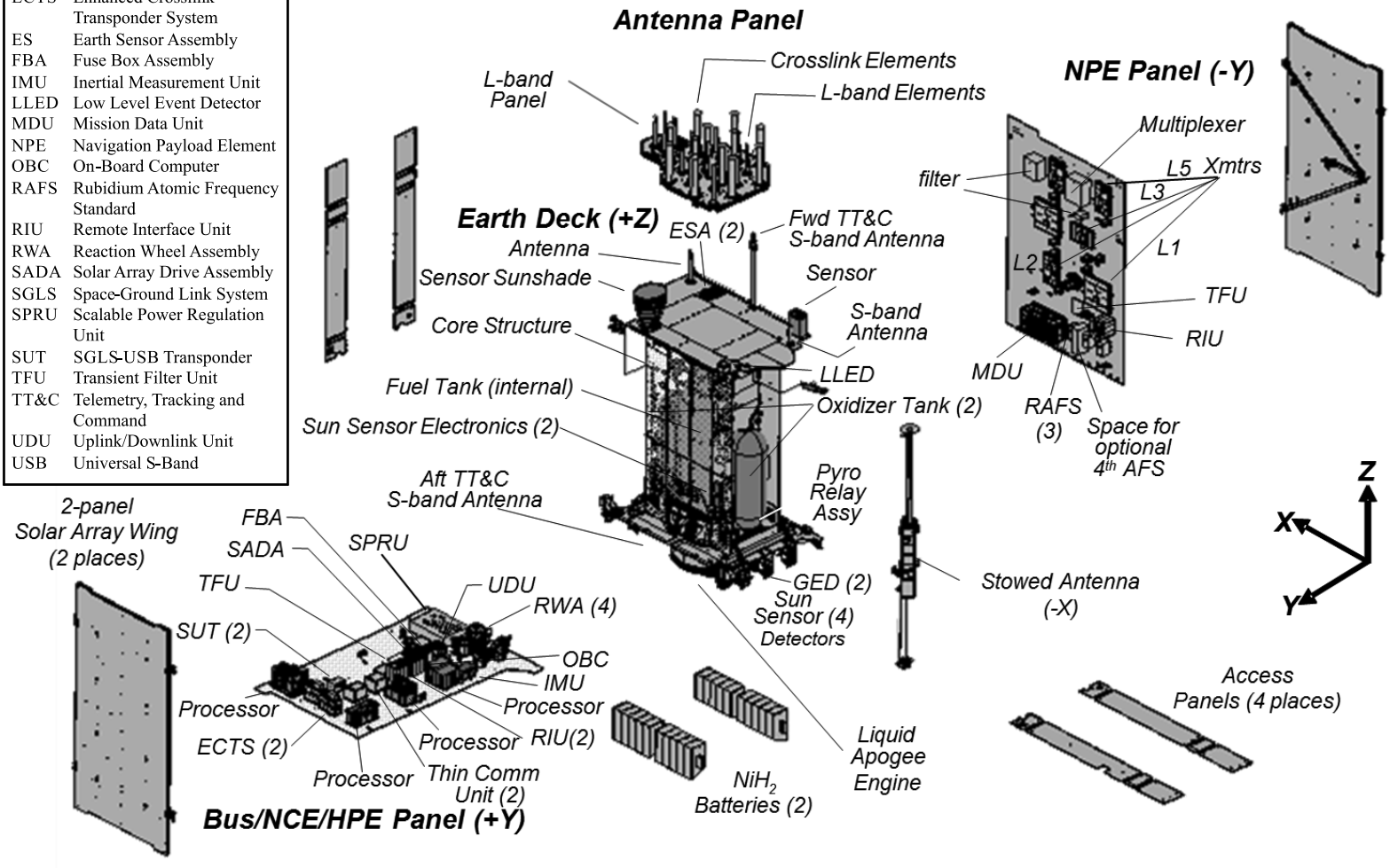


Figure 3.15 GPS III expanded view.

design, such as a hydrogen MASER. GPS III has a significant capability to operate and monitor a backup AFS, including the experimental AFS, for stability performance measurement and characterization. Redundant time-keeping system loops allow independent operation of the accurate hardware/software control loops. This capability is not available on any of the previous generations of GPS satellites [27].

The NCE provides communications capability to the SV. It consists of the enhanced crosslink transponder subsystem (ECTS), the thin communications unit (TCU) to distribute commands and collect telemetry, and the S-band boxes.

The HPE hosts several government-furnished equipment (GFE) items provided to GPS III. The antenna subsystem consists of the Earth coverage L-band antenna panel (based on IIR-M technology [18, 28]), the S-band antennas, and the UHF antennas.

The vehicle bus element includes numerous subsystems: the attitude control subsystem (ACS), the electrical power subsystem (EPS), the thermal control subsystem (TCS), the TT&C, the propulsion subsystem (PSS), and the mechanical subsystem (MSS). These subsystems will now be highlighted.

The ACS maintains the attitude knowledge and controls the pointing of the SV. This includes the nominal Earth pointing of the L-band antenna panel, and pointing the solar arrays at the Sun. It also controls thrust direction for propulsion subsystem maneuvers. The ACS consists of a set of sensors and actuators: Sun sensors, Earth sensors, the inertial measurement unit (IMU), reaction wheels for fine attitude control, magnetic torque rods for momentum unloading, 0.2-lbf thrusters for periodic station-keeping, and 5.0-lbf thrusters for more coarse maneuvers.

The EPS provides the stable electrical power for the entire SV, including during eclipse events. It consists of the solar arrays, nickel-hydrogen (NiH2) batteries, and the power regulation unit. The TCS maintains the proper temperature of the various SV components within safe limits. It consists of insulation, reflectors, heaters, radiators, heat pipes, and thermistors.

The TT&C subsystem consists of the bus computer [the on-board computer (OBC)], the uplink/downlink unit (UDU) for commanding and telemetry communication with the CS, remote interface units, deployment device control, and event detectors.

The PSS provides the thrust capability to alter the position and attitude of the SV. It consists of the Liquid Apogee Engine for final orbit insertion following launch, the 5-lbf thrusters for large on-orbit maneuvers, and the 0.2-lbf thrusters for attitude and station-keeping maneuvers.

The MSS consists of the basic SV structure, hinges, and the deployable elements.

GPS III Advanced Capabilities and Capability Insertion

GPS III will bring new capabilities to the user community beyond the current GPS generations, notably the new L1C signal.

GPS III is the first version of GPS SVs capable of selecting pseudo-random number (PRN) settings in the range of 38 to 63. This allows for more than 32 active SVs in the constellation, a limitation of the current GPS satellites. This capability complies with the latest IS-GPS-200 specification [29] and results in improved accuracy and greater coverage for all users.

GPS III can also host a fourth advanced technology clock. This could be used as a technology demonstrator or on-orbit performance validation for future clock designs which will likely be critical to future advances of the GPS constellation accuracy.

Central to the GPS III advanced design is ensuring that the navigation signals provided by the SV meet the defined criteria for safety-of-life applications for navigation users (i.e., signal integrity and continuity). This includes accuracy, availability, and protection from misleading signals. This capability is defined by the Positioning Signal Integrity and Continuity Assurance (PSICA) requirements. PSICA affects various SV subsystems such as design of the OBC and NPE [30].

GPS III is designed with capability insertion in mind. This is enabled by the modern, scalable bus design, active parts suppliers, engaged subcontractors, and an existing production line. Numerous important capabilities are already being pursued for near-term incorporation.

High on the list for capability insertion to GPS III is the Search and Rescue (SAR) payload, called the SAR/GPS, which relays distress signals from emergency beacons to search and rescue operations [31]. A Laser Reflector Array (LRA) capability will allow scientists to accurately measure the GPS III SV orbit, leading to more accurate modeling of the Earth's gravitational field and the effects of special relativity [10].

Other planned enhancements are currently in development for implementation. A Digital Waveform Generator (DWG) will replace the analog boxes creating a fully digital navigation payload capable of generating new navigation signals on-orbit. Lithium-ion (Li-ion) batteries will reduce SV weight and provide better EPS performance. Additionally, more M-code power will be added to provide higher power modernized signals to the military.

The New L1C Signal

GPS III will be the first GPS SV to broadcast the new L1C signal [32, 33] (Section 3.7.2.4 contains technical details of the L1C signal). This will be the fourth civilian signal (in addition to L1 C/A, L2C, and L5) and implements the second-generation CNAV-2 modernized navigation message. This signal will be interoperable (i.e., spectrally compatible) with other SATNAV systems, such as Europe's Galileo, Japan's Quasi-Zenith Satellite System (QZSS), and China's Beidou. (L1C interoperability (i.e., spectral compatibility) with these systems is discussed in Section 3.7.2.)

As with other modernized GPS signals, the new L1C signal structure provides for improved acquisition and tracking, faster data download and a more accurate ranging signal. It brings the modernized structure to the L1 frequency.

The L1C benefits include increased accuracy, acquisition, and tracking. Additional data bits in the message provide for greater PNT accuracy in the CNAV-2 message. Overall, navigation and timing users throughout the globe will benefit significantly from having GPS broadcast the L1C signal.

Pathfinder Satellite and Simulators

The GPS III Non-Flight Satellite Testbed (GNST) is the pathfinder unit for the GPS III program [34]. A full-sized version of the GPS III SV, populated with fully

functional non-flight boxes and loaded with the operational flight software, the GNST serves as a risk reduction platform for GPS III. It has provided for physical fit-checks at the factory and at the launch site, as well as electrical and flight software functional verification. It was an important platform for development and validation of factory build-and-test procedures. This has significantly reduced the risk for SV assembly, test, prelaunch operations, and capability insertion. It will now serve as a long-term test article for the entire life of the GPS III program, providing SV-level validation, early verification of ground, support, and test equipment, and early confirmation and rehearsal of transportation operations.

A number of high-fidelity and low-fidelity GPS III simulators have been developed and delivered. This includes the GPS III Spacecraft Simulator (G3SS) located at Cape Canaveral, a bus real-time simulator, the Integrated Software Interface Test Environment (InSite), and the Space Vehicle Subsystem Models and Simulation (SVSMS). These simulators, in both hardware and software implementations, provide support for ground and vehicle system checkout, launch readiness, and SV on-orbit maintenance.

Current Status

The first GPS III Space Vehicle has completed design qualification, environmental testing, and is now available for launch to begin a new era in GPS performance and capability. It will have advanced and expandable capabilities and will provide increased performance to the GPS user. GPS III will provide PVT services and advanced antijam capabilities yielding superior system security, accuracy, and reliability worldwide.

GPS III will sustain the GPS constellation, replacing older SVs that are well past their expected lives. All users, civilian and military, will benefit from the improved performance and advanced capabilities of GPS III for the several decades to come. The GPS III capabilities with the new L1C signal, higher signal power, greater accuracy, longer SV lifetime, and higher signal availability will maintain GPS as the gold standard for worldwide satellite navigation systems [35].

3.3 Control Segment Description

The GPS control segment (CS) provides capabilities for monitoring, commanding, and controlling the GPS satellite constellation. Functionally, the CS monitors the downlink L-band navigation signals, generates and updates the navigation messages, and is used to resolve satellite anomalies. Additionally, the CS monitors each satellite's state of health, manages tasks associated with satellite station-keeping maneuvers and battery recharging, and commands the satellite bus and payloads, as required [36]. After a quick overview, this section will present a detailed discussion of the CS's current configuration and its primary functions, including data processing of the navigation mission of GPS. This will be followed by a short summary of recent GPS CS changes, and then a discussion of near-term planned upgrades. The satellite commanding and maintenance activities will not be discussed.

The GPS CS, the operational control system (OCS), consists of three subsystems: the MCS, L-band monitor stations (MSs), and S-band ground antennas (GAs). The

operation of the OCS is performed at the MCS, under the operation of the U.S. Air Force Space Command, Second Space Operations Squadron (2 SOPS), located at Schriever Air Force Base (AFB) in Colorado Springs, Colorado. The MCS provides for continuous GPS services, 24 hours per day, 7 days a week, and serves as the mission control center for GPS operations. The Alternate MCS (AMCS), located at Vandenberg AFB, California, provides redundancy for the MCS. The major subsystems of the OCS and their functional allocation are shown in Figure 3.16.

The 2 SOPS supports all crew-action-required operations of the GPS constellation, including daily uploading of navigation information to the satellites and monitoring, diagnosis, reconfiguration, and station-keeping of all satellites in the GPS constellation. Spacecraft prelaunch, launch, and insertion operations are performed by a different control segment element, the launch, anomaly, and disposal operations (LADO) system with support from the reserve squadron, Nineteenth Space Operations Squadron (19 SOPS), also located at Schriever AFB. If a given SV is determined to be incapable of normal operations, the satellite commanding can be transferred to LADO for anomaly resolution or test monitoring.

3.3.1 OSC Current Configuration

At the time of this writing, the OCS configuration consists of dual MCSs, six OCS MSs, 10 National Geospatial-Intelligence Agency (NGA) MSs, and four GAs (see Figure 3.17 [37, 38]). The MCS data processing software, hosted on what is called the Architecture Evolution Plan (AEP) client-server platform running a POSIX-compliant operating system [39], commands and controls the OCS with multiple high-definition graphical and textual displays. The transition to the AEP version

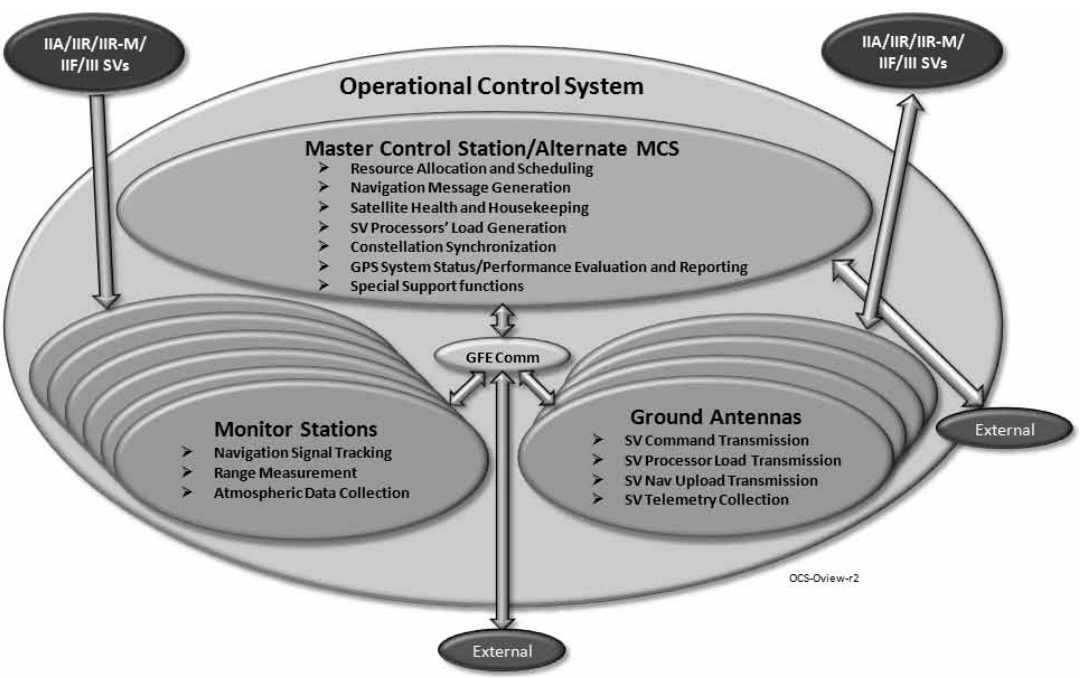


Figure 3.16 OCS overview.

Figure 3.17 Geographic distribution of OCS facilities [37].

of the OCS, and its foundation in the Legacy Accuracy Improvement Initiative (L-AII) upgrade as well as the full Accuracy Improvement Initiative (AII), will be highlighted in Section 3.3.2.

The OCS MSs and GAs are operated remotely from the active MCS (maintenance personnel are available at each site). The OSC MSs' and GAs' data processing software, hosted on a mixed set of computing platforms, communicates with the MCS using transmission control protocol/Internet protocol (TCP/IP) communication protocols. The MCS also has numerous internal and external communication links, also using TCP/IP.

3.3.1.1 MCS Description

The MCS provides the central command and control of the GPS constellation. Specific functions include:

- Monitoring and maintaining satellite state of health;
- Monitoring the satellite orbits;
- Estimating and predicting satellite clock and ephemeris parameters;
- Estimating the MS clock states;
- Generating GPS navigation messages;
- Maintaining the GPS timing service and its synchronization to UTC (USNO);
- Monitoring the navigation service integrity;
- End-to-end verifying and logging the navigation data delivered to the GPS user;

- Commanding satellite maneuvers to maintain the GPS orbit and repositioning due to vehicle failures.

All ground facilities necessary to support the GPS constellation are contained within the OCS, as shown in Figure 3.16. The OCS shares other ground antennas, the automated remote tracking stations (ARTS), from the Air Force Satellite Control Network (AFSCN) and additional MSs with NGA, in AEP. The MCS consists of data processing, control, display, and communications equipment to support these functions.

The primary task of the MCS is to generate and distribute the navigation data messages (sometimes referred to as the NAV Data messages). [Details of the NAV Data messages are contained in Sections 3.7.1.3 (Legacy NAV) and 3.7.3 (CNAV).] The MCS uses a sequence of steps, including collecting and processing the MS measurements, generating satellite ephemeris and clock estimates and predictions, and constructing and distributing the NAV Data messages. The MSs provide the raw pseudorange, carrier phase, and meteorological measurements that are smoothed by the MCS. A Kalman filter generates the precise satellite ephemeris and clock estimates, using these smoothed measurements. It is an epoch-state filter with the epoch of the state estimates at a different time than that of the measurements. The MCS filter is a linearized Kalman filter, with the ephemeris estimates linearized around a nominal reference trajectory. The reference trajectory is computed using accurate models to describe each satellite's motion. These ephemeris estimates, together with the reference trajectory, construct the precise ephemeris predictions that form the basis of the NAV Data message ephemeris parameters. Specifically, a least-squares-fit routine converts the predicted positions into the navigation orbital elements, in accordance with IS-GPS-200 [29]. The resulting orbital elements are uploaded into the satellite's navigation payload memory and transmitted to the GPS user.

Fundamentally, GPS navigation accuracy is derived from a coherent time scale, known as GPS system time, with one of the critical components being the satellite's AFS, which provides the stable reference for the satellite clock. As discussed earlier, each satellite carries multiple AFSs. The MCS commands the satellite AFSs, monitors their performance, and maintains estimates of satellite clock bias, drift, and drift rate (for rubidium only) to support the generation of clock corrections for the NAV Data messages. GPS system time is defined relative to an ensemble of selected active SV and MS AFSs [54]. The ensemble or composite AFS improves GPS time stability and minimizes its vunerability to any single AFS failure in defining such a coherent time scale.

Another important task of the MCS is to monitor the integrity of the navigation service. This is part of the L-band monitor processing (LBMON) [40]. Throughout the entire data flow from MCS to satellite and back, the MCS ensures the NAV Data message parameters are uploaded and transmitted correctly. The MCS maintains a complete memory image of the NAV Data message and compares downlink messages (received from its MSs) against the expected messages. Significant differences between the downlink versus expected navigation message result in an alert and corrective action by 2 SOPS. Along with navigation bit errors, the MCS monitors the L-band ranging data for consistency across satellites and across MSs.

When an inconsistency is observed across satellites or MSs, the MCS generates an L-band alert within 60 seconds of detection [40].

The OCS depends on external data from USNO and NGA including coordination with the UTC (USNO) absolute time scale, precise MS coordinates, and Earth orientation parameters.

Role of NGA in GPS

The following has been extracted from [41]:

> NGA and its predecessor organizations have operated Global Positioning System (GPS) Monitor Stations for more than 20 years. The NGA GPS Monitor Station Network (MSN) supports the DoD reference system WGS 84 and has expanded to include direct support to the Air Force (AF) Operational Control Segment. NGA stations are located around the world, strategically placed to complement the more equatorial Air Force Monitor Stations. These stations use geodetic quality GPS receivers and high performance cesium clocks. NGA monitors the behavior of its stations remotely to ensure their data integrity and high rate of availability. The NGA GPS stations are tightly configured and controlled to achieve the highest accuracy possible.
>
> NGA plays a vital role in GPS integrity monitoring. Since September 2005, the AF GPS MCS Kalman filter has used data from NGA GPS Monitor Stations to determine the broadcast orbits of the GPS satellites. NGA data also gives the MCS visibility of the satellites from at least two stations at all times. Sending NGA data to the MCS in near real time has enhanced both the accuracy and integrity of GPS. The computation of the GPS Precise Ephemeris, which is considered DoD truth, is an integral part of supporting WGS 84. The ephemeris, computed using NGA, AF and a few International GNSS Service (IGS) stations, is provided to the AF and posted on NGA's website for all GPS users.

3.3.1.2 Monitor Station Description

To perform the navigation tracking function, the OCS has a dedicated, globally distributed, L-band MS network. At the time of this writing, the OCS network consists of 16 MSs, with coverage, as shown in Figure 3.18 (coverage shown between plus and minus 55° latitude) [42]. The six OCS MSs are Ascension Island, Diego Garcia, Kwajalein, Hawaii, Colorado Springs, and Cape Canaveral. The 10 NGA stations are Bahrain, Australia, Ecuador, the United States Naval Observatory (USNO, Washington, D.C.), Uruguay, United Kingdom, South Africa, South Korea, New Zealand, and Alaska. The OCS MSs are located near the equator and the NGA MSs provide mid-latitude and high-latitude (both North and South) locations to maximize L-band coverage.

Each OCS MS operates under the control of the MCS and consists of the equipment and computer programs necessary to collect satellite-ranging data, satellite status data, and local meteorological data. This data is forwarded to the MCS for processing. Specifically, an OCS MS consists of a single dual-frequency receiver, dual cesium AFSs, meteorological sensors, and local workstations and communication equipment. Each receiver's antenna element consists of a conical ground plane with annular chokes at the base to produce a 14-dB multipath-to-direct signal

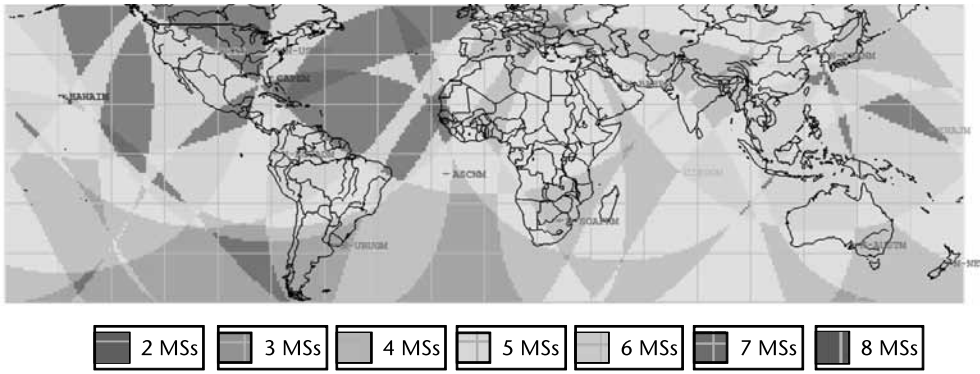


Figure 3.18 OCS and NGA monitoring station coverage [42].

rejection ratio for signal paths above 15° elevation. (An in-depth discussion on multipath is contained in Chapter 9.) The HP5071 cesium AFSs provide a 5-MHz reference to the receiver. Continuous-phase measurements between the AFSs are provided to the MCS for independent monitoring of the active atomic clock and for support of AFS switchovers. The MCS maintains a coherent MS time scale. At AFS switchovers, the MCS provides the phase and frequency difference estimates (between AFSs) to the MCS Kalman filter to minimize any time scale disruptions. Meteorological sensors provide surface pressure, temperature, and dew point measurements to the MCS Kalman filter to model the troposphere delay. However, these meteorological sensors are in disrepair, and their measurements have been replaced by monthly tabular data [43]. The local workstations provide commands and data collection between the OCS MS and the MCS.

The OCS MSs use a 12-channel, survey-grade, all-in-view receiver. These receivers, developed by Allen Osbourne Associates (AOA) (now Harris), are based on proven Jet Propulsion Laboratory (JPL) Turbo Rogue technology. The AOA receiver is designed with complete independence between the L1 and L2 tracking loops, with each tracking loop commanded by the MCS under various track acquisition strategies. With such a design, the overall receiver tracking performance can be maintained, even when tracking abnormal satellites (e.g., nonstandard code or satellite initialization, which require additional acquisition processing). These all-digital receivers have no detectable interchannel bias errors. (An earlier OCS receiver required external interchannel bias compensation due to its analog design with separate correlation and data processing cards. Interchannel bias is a time-delay difference incurred when processing a common satellite signal through different hardware and data processing paths in a receiver.)

The OCS MS receivers differ from normal receivers in several areas. First, these receivers require external commands for acquisition. Although most user equipment is only designed to acquire and track GPS signals that are in compliance with applicable specifications, the OCS receiver needs to track signals even when they are not in compliance. The external commands allow the OCS receiver to acquire and track abnormal signals from unhealthy satellites. Second, all measurements are time tagged to the satellite X1 epoch (see Section 3.7.1.1 for further details on the

X1 epoch), whereas a typical user receiver time tags range measurements relative to the receiver's X1 epoch. Synchronizing measurements relative to the satellite's X1 epochs facilitates the MCS's processing of data from the entire distributed OCS L-band MS network. The OCS receivers provide the MCS with 1.5-second pseudorange and accumulated delta range measurements [also known as P(Y)-code and carrier phase measurements, respectively]. Third, the MCS receives all of the raw demodulated navigation bits from each MS (without processing of the Hamming code used for error detection) so that problems in the NAV Data message can be observed. The returned NAV Data message is compared bit by bit against expected values to provide a complete system-level verification of the MCS-GA-satellite-MS data path (part of LBMON). Additionally, the OCS receivers provide the MCS with various internal signal indicators, such as time of lock of the tracking loops and internally measured signal-to-noise ratio (SNR). This additional data is used by the MCS to discard questionable measurements from the OCS Kalman filter. As noted earlier, the OCS maintains the MS time scale to accommodate station time changes, failures, and reinitialization of the station equipment. The Air Force MS coverage of the GPS satellites is shown in Figure 3.18, with the grayscale code denoting the number of MSs visible to a satellite [42]. Satellite coverage varies from one in the region west of South America to as many as five in the continental United States.

The NGA MSs use 12-channel GPS receivers, developed by ITT Industries (now Harris), that are similar to the OCS MS receivers. Indeed, the original NGA MS receiver contract was awarded to AOA in July 2002, but AOA was subsequently acquired by ITT in 2004.

3.3.1.3 Ground Uplink Antenna Description

To perform the satellite commanding and data transmission function, the OCS includes a dedicated, globally distributed, GA network. Currently, the OCS network, collocated with the Air Force MSs, consists of Ascension Island, Diego Garcia, Kwajalein, and Cape Canaveral. The Cape Canaveral facility also serves as part of the prelaunch compatibility station supporting prelaunch satellite compatibility testing. Additionally, several ARTS GAs located around the world, from the AFSCN, serve as GPS GAs, when scheduled. These GAs provide the TT&C interface between the OCS and the space segment, and for uploading the navigation data.

These GAs are full-duplex, S-band communication facilities that have dedicated command and control sessions with a single SV at a time. Under MCS control, multiple simultaneous satellite contacts can be performed. Each GA consists of the equipment and computer programs necessary to transmit commands, navigation data uploads, and payload control data received from the MCS to the satellites and to receive satellite telemetry data that is forwarded to the MCS. All OCS GAs are dual-threaded for system redundancy and integrity. The AFSCN ARTS GAs can also support S-band ranging. The S-band ranging provides the OCS with the capability to perform satellite early orbit and anomaly resolution support. The GA coverage of the GPS satellites, between plus and minus 55° latitude, is shown in Figure 3.19, with the grayscale code denoting the number of GAs visible to a satellite [42].

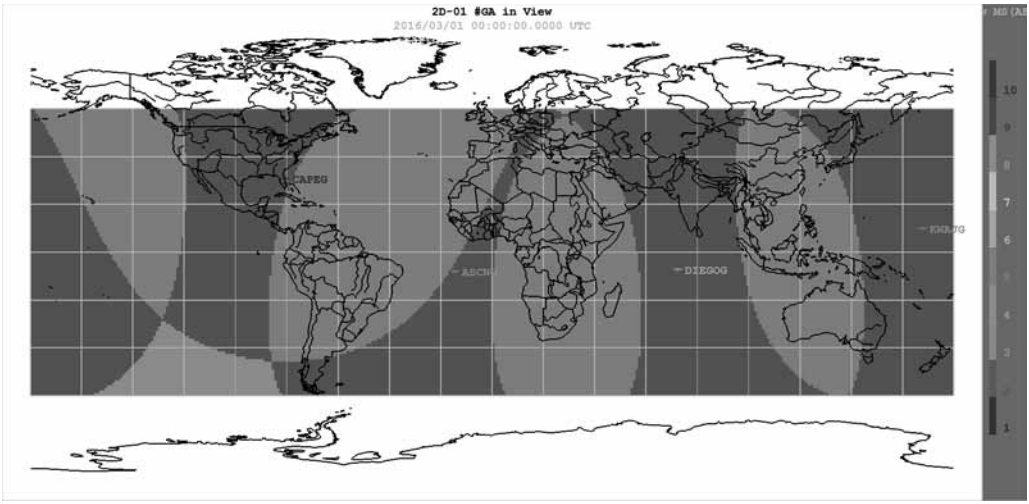


Figure 3.19 Control segment ground antenna coverage [42].

3.3.1.4 MCS Data Processing

MCS Measurement Processing

To support the MCS estimation and prediction function, the OCS continuously tracks the L1 and L2 P(Y) codes. At track acquisition, the L1 C/A code is sampled during the handover to P(Y) code to ensure that it is being broadcast (however, the OCS does not continuously track the L1 C/A code). The raw 1.5-second L1 and L2 pseudorange and carrier phase (also known as accumulated delta range) measurements are converted at the MCS into 15-minute smoothed measurements. The smoothing process uses the carrier phase measurements to smooth the pseudorange data to reduce the measurement noise. The process provides smoothed pseudorange and sampled carrier phase measurements for use by the MCS Kalman filter.

The smoothing process consists of data editing to remove outliers and cycle slips, converting raw dual-frequency measurements to ionosphere-free observables, and generating smoothed measurements once a sufficient number of validated measurements are available. Figure 3.20 shows a representative 15-minute data smoothing interval consisting of 600 pseudorange and carrier phase observations, with 595 observations used to form a smoothed pseudorange minus carrier phase offset and the 5 remaining observations used to form a carrier phase polynomial.

The MCS data editing limit checks the pseudoranges and performs third-difference tests on the raw L1 and L2 observables. The third-difference test compares consecutive sequences of L1 and L2 observables against thresholds. If the third-difference test exceeds these thresholds, then those observables are discarded for subsequent use in that interval. Such data editing protects the MCS Kalman filter from questionable measurements. Ionosphere-corrected, L1 pseudorange and phase measurements, ρ_c and ϕ_c, respectively, are computed using the standard ionosphere correction (see Section 10.2.4.1):

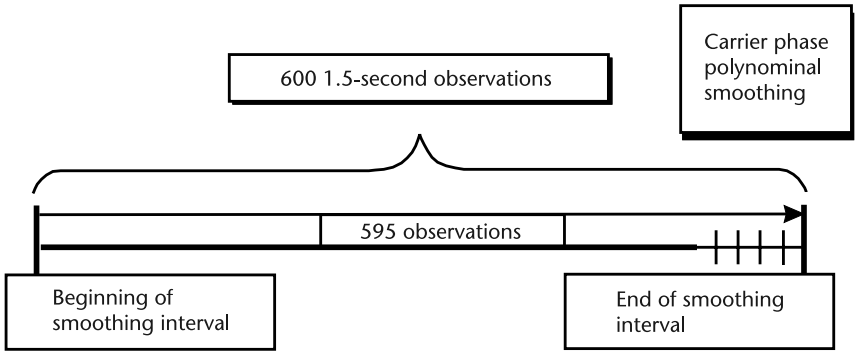


Figure 3.20 Representative MCS data-smoothing interval.

$$\begin{aligned}\rho_c &= \rho_1 - \frac{1}{(1-\alpha)} \cdot (\rho_1 - \rho_2) \\ \phi_c &= \phi_1 + \frac{1}{(1-\alpha)} \cdot (\phi_1 - \phi_2)\end{aligned} \tag{3.1}$$

where $\alpha = (154/120)^2$, and ρ_i and ϕ_i for $i = 1, 2$ are the validated L1 and L2 pseudorange and phase measurements, respectively.

Ionosphere-corrected pseudorange and carrier phase measurements are related by a constant offset. By exploiting this fact, a smoothed pseudorange measurement, $\overline{\rho}_c$, is formed from a carrier phase as follows:

$$\overline{\rho}_c = \phi_c + B \tag{3.2}$$

where B is an unknown constant computed by averaging the L1 ionosphere-corrected pseudorange and carrier phase measurement, ρ_c and ϕ_c, differences

$$B = \sum \left(\rho_c(z_j) - \phi_c(z_j) \right) \tag{3.3}$$

over all validated measurements in the smoothing interval. The MCS pioneered such carrier-aided smoothing of pseudoranges in the early 1980s [49].

The MCS Kalman filter performs measurement updates every 15 minutes based on its uniform GPS time scale (i.e., GPS system time). The smoothing process generates second-order pseudorange and carrier phase measurement polynomials in the neighborhood of these Kalman update times. A phase measurement polynomial, consisting of bias, drift and drift rate, $\hat{\mathbf{X}}_c$, is formed using a least-squares fit of the last five phase measurements in the smoothing interval, $\vec{\boldsymbol{\phi}}_c$:

$$\hat{\mathbf{X}}_c = \left(\mathbf{A}^T \, \mathbf{W} \, \mathbf{A} \right)^{-1} \cdot \quad \mathbf{A}^T \, \mathbf{W} \, \vec{\boldsymbol{\phi}}_c \tag{3.4}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & -2\tau & 4\tau^2 \\ 1 & -\tau & \tau^2 \\ 1 & 0 & 0 \\ 1 & \tau & \tau^2 \\ 1 & 2\tau & 4\tau^2 \end{bmatrix}, \quad \vec{\boldsymbol{\phi}}_c = \begin{bmatrix} \phi_c(z_{-2}) \\ \phi_c(z_{-1}) \\ \phi_c(z_0) \\ \phi_c(z_1) \\ \phi_c(z_2) \end{bmatrix} \tag{3.5}$$

where τ equals 1.5 seconds and $\{z_i$, for $i = -2, -1, 0, 1, 2\}$ denotes the time tags associated with the last five phase measurements in the interval. The weighting matrix, $\mathbf{W}$ in (3.4), is diagonal with weights derived from the receiver's reported SNR value. The pseudorange measurement polynomial, $\hat{\mathbf{X}}_p$, is formed using the constant offset in (3.3) as follows:

$$\hat{\mathbf{X}}_p = \hat{\mathbf{X}}_c + \begin{bmatrix} B \\ 0 \\ 0 \end{bmatrix} \tag{3.6}$$

These smoothed pseudorange and phase measurements, in (3.6) and (3.4), respectively, are interpolated by the MCS Kalman filter to a common GPS time scale, using the satellite clock estimates.

MCS Ephemeris and Clock Processing

The MCS ephemeris and clock processing software continuously estimates the satellite ephemeris, clock, and MS states, using a Kalman filter with 15-minute updates based on the smoothed measurements described above. The MCS ephemeris and clock estimates are used to predict the satellite's position and clock at future times to support the generation of the NAV Data message.

The MCS ephemeris and clock processing is decomposed into two components: offline processing for generating reference trajectories and real-time processing for the inertial-to-geodetic coordinate transformations, the Sun/Moon ephemeris, and for maintaining the MCS Kalman filter estimates. The MCS offline processing depends on highly accurate models. The MCS reference trajectory force models [44, 45] include the 1996 Earth Gravitation Model (EGM 96) with gravitational harmonics truncated to degree 12 and order 12, the satellite-unique solar radiation models, the solar and lunar gravitational effects (derived from the JPL Solar Ephemeris, DE200) and the IERS 2003 solar and lunar solid tidal effects, including vertical and horizontal components. The magnitude of these various forces and their corresponding effect on the GPS orbits has been analyzed and is summarized in Table 3.9 [46].

The differences on the left and right sides of Table 3.9 quantify the positional error due to that component on the ephemeris trajectory and orbit determination, respectively. Since the equations of motion describing GPS orbits are nonlinear, the MCS linearizes the ephemeris states about a nominal reference trajectory [47, 49]. To support ephemeris predictions, these ephemeris estimates are maintained

Table 3.9 Acceleration Forces Perturbing the Satellite Orbit

	RMS Orbit Differences over Three Days (m)				*RMS Orbit Determination (m)*			
Perturbing Acceleration	*Radial*	*Along Track*	*Cross Track*	*Total*	*Radial*	*Along Track*	*Cross Track*	*Total*
Earth oblateness (C_{20})	1,341	36,788	18,120	41,030	1,147	1,421	6,841	7,054
Moon gravitation	231	3,540	1,079	3,708	87	126	480	504
Sun gravitation	83	1,755	431	1,809	30	13	6	33
C_{22}, S_{22}	80	498	10	504	3	3	4	5
C_{nm}, S_{nm} (n,m=3..8)	11	204	10	204	4	13	5	15
C_{nm}, S_{nm} (n,m=4..8)	2	41	1	41	1	2	1	2
C_{nm}, S_{nm} (n,m=5..8)	1	8	0	8	0	0	0	0
Solar radiation pressure	90	258	4	273	0	0	0	0.001

relative to the reference trajectory's epoch states and the trajectory partials (relative to the epoch) used to propagate to current or future times.

The MCS Kalman filter tracks the satellite ephemeris in Earth Center Inertial (ECI) coordinates and transforms the satellite positions into Earth-centered, Earth-fixed (ECEF) coordinates using a series of rotation matrices. These ECI-to-ECEF coordinate rotation matrices account for luni-solar and planetary precession, nutation, Earth rotation, polar motion and UT1-UTC effects [65]. (Polar motion and UT1-UTC Earth orientation predictions are provided daily to the OCS by the NGA.)

The MCS Kalman state estimate consists of three ECI positions and velocities, two solar pressures and up to three clock states for each satellite, and a tropospheric wet height and two clock states for each MS. The two solar pressure states consist of a scaling parameter to the a priori solar pressure model and a Y-body axis acceleration. The Kalman filter clock states include a bias, drift and drift rate (for Rubidium only). To avoid numerical instability, the MCS Kalman filter is formulated in **U-D** factored form, where the state covariance (e.g., **P**) is maintained as:

$$\mathbf{P} = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{U}^T \tag{3.7}$$

with **U** and **D** being upper triangular and diagonal matrices, respectively [48]. The **U-D** filter improves the numerical dynamic range of the MCS filter estimates whose time constants vary from several hours to several weeks. The MCS Kalman time update has the form:

$$\tilde{\mathbf{U}}(t_{k+1})\,\tilde{\mathbf{D}}(t_{k+1})\,\tilde{\mathbf{U}}(t_{k+1})^T = \left[\mathbf{B}(t_k) \,\middle|\, \hat{\mathbf{U}}(t_k)\right] \left[\begin{array}{c|c} \mathbf{Q}(t_k) & \\ \hline & \hat{\mathbf{D}}(t_k) \end{array}\right] \left[\begin{array}{c} \mathbf{B}(t_k)^T \\ \hline \hat{\mathbf{U}}(t_k)^T \end{array}\right] \tag{3.8}$$

where $\hat{\mathbf{U}}(\cdot)$, $\hat{\mathbf{D}}(\cdot)$ and $\tilde{\mathbf{U}}(\cdot)$, $\tilde{\mathbf{D}}(\cdot)$ denote the a priori and a posteriori covariance factors, respectively, $\mathbf{Q}(\cdot)$ denotes the state process noise matrix, and $\mathbf{B}(\cdot)$ denotes the matrix that maps the process noise to the appropriate state domain. The MCS process noises include the satellite and MS clocks, troposphere-wet height, solar pressure, and ephemeris velocity (with the latter being in radial, along-track, and

cross-track coordinates [49]). Periodically, the 2 SOPS retunes the satellite and MS clock process noises, using on-orbit GPS Allan and Hadamard clock characterization, as provided by the Naval Research Laboratory [50, 51]. The MCS Kalman filter performs scalar measurement updates, with a statistically consistent test to detect outliers (based on the measurement residuals or innovation process [47]). The MCS measurement model includes a clock polynomial model (up to second-order), the Neill/Saastamoinen troposphere model [52, 53], the IERS 2003 station tide displacement model (both vertical and horizontal components), periodic relativity, and satellite phase center corrections.

Since a pseudorange measurement is simply the signal transit time between the transmitting satellite and the receiving MS, the MCS Kalman filter can estimate both the ephemeris and clock errors. However, any error common to all the clocks remains unobservable. Essentially, given a system of n clocks, there are only equivalently $n - 1$ separable clock observables, leaving one unobservable state. An early MCS Kalman filter design avoided this unobservablity by artificially forcing a single MS clock as the master and referencing all MCS clock estimates to that station. Based on the theory of composite clocks, developed in [54], the MCS Kalman filter was upgraded to exploit this unobservability and established GPS system time as the ensemble of selected active AFSs. At each measurement update, the composite clock reduces the clock estimate uncertainties [49]. Also with the composite clock, GPS time is steered to the UTC(USNO) absolute time scale (accounting for current leap second differences) for consistency with other timing services. Common view of the satellites from multiple MSs is critical to the estimation process. This closure of the time-transfer function provides the global time scale synchronization necessary to achieve submeter estimation performance. Given such advantages of the composite clock, the International GPS Service (IGS) transitioned its products to IGS system time along the lines of the composite clock [55].

The MCS Kalman filter has several unique features. First, the MCS Kalman filter is decomposed into smaller minifilters, known as partitions. The MCS partitioned Kalman filter was required due to computational limitations in the 1980s, but now provides flexibility to remove lesser-performing satellites from the primary partition. In a single partition, the Kalman filter estimates up to 32 satellites and all MS states, with logic across partitions to coordinate the alignment of the redundant ground estimates. Second, the MCS Kalman filter has constant state estimates, that is, filter states with zero covariance. (This feature is used in the cesium and rubidium AFS models, which are linear and quadratic polynomials, respectively.) Classically, Kalman theory requires the state covariance to be positive-definite. However, given the **U-D** time update in (3.8) and its associated Gram-Schmidt factorization [48], the a posteriori covariance factors, $\tilde{\mathbf{U}}(\cdot)$, $\tilde{\mathbf{D}}(\cdot)$, are constructed to be positive semidefinite with selected states having zero covariance. Third, the MCS Kalman filter supports Kalman backups. The MCS Kalman backup consists of retrieving prior filter states and covariances (up to the past 54 hours) and reprocessing the smoothed measurements under different filter configurations. This backup capability is critical to 2 SOPS for managing satellite, GA, or operator-induced abnormalities. The MCS Kalman filter has various controls available to 2 SOPS to manage special events including AFS runoffs, autonomous satellite jet firings, AFS reinitializations and switchovers of AFSs, reference trajectories, and Earth orientation parameter changes. The MCS Kalman filter has been continuously running

since the early 1980s with no filter restarts, including the transition from legacy to the AEP system in 2007.

MCS Upload Message Formulation

The MCS upload navigation messages are generated by a sequence of steps. First, the MCS generates predicted ECEF satellite antenna phase center positions, denoted as $\left[\tilde{r}_{sa}(\bullet \mid t_k)\right]_E$, using the most recent Kalman filter estimate at time, t_k. Next, for the legacy GPS signals, the MCS performs a least-squares fit of these predicted positions using the NAV Data message ephemeris parameters. The least-squares fits are over either 4-hour or 6-hour time intervals, also known as a subframe. (Note that the subframe fitting intervals are longer for the extended operation uploads.) The 15 orbital elements [29] can be expressed in vector form as

$$\mathbf{X}(t_{oe}) \equiv \left[\sqrt{a},\, e,\, M_0,\, \omega,\, \Omega_0,\, i_0,\, \dot{\Omega},\, \dot{i},\, \Delta n,\, C_{uc},\, C_{us},\, C_{ic},\, C_{is},\, C_{rc},\, C_{rs}\right]^T \tag{3.9}$$

with an associated ephemeris reference time, t_{oe}, and are generated using a nonlinear weighted least-squares fit.

For a given subframe, the orbital elements, $\mathbf{X}(t_{oe})$, are chosen to minimize the performance objective:

$$\sum_{\ell} \left\{ \begin{array}{l} \left(\left[\tilde{r}_{sa}(t_\ell \mid t_k)\right]_E - g_{eph}(t_\ell, \mathbf{X}(t_{oe}))\right)^T \mathbf{W}(t_\ell) \\ \quad \left(\left[\tilde{r}_{sa}(t_\ell \mid t_k)\right]_E - g_{eph}(t_\ell, \mathbf{X}(t_{oe}))\right) \end{array} \right\} \tag{3.10}$$

where $g_{eph}(\ )$ is a nonlinear function mapping the orbital elements, $\mathbf{X}(t_{oe})$, to an ECEF satellite antenna phase center position and $\mathbf{W}(\cdot)$ is a weighting matrix [29].

As defined in (3.10), all position vectors and associated weighting matrices are in ECEF coordinates. Since the MCS error budget is defined relative to the URE, the weighting matrix is resolved into radial, along-track and cross-track (RAC) coordinates, with the radial given the largest weight. The weighting matrix of (3.10) has the form:

$$\mathbf{W}(t_\ell) = \mathbf{M}_{E \leftarrow RAC}(t_\ell) \cdot \mathbf{W}_{RAC}(t_\ell) \cdot \mathbf{M}_{E \leftarrow RAC}(t_\ell)^T \tag{3.11}$$

where $\mathbf{M}_{E \leftarrow RAC}(\cdot)$ is a coordinate transformation from RAC-to-ECEF coordinates and $\mathrm{W}_{\mathrm{RAC}}$ is a diagonal RAC weighting matrix.

For the orbital elements in (3.9), the performance objective in (3.10) can become ill-conditioned for small eccentricity, e. An alternative orbital set is introduced to remove such ill-conditioning; specifically, three auxiliary elements defined as follows:

$$\alpha = e\cos\omega, \qquad \beta = e\sin\omega, \qquad \gamma = M_0 + \omega \tag{3.12}$$

Thus, the objective function in (3.10) is minimized relative to the alternative orbital elements, $\bar{\mathbf{X}}(\bullet)$ having the form:

$$\bar{\mathbf{X}}(t_{oe}) \equiv \left[\sqrt{a},\, \alpha,\, \beta,\, \gamma,\, \Omega_0,\, i_0,\, \dot{\Omega},\, \dot{i},\, \Delta n,\, C_{uc},\, C_{us},\, C_{ic},\, C_{is},\, C_{rc},\, C_{rs}\right]^T \tag{3.13}$$

The three orbital elements (e, M_0, ω) are related to the auxiliary elements (α, β, γ) by the inverse mapping

$$e = \sqrt{\alpha^2 + \beta^2}, \quad \omega = \tan^{-1}(\beta/\alpha), \quad M_0 = \gamma - \omega \tag{3.14}$$

The advantage of minimizing (3.10) with respect to $\bar{\mathbf{X}}(\bullet)$ in (3.13) versus $\mathbf{X}(\cdot)$ in (3.9) is that the auxiliary orbital elements are well defined for small eccentricity.

The minimization problem in (3.10) and (3.14) is simplified by linearizing $g_{eph}(\ )$ about a nominal orbital element set, denoted by $\bar{\mathbf{X}}_{nom}(t_{oe})$ such that

$$g_{eph}(t_\ell, \bar{\mathbf{X}}(t_{oe})) = g_{eph}(t_\ell, \bar{\mathbf{X}}_{nom}(t_{oe})) + \left.\frac{\partial\, g_{eph}(t_\ell, \lambda)}{\partial\, \lambda}\right|\lambda = \bar{\mathbf{X}}_{nom}(t_{oe}) \bullet \left(\bar{\mathbf{X}}(t_{oe}) - \bar{\mathbf{X}}_{nom}(t_{oe})\right) \tag{3.15}$$

and then (3.10) becomes equivalently

$$\sum_\ell \left\{ \begin{array}{l} \left(\left[\Delta\tilde{r}_{sa}(t_\ell \mid t_k)\right]_E - \mathbf{P}\left(t_\ell, \bar{\mathbf{X}}_{nom}(t_{oe})\right) \bullet \Delta\bar{\mathbf{X}}(t_{oe})\right)^T \\ \bullet \mathbf{W}(t_\ell) \bullet \left(\left[\Delta\tilde{r}_{sa}(t_\ell \mid t_k)\right]_E - \mathbf{P}\left(t_\ell, \bar{\mathbf{X}}_{nom}(t_{oe})\right) \bullet \Delta\bar{\mathbf{X}}(t_{oe})\right) \end{array} \right\} \tag{3.16}$$

where

$$\left[\Delta\tilde{r}_{sa}(t_\ell \mid t_k)\right]_E = \left[\tilde{r}_{sa}(t_\ell \mid t_k)\right]_E - g_{eph}(t_\ell, \bar{\mathbf{X}}_{nom}(t_{oe})) \tag{3.17}$$

$$\mathbf{P}\left(t_\ell, \bar{\mathbf{X}}_{nom}(t_{oe})\right) = \left.\frac{\partial\, g_{eph}(t_\ell, \lambda)}{\partial\, \lambda}\right|\lambda = \bar{\mathbf{X}}_{nom}(t_{oe}) \tag{3.18}$$

$$\Delta\bar{\mathbf{X}}(t_{oe}) = \bar{\mathbf{X}}(t_{oe}) - \bar{\mathbf{X}}_{nom}(t_{oe}) \tag{3.19}$$

Following classical least square techniques (see description in Appendix A) applied to the performance objective in (3.16) yields

$$\begin{aligned} &\sum_\ell \left\{\mathbf{P}\left(t_\ell, \bar{\mathbf{X}}_{nom}(t_{oe})\right)^T\ \mathbf{W}(t_\ell)\ \mathbf{P}\left(t_\ell, \bar{\mathbf{X}}_{nom}(t_{oe})\right)\right\}\Delta\bar{\mathbf{X}}(t_{oe}) \\ &= \sum_\ell \left\{\mathbf{P}\left(t_\ell, \bar{\mathbf{X}}_{nom}(t_{0r})\right)^T\ \mathbf{W}(t_\ell)\left[\Delta\tilde{r}_{sa}(t_\ell \mid t_k)\right]_E\right\} \end{aligned} \tag{3.20}$$

where the solution, $\Delta\bar{\mathbf{X}}(t_{oe})$, is referred to as the differential correction. Since $g_{eph}(\ )$ is nonlinear, the optimal orbital elements in (3.16) are obtained by successive iterations: first, a nominal orbital vector, $\bar{\mathbf{X}}_{nom}(t_{oe})$ followed by a series of the differential corrections, $\Delta\bar{\mathbf{X}}(t_{oe})$ using (3.20), until the differential correction converges to zero.

Following a similar approach, the almanac navigation parameters are also generated [29]. These resulting orbital elements, $\bar{\mathbf{X}}(\bullet)$, are then scaled and truncated in compliance with the NAV Data message format. Note, these orbital elements, $\bar{\mathbf{X}}(\bullet)$, are quasi-Keplerian and represent a local fit of the satellite ECEF trajectory and are not acceptable for overall orbit characterization.

Navigation Upload Curve Fit Errors

The generation of the navigation upload results in some errors from the least-squares fit and the LSB representation of the broadcast message. These error sources exist for both the legacy LNAV and the modernized CNAV navigation messages. Actual measured satellite clock offset and ephemeris fit errors per each SV for a single day, associated with the legacy NAV data message generation as described above, are shown in Figure 3.21 [56]. For 2-hour broadcast intervals, 4-hour utilization (fit) intervals, from August 17, 2016, eight performance metrics are depicted:

- For the orbit fit:
  - Average upper bound error (AVG UBE) per SV;
  - Maximum UBE error (MAX UBE) per SV;
  - Average error (AVG UBE) across all SVs;
  - Maximum error (MAX UBE) across all SVs.

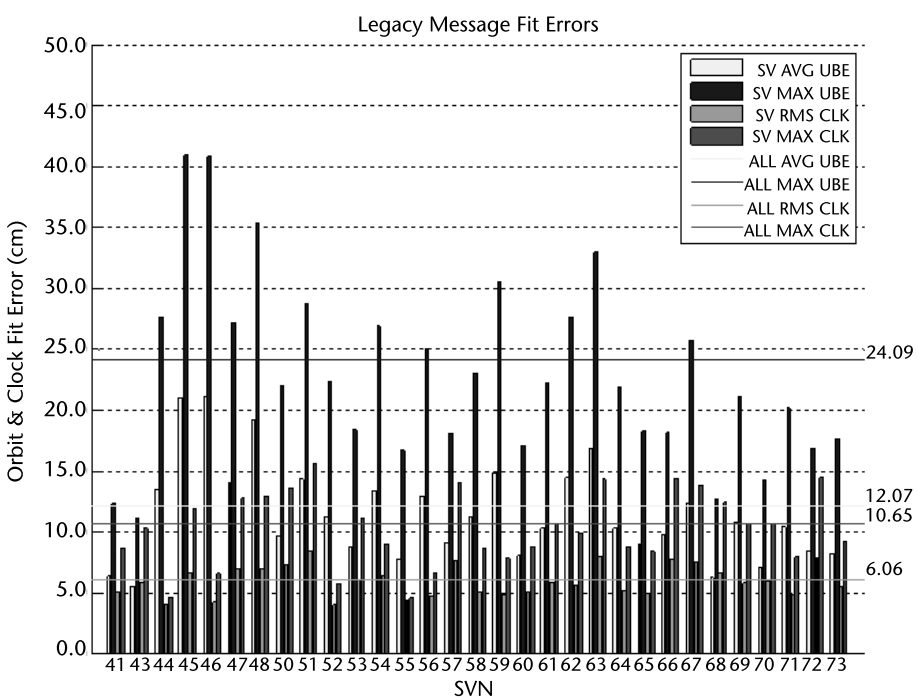


Figure 3.21 Legacy NAV message clock and orbit fit errors [56].

- For the satellite clock offset fit:
  - RMS (RMS CLK) per SV;
  - Maximum (MAX CLK) per SV;
  - RMS (RMS CLK) across all SVs;
  - Maximum (MAX CLK) across all SVs.

In Figure 3.21, the orbit fit RMS AVG and the RMS maximum error were 12.07 and 24.09 cm, respectively. For the SV clock offset fit, the RMS CLK and MAX CLK were 6.06 and 10.65 cm, respectively. Since this measured fit error data is for a single day, no daily/seasonal variations are included.

Regarding the modernized signals, an CNAV data message representation has been implemented with additional parameters and reduced quantization errors. Actual measured SV clock offset and ephemeris fit errors associated with the modernized NAV data message are shown in Figure 3.22 [56]. For 2-hour broadcast intervals, 3-hour utilization (fit) intervals, from June 7, 2016, the same eight performance metrics as in Figure 3.21 are depicted. In Figure 3.22, the orbit fit RMS AVG and the RMS maximum error were 0.47 and 0.82 cm, respectively. For the SV clock offset fit, the RMS CLK and MAX CLK were 0.28 and 0.42 cm, respectively. Similarly, with respect to the legacy NAV data, the modernized measured fit error data is for a single day; thus, no daily/seasonal variations are included. A comparison with the results of Figure 3.21 shows that the modernized fit errors are significantly reduced.

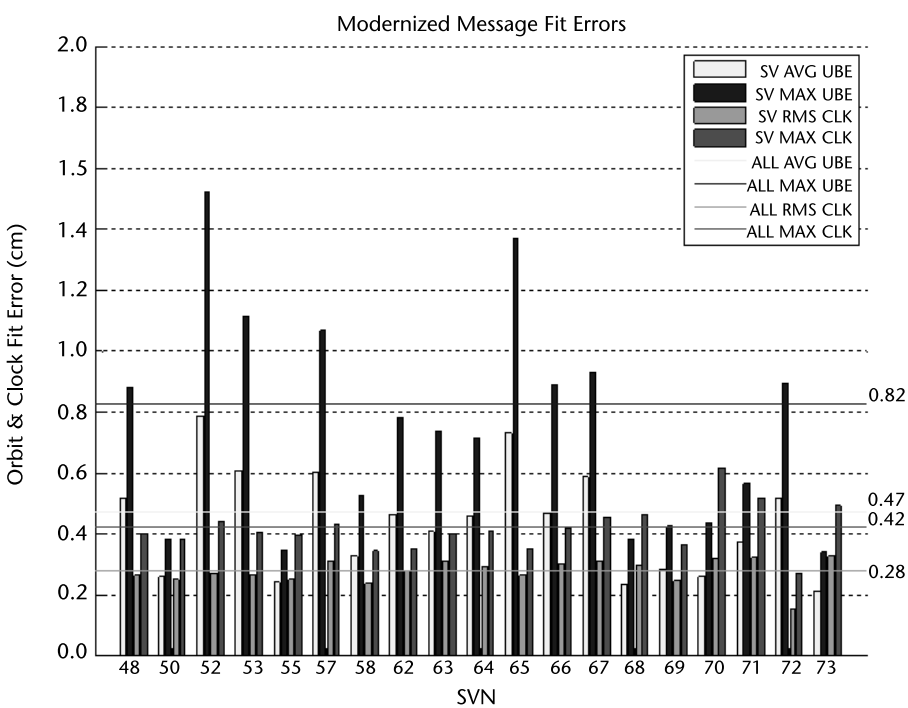


Figure 3.22 Modernized NAV message clock and orbit fit errors [56].

MCS Upload Message Dissemination

Nominally, each satellite's NAV Data message is uploaded at least once per day. The Legacy LNAV and CNAV Data messages (Sections 3.8.1 and 3.8.2) are generated in accordance with IS-GPS-200 [29]. Additionally, the MCS-GA-satellite uploads are checked after the navigation data is loaded into the satellite's memory. Error-checking processes exist along the entire path of navigation service for integrity. The satellite upload communication protocol maximizes the probability of successful transmission of data content to the satellite onboard memory.

The NAV Data is based on predictions of the MCS Kalman filter estimates, which degrade with the age of data. The 2 SOPS monitors the navigation accuracy and performs contingency uploads when the accuracy exceeds specific thresholds. Unfortunately, the dissemination of the NAV Data message is a trade-off of upload frequency to navigation accuracy. Various upload strategies have been evaluated to minimize upload frequency while maintaining an acceptable navigation service [15, 72]. GPS navigation accuracy depends on many factors, including performance of the satellite AFSs, the number and placement of the MSs, measurement errors, ephemeris modeling, and filter tuning. A significant improvement in the zero-age-of-data (ZAOD) accuracy results from the addition of the NGA MSs with L-AII and then AEP (as will be shown in the next section), along with more stable AFSs in the IIR and IIF satellites.

3.3.2 OCS Transition

The MCS operational software had been hosted on an IBM mainframe since the early 1980s. In September 2007, the MCS operation on the legacy mainframe was transferred to the AEP OCS, hosted on a distributed architecture of POSIX clients and servers [57–60]. The AEP update is an object-oriented software design using TCP/IP communication protocols across workstations connected by a 1-GB Ethernet local area network (LAN). The AEP distributed architecture maintains the MCS operational data in an Oracle database (with a failover strategy).

Major upgrades formed the basis of this transition. The L-AII was added to the legacy MCS in 2005 [38]. The L-AII upgrade provided additional capabilities to the legacy system and then aided the transition to AEP which encompassed the full AII capability [61, 62]. With L-AII and then AEP, the number of MSs could be increased to a total of 6 Air Force OCS MSs plus as many as 14 NGA MSs processed by the OCS. These additional NGA MSs provide the OCS with continuous L-band tracking coverage of the constellation. Prior to L-AII, there had been a satellite L-band coverage outage of up to 2 hours. This South Pacific gap no longer exists.

The L-AII upgrade modified the existing MCS mainframe implementation to support additional MSs and satellites in a partitioned Kalman filter. Since the 1980s, the MCS has used a partitioned Kalman filter consisting of up to six satellites and up to six MSs per partition. This partition filter design was due to computational limitations and hindered MCS navigation accuracy. The L-AII upgrade enabled the MCS to support up to 20 MSs and up to 32 satellites in a partition. (Note that the L-AII MCS Kalman filter maintained the partitioning and back-up capabilities to support satellite abnormalities.) NGA provided additional MSs for the MCS with 15-minute smoothed and 1.5-second raw pseudorange and

carrier phase measurements from Harris (formerly ITT) MS SAASM-based receivers. These smoothed and raw measurements are used in the MCS Kalman filter and LBMON [40] processing, respectively. With these improvements operational, the MCS Kalman filter zero-age-of-data URE reduced approximately by one-half [15, 64] and the L-band monitor visibility coverage increased from 1.5 MSs/satellite to 3 to 4 MSs/satellite. The combined OCS and NGA MS network was shown in Figure 3.18.

The L-AII upgrade (and AEP) included several model improvements to the MCS processing. The legacy and AEP models are summarized in Table 3.10. Various U.S. government agencies, research laboratories, and the international GPS community have developed improved GPS models over the past 30 years. These L-AII/AEP model updates of geopotential, station-tide displacement, and Earth orientation parameter enable the MCS processing to be compliant with the conventions of the IERS [65]. The JPL-developed solar pressure model improves the satellite ephemeris dynamic modeling with the inclusion of Y-axis, β-dependent force, where β is the angle between the Sun-Earth line and the satellite orbital plane. The Neill/Saastamoinen model improves tropospheric modeling at low elevations [67].

Zero Age of Data

A primary method for analyzing the performance of the MCS modeling, estimation, and upload generation is through the computation of the ZAOD [37, 70]. The ZAOD measurement represents the best estimate of the performance floor prior to the propagation to a future epoch (i.e., prediction for a navigation message). The ZAOD analysis compares the MCS Kalman filter states at a particular time to an independent truth standard. This metric is computed in terms of the URE. The ZAOD URE performance metric is valuable since it represents the base accuracy of the MCS Kalman filter. It is the highest-quality ephemeris and timing the GPS NAV Data message can provide to the GPS User. ZAOD URE aids in differentiating between problems with the MCS Kalman filter states and problems in the broadcast NAV Data message and how they will contribute to the GPS user's overall error.

Table 3.10 Legacy and AEP Model Upgrades

Model	*Legacy MCS Capability [44, 49]*	*AEP MCS Upgrade*
Geopotential model	WGS84 (8×8) gravitational harmonics	EGM 96 (12×12) gravitational harmonics [65]
Station tide displacement	Solid tide displacement accounting for lunar and solar vertical component only	IERS 2003, including vertical and horizontal components [65]
Earth orientation parameters	No zonal or diurnal/semidiurnal tidal compensation	Restoration of zonal tides and application of diurnal/semidiurnal tidal corrections [65]
Solar radiation pressure model	Rockwell Rock42 model for Block II/IIA and Lockheed Martin Lookup model for IIR [66]	JPL empirically derived solar pressure model [67]
Troposphere model	Hopfield/Black model [68, 69]	Neill/Saastamoinen model [52, 53]

Figure 3.23 shows the ZAOD URE improvement due to the installation of the L-AII software and hardware changes in 2005 [37, 71]. The total, ephemeris, and clock RMS URE values are shown along with a 7-day moving average of the total URE. Several L-AII milestones are labeled to provide an indication of where these milestones are located with respect to the data. Prior to L-AII, the average URE was around 0.45m. Following L-AII install, it improved to 0.25m.

Once the NAV Data message is generated, even prior to upload to the SV, the ZAOD state is left behind and the age of the data (and its NAV Data errors) grows. In general, GPS uploads occur on a daily basis. Several analyses and studies have shown that users benefit from reduced navigation errors with more frequent uploads, thus reducing the upload age of data and accompanying broadcast navigation message errors [72, 73].

Recent OCS Improvements

AEP provides an integrated suite of commercial off-the-shelf (COTS) products and improved graphical user interface displays. AEP was designed with hardware and software extensibility in mind. This enabled the addition of support for the IIF satellites and the modernized signals. In recent years, OCS extensibility has enabled the following additional upgrades:

- The SAASM implementation provided the next generation of security to the GPS system.
- The AEP MODNAV modification incorporated the capability for modernized navigation signals as a series of government off-the-shelf (GOTS) product implementations.

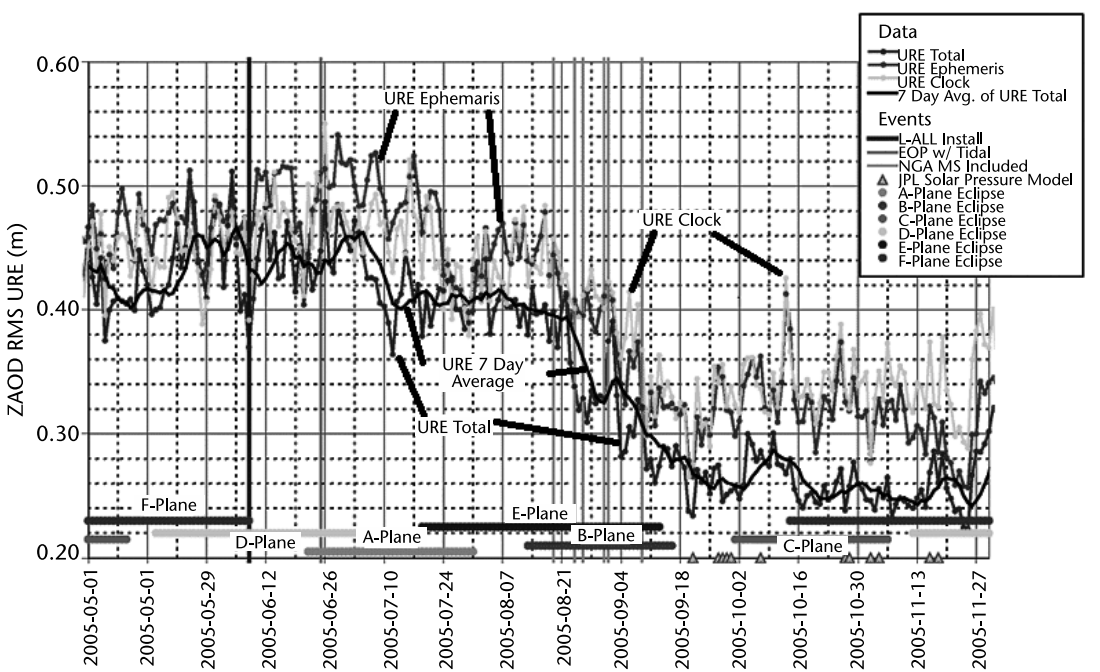


Figure 3.23 ZAOD improvement due to L-AII installation [37].

- The GPS Intrusion Protection Reinforcement (GIPR) modification provided a multitiered communications security capability for the protection of data and infrastructure and the enhancement of the sustainability of the system and to meet future GPS operational requirements [74].
- The COTS UPgrade (CUP) modification accomplished the initial phase of an ongoing OCS sustainment to update or replace obsolete COTS computer hardware and software products.
- AEP was modified to add the capability to ingest the Earth Orientation prediction data consistent to IERS Technical Note No. 36 (TN36) [75]. In the future, AEP will accommodate TN36 coordinate conversions for the purpose of building GPS III SV uploads. Full incorporation of TN36 conversion will be included in a future update of the MCS.
- The CUP Phase 2 (CUP 2) full modification of software and hardware improves the OCS's information assurance posture and extensibility. It affects the server hardware, as well as the operating system, database, network management, and command and control software.
- MODNAV Phase 2 removed the GOTS and integrated the capability into the OCS.

As AEP continues to evolve, the OCS will add additional features and functionality.

3.3.3 OCS Planned Upgrades

Over the next several years, the OCS will field several major upgrades:

- The CUP Phase 3 (CUP 3) modification will complete the modernization of AEP servers and will provide integration into the OCS to improve the OCS's extensibility.
- The GPS Ground Antenna/AFSCN Interface Technology Refresh (GAITR) modification will replace obsolete hardware for GA computing equipment.
- The Monitor Station Technology Improvement and Capability (MSTIC) will provide a software-based receiver to replace the existing OCS MS receivers.
- The Next Generation Operational Control Segment (OCX) will be a complete update and replacement of the current OCS, providing command, control, and navigation data uploads to the GPS Block IIR/IIR-M and Block IIF satellites, as well as full functional support to the new GPS III satellites.
- The initial phase of OCX will support GPS III launch, early on-orbit test, and anomaly resolution. This phase is called the Launch and Checkout Capability/Launch and Checkout System (LCC/LCS).
- The GPS III Contingency Operations (COps) update will provide telemetry, commanding, and navigation data upload capability to AEP to support the first several GPS III satellites [76].

3.4 User Segment

While this chapter focuses on GPS, the trend in user receiving equipment is for a receiver to utilize signals from one or more: constellations, SBAS services and/or regional SATNAV systems [77]. Therefore, we now refer to the receiver as a GNSS receiver. Technology trends in component miniaturization and large-scale manufacturing have led to a proliferation of low-cost GNSS receiver components. GNSS receivers are embedded in many of the items we use in our daily lives. These items include cellular telephones, cameras, and automobiles. This is in contrast to the initial receiving sets manufactured in the mid-1970s as part of the system concept validation phase. These first receivers were primarily analog devices for military applications and were large, bulky, and heavy. Depending on the application, a GNSS receiver is realized in many forms including single chip devices, chipsets, OEM (original equipment manufacturer) boards and standalone units. As mentioned above, a GNSS receiver processes signals from multiple global constellations, SBASs and regional SATNAV systems. An example is the Hemisphere Vector VS330 GNSS COMPASS that provides precision heading information using GPS, GLONASS and SBAS signals. Another example is the Qualcomm single chip SiRF AtlasVI which utilizes signals from GPS, SBAS, GLONASS, Galileo, BeiDou, and QZSS. At the time of this writing, the majority of GNSS receivers were chipsets or single chip devices integrated into the billions of cellphones in use worldwide (3.08 billion in 2014 [77]). Many of these single-chip GNSS receivers leverage low voltage bipolar complementary metal oxide semiconductor (BiCMOS) silicon germanium (SiGe) processes. The BiCMOS SiGe processes enable RF, analog and digital devices to be integrated on a single chip that incorporate onboard power management techniques to meet the need for small size and low battery drain of handheld devices [78]. Selection of a GNSS receiver depends on the user's application (e.g., civilian versus military, platform dynamics, shock and vibration environment, required accuracy, use in assisted GPS application). Following an overview of a typical receiver's components, selection criteria are addressed. Chapter 8 provides an in-depth technical description of GNSS receiver architectures and signal processing. GNSS receiver architectures/integrations for cellular telephone and automotive applications are contained in Chapter 13.

3.4.1 GNSS Receiver Characteristics

A block diagram of a GNSS receiver is shown in Figure 3.24. The GNSS receiver consists of five principal components: antenna, receiver front end, processor, input/output (I/O) device such as a control display unit (CDU), and a power supply.

Antenna

Satellite signals are received via the antenna, which is right-hand circularly polarized (RHCP) and provides near hemispherical coverage. As shown in Figure 3.25, typical coverage is 160° with gain variations from about 2.5 dBic at zenith to near unity at an elevation angle of 15°. (The RHCP antenna unity gain also can be expressed as 0 dBic = 0 dB with respect to an isotropic circularly polarized antenna.)

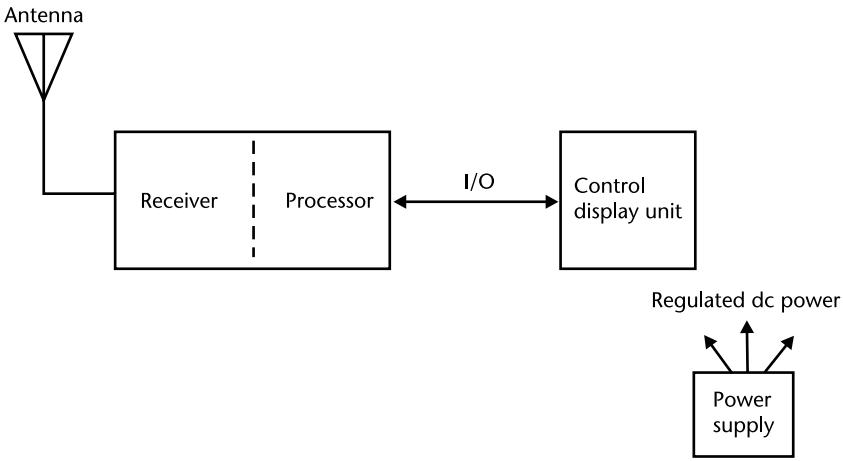


Figure 3.24 Principal GNSS receiver components.

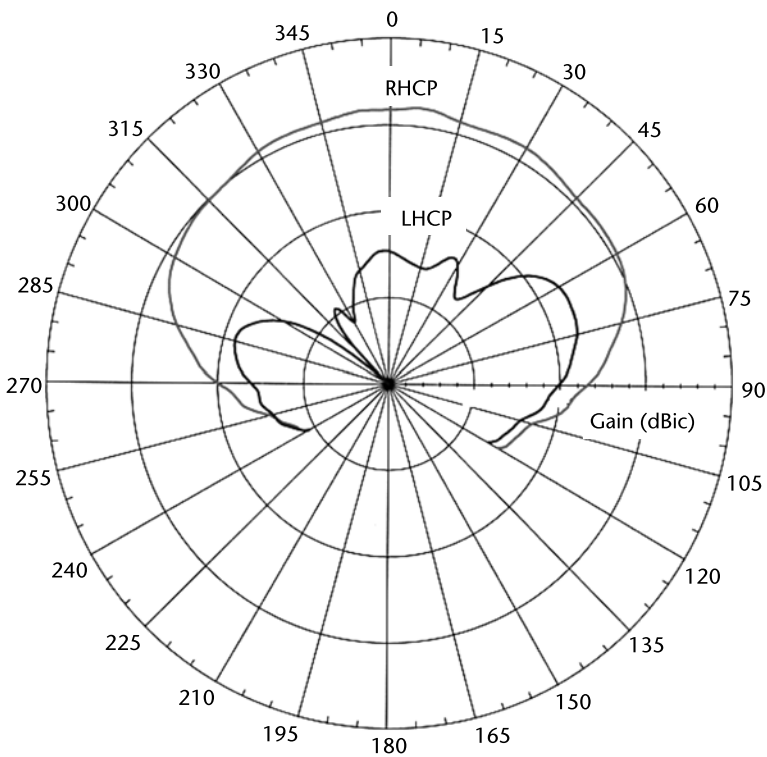


Figure 3.25 Example of RHCP hemispherical antenna pattern.

Below 15°, the gain is usually negative. Section 8.2 provides a detailed description of the various types of GNSS antennas and their respective applications.

Receiver

A detailed description of receiver signal acquisition and tracking operation is provided in Chapter 8; however, some high-level aspects are described herein to aid

our discussion. As stated above, most GNSS receivers will process signals from one or more: GNSS constellations, SBAS services and regional SATNAV systems. It is envisioned that a number of receiver types will be available. Most likely, these will be dual or tri-band to achieve ionospheric compensation and increased interference immunity. For safety-of-life applications, Aeronautical Radionavigation Service (ARNS) band users will require dual band (L1/E1 and L5/E5) receivers and antennas. High-accuracy applications using carrier phase measurements utilize signals from two or three frequency bands. Utilizing the carrier phase as a measurement observable enables centimeter-level (or even millimeter-level) measurement accuracy. (Carrier-phase measurements are described extensively in Section 12.3.1.2.)

Most receivers have multiple channels whereby each channel tracks the transmission from a single satellite on a single frequency. A simplified block diagram of a multichannel generic GNSS receiver is shown in Figure 3.26. The received RF CDMA satellite signals are usually filtered by a passive bandpass prefilter to reduce out-of-band RF interference. (Note that multiple prefilters may be required to receive signals from two or more frequency bands; that is one per frequency band.)

The prefilter is normally followed by a preamplifier. The RF signals are then downconverted to an intermediate frequency (IF). The IF signals are sampled and digitized by an analog-to-digital (A/D) converter. The A/D sampling rate is typically 2 to 20 times the PRN code chipping rate. The minimum sampling rate is twice the stopband bandwidth of the codes to satisfy the Nyquist criterion. Oversampling reduces the receiver sensitivity to A/D quantization noise, thereby reducing the number of bits required in the A/D converter. The samples are forwarded to the digital

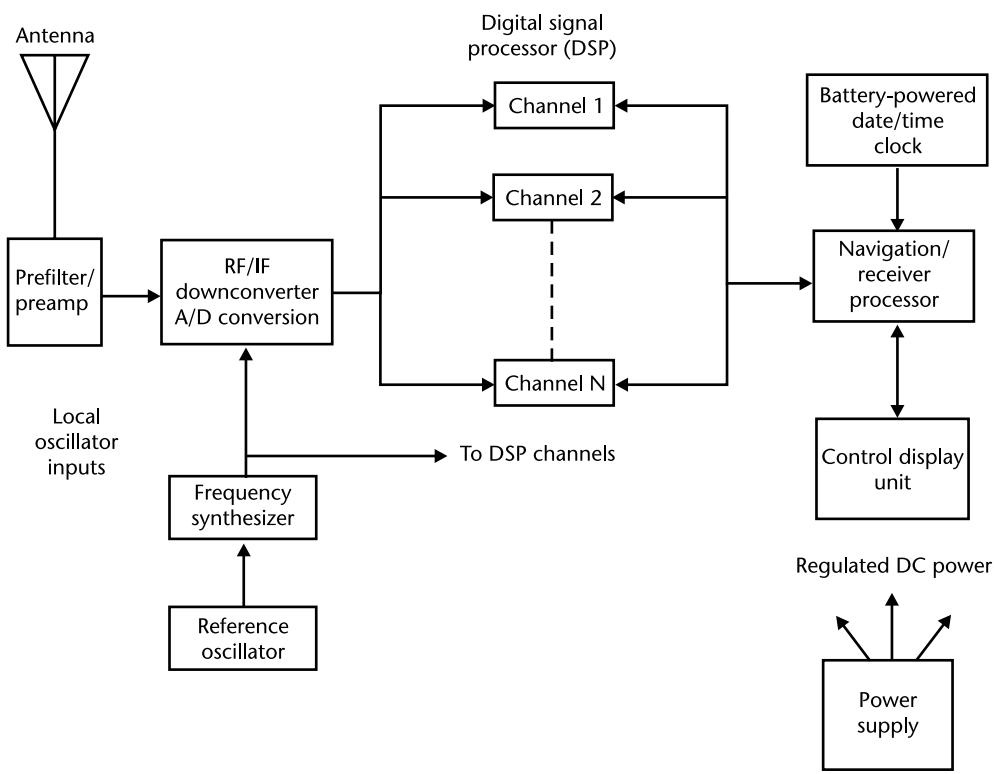


Figure 3.26 Generic GNSS receiver.

signal processor (DSP). The DSP contains N parallel channels to simultaneously track the carriers and codes from up to N satellites and corresponding frequencies. (N generally ranges from 12 to more than 100 in today's receivers.) Some receivers have a configurable number of channels depending on the particular user application [79]. Each channel contains code and carrier tracking loops to perform code and carrier-phase measurements as well as navigation message data demodulation. The channel may compute three different satellite-to-user measurement types: pseudoranges, carrier phase delta ranges (sometimes referred to as delta pseudorange), and integrated Doppler, depending on the implementation. The desired measurements and demodulated navigation message data are forwarded to the processor.

Note that GNSS receivers designed for use in handheld devices need to be power efficient. These receivers trade off susceptibility to high-power, in-band interferers to achieve minimum power supply (e.g., battery) drain. High dynamic range receiver front ends are needed in interference-resistant receivers and the necessary components (e.g., amplifiers and mixers with high intermodulation product levels) require high bias voltage levels. Also, the number of RF front ends and digital channels also are part of the performance versus power efficiency trade.

Navigation/Receiver Processor

A processor is generally required to control and command the receiver through its operational sequence, starting with channel signal acquisition and followed by signal tracking and NAV data collection. (Some GNSS receivers have an integral processing capability within the channel circuitry to perform these signal processing functions.) In addition, the processor may also form the PVT solution from the receiver measurements. In some applications, a separate processor may be dedicated to the computation of both PVT and associated navigation functions. Most processors provide an independent PVT solution on a 1-Hz basis. However, receivers designated for autoland aircraft precision approach and other high-dynamic applications normally require computation of independent PVT solutions at a minimum of 5 Hz. The formulated PVT solution and other navigation-related data are forwarded to the I/O device.

Input/Output Device

The input/output (I/O) device is the interface between the GNSS set and the user. I/O devices are of two basic types: integral or external. For many applications, the I/O device is a CDU. The CDU permits operator data entry, displays status and navigation solution parameters, and usually accesses numerous navigation functions such as waypoint entry and time-to-go. Most handheld units have an integral CDU. Other installations, such as those onboard an aircraft or ship, may have the I/O device integrated with existing instruments or control panels. In addition to the user and operator interface, applications such as integration with other sensors (e.g., INS) require a digital data interface to input and output data. Common interfaces include Bluetooth, USB, UART, Ethernet, ARINC 429, MIL-STD-1553B, RS-232, and RS-422.

Power Supply

The power supply can be integral, external, or a combination of the two. Typically, alkaline or lithium batteries are used for integral or self-contained implementations, such as handheld portable units; whereas an existing power supply is normally used in integrated applications, such as a board-mounted receiver installed within a server to provide accurate time. Airborne, automotive, and shipboard GNSS set installations normally use platform power but typically have built-in power converters (ac to dc or dc to dc) and regulators. There usually is an internal battery to maintain data stored in volatile random access memory (RAM) integrated circuits (ICs) and to operate a built-in timepiece (date/time clock) in the event platform power is disconnected.

3.4.1.1 GNSS Receiver Selection

At the time of this writing, there were over 45 GNSS receiver providers in the world [79]. While some like Qualcomm offer a few different chip set receivers for integration with other electronic functions, other companies like GARMIN and Trimble Navigation have many different end products ranging from handhelds to automobile and aircraft navigators to complex survey receivers. GNSS receiver selection is dependent on user application. The intended application strongly influences receiver design, construction, and capability. For each application, numerous environmental, operational, and performance parameters must be examined. A sampling of these parameters is provided here:

- Shock and vibration requirements, temperature and humidity extremes, as well as atmospheric salt content.
- The necessary independent PVT update rate must be determined. As an example, this rate is different for aircraft precision approach than for marine oil tanker guidance.
- Will the receiver be used in a network-assisted GNSS application (e.g., cellphone)? If so, is this in a mobile station assisted or mobile station based configuration? If mobile station assisted, the position solution is computed in the network. Here a network-based processor handles some of the functions of a traditional GNSS receiver. For a mobile station based configuration, the position solution is computed within the handset. (Elaboration on network assisted GNSS is provided in Chapter 13.)
- Will the receiver have to operate in a high multipath environment (i.e., near buildings or on an aircraft where satellite signals are reflected by various fuselage surfaces)? If so, multipath mitigation signal processing techniques and a choke ring antenna may be required. (Detailed descriptions of multipath and multipath mitigation techniques are contained in Chapter 9.)
- Under what type of dynamic conditions (e.g., acceleration, velocity) will the receiver have to operate? GNSS receivers for fighter aircraft applications are designed to maintain full performance even while experiencing multiple "g's" of acceleration, whereas sets designated for surveying are not normally de-

signed for severe dynamic environments. (Chapter 8 provides GNSS receiver design guidelines to accommodate anticipated dynamics.)

- Is a differential GNSS (DGNSS) capability required? (DGNSS is an accuracy enhancement technique covered in Chapter 12.) DGNSS provides greater accuracy than standalone GNSS operation. Most receivers are manufactured with a DGNSS capability.
- Does the application require reception of a geostationary satellite-based overlay service referred to as SBAS broadcasting satellite integrity, ranging, and/or DGNSS information? (SBAS is discussed in Chapter 12.) There are also commercial geostationary satellite services such as the NavCom Starfire system that provide corrections worldwide that can be received and processed in the same receiver. This provides centimeter-level precision in what appears to be a stand-alone receiver system, but a large, ground-based monitoring network and upload system are involved.
- Waypoint storage capability and the number of routes and legs need to be assessed.
- Does the GNSS receiver have to operate in an environment that requires enhanced interference rejection capabilities? Chapter 9 describes several techniques to achieve this.
- If the receiver has to be interfaced with an external system, does the proper I/O hardware and software exist? An example would be if the user requires a blended solution consisting of GNSS and other sensors such as an IMU and/or vision system.
- In terms of data input and display features, does the receiver require an external or integral CDU capability. Some aircraft and ships use repeater units such that data can be entered or extracted from various physical locations. Display requirements such as sunlight-readable or night-vision-goggle-compatible must be considered.
- Are local datum conversions required, or is WGS-84 sufficient? If so, does the receiver contain the proper transformations?
- Is portability for field use required?
- Economics, physical size, and power consumption must also be considered.

As stated above, these are only a sampling of GNSS set selection parameters. One must carefully review the requirements of the user application prior to selecting a receiver. In most cases, the selection will be a trade-off that requires awareness of the impact of any GNSS set deficiencies for the intended application.

3.5 GPS Geodesy and Time Scale

3.5.1 Geodesy

3.5.1.1 The GPS ECEF Reference Frame: WGS 84

As discussed in Section 2.2.7, SATNAV system operators may run their own tracking networks and may establish their own ECEF reference frame. This is the case for

GPS. The ECEF reference frame inherent in the GPS broadcast orbits and clocks is the DoD's World Geodetic System 1984 (WGS 84) [80].

It is useful to understand that there have been six realizations of WGS 84 as of this edition. The original WGS 84 was used for the broadcast GPS orbit beginning January 23, 1987. WGS 84 (G730), where the G730 denotes GPS week, was used beginning on June 29, 1994. WGS 84 (G873) started on January 29, 1997. WGS 84 (G1150) began on January 20, 2002. WGS 84 (G1674) started on February 8, 2012, and the current frame, WGS 84 (G1762), was introduced on October 16, 2013. These reference frame realizations have brought the WGS 84 into extremely close coincidence with the International Terrestrial Reference Frame (ITRF) described in Section 2.2.7. The RMS accuracy between WGS 84 (G1762) and the ITRF2008 frame is 1-cm overall [80].

The fact that there have been six realizations of WGS 84 has led to some confusion regarding the relationship between WGS 84 and other reference frames. In particular, care must be used when interpreting older references. For example, the original WGS 84 and the North American Datum 1983 (NAD 83) were made coincident [81], leading to an assertion that the WGS 84 and NAD 83 frames were identical. However, as stated above, WGS 84 (G1762) is coincident with ITRF2008. It is known that NAD 83 is offset from ITRF2008 by about 2.2m. Hence, the NAD 83 reference frame and the current realization of WGS 84 can no longer be considered identical. The National Geodetic Survey, NOAA is working towards a new reference frame to replace NAD 83. It has been anticipated that this new frame will be available in 2022 and will be aligned with the latest ITRF.

WGS 84 also defines its own ellipsoid. Quantities suitable for use with coordinate conversion by Table 2.1 are provided in Table 3.11.

It should be noted that this ellipsoid is extremely close, but not identical, to the Geodetic Reference System 1980 (GRS 80) ellipsoid described in Section 2.2.7. These GRS 80 and the WGS 84 ellipsoids only differ by 0.1 mm in the semiminor axis, b.

The GPS CNAV message, Type 32, transmits the Earth orientation components described in Section 2.2.2.2 [29]. This supports transformations between ECI and ECEF frames. For most terrestrial applications one may solve the GPS navigation problem in ECEF as discussed in Section 2.2.2.

3.5.2 Time Systems

3.5.2.1 GPS System Time

Each SATNAV system maintains its own internal reference time scale. For GPS, this is referred to as GPS system time (see Section 2.1). GPS system time is a "paper" time scale based on statistically processed readings from the atomic clocks in the

Table 3.11 Quantities for the WGS 84 Ellipsoid

Semimajor axis	$a = 6{,}378.137$ km
Semiminor axis	$b = 6{,}356.7523142$ km
Square eccentricity	$e^2 = 0.00669437999014$
Square second eccentricity	$e'^2 = 0.00673949674228$

GPS satellites and at various GPS ground control segment components. GPS system time is a continuous time scale that is not adjusted for leap seconds.

3.5.2.2 UTC(USNO)

As mentioned in Section 2.7.2, each SATNAV system disseminates a realization of UTC. The U.S. Naval Observatory (USNO) supports GPS by providing its underlying UTC timing reference. This form of UTC is denoted as UTC(USNO). GPS system time and UTC(USNO) were coincident at 0 hours, January 6, 1980. At the time of this writing, GPS system time led UTC(USNO) by 18 seconds. The GPS control segment is required to steer GPS system time within 40 ns (95%) of UTC(USNO) (modulo 1 s), but real performance has been better than 2 ns (modulo 1 second) for the past 15 years (<750 ps since November 2010 [82]. An epoch in GPS system time is distinguished by the number of seconds that have elapsed since Saturday/Sunday midnight and the GPS week number. GPS weeks are numbered sequentially and originate with week 0, which began at 0 hours, January 6, 1980 [29].

Receiver Computation of UTC(USNO)

Static Users

It can be observed from (2.44) that if the user's position (x_u, y_u, z_u) and satellite ephemerides (x_1, y_1, z_1) are known, then a static receiver can solve for t_u by making a single pseudorange measurement, ρ_1. Once t_u is determined, it can be subtracted from the receiver clock time, t_{rcv}, to obtain GPS system time, t_E. (Note that in the development of the user position solution in Section 2.5, GPS system time was denoted as T_u, which represented the instant in system time when the satellite signal reached the user receiver. However, we need to represent GPS system time at any particular time and will use the parameter t_E to do so.)

Expressing receiver clock time at any particular time:

$$t_{rcv} = t_E + t_u$$

So that:

$$t_E = t_{rcv} - t_u$$

From IS-GPS-200 [29], UTC(USNO), t_{UTC}, is computed as follows:

$$t_{UTC} = t_E - \Delta t_{UTC}$$

where Δt_{UTC} represents the number of integer leap seconds Δt_{LS} and a fractional estimate of the difference between GPS system time and UTC(USNO) modulo 1 s denoted herein as δt_A. [The control segment provides polynomial coefficients (a_0, a_1, and a_2) in the navigation data message that are used to compute the fractional difference between GPS system time and UTC(USNO) [29].]

Therefore, UTC(USNO), t_{UTC}, can be computed by the receiver as follows:

$$
\begin{aligned}
t_{UTC} &= t_E - \Delta t_{UTC} \\
&= t_{rcv} - t_u - \Delta t_{UTC} \\
&= t_{rcv} - t_u - \Delta t_{LS} - \delta t_A
\end{aligned}
$$

Mobile Users

Mobile users compute UTC(USNO) using the exact methodology described above except that they need to solve the system of (2.44) through (2.47) to determine the receiver clock offset, t_u.

3.6 Services

GPS is a dual-use system. That is, it provides separate services for civil and military users. These are called the Standard Positioning Service (SPS) and the Precise Positioning Service (PPS), respectively. The SPS is designated for the civil community and, at the time of this writing, was the predominant satellite navigation service in use by millions throughout the world. The PPS is available primarily to the military of the United States and its allies for users properly equipped with PPS receivers [5]. Access to the GPS PPS is controlled through cryptography.

The U.S. government guarantees specific levels of performance for both the SPS and PPS. These performance levels are formally documented in the SPS Performance Specification [3] and PPS Performance Specification [5].

As we will see in later parts of this book, in particular Chapter 11, GNSS position and time accuracy is a function of error contributions from all three system segments: space, control and user. In most cases, only the space and control segment error contributions are under the GNSS provider control. The reason being is that user equipment (i.e., the GNSS receiver) can range from inexpensive single chip devices for cellphone use to high precision receivers for survey applications. In light of this, the U.S. government only guarantees the accuracy and integrity of the GPS signal-in-space (SIS). Key attributes of both the SPS and PPS performance standards (PSs) are provided next.

3.6.1 SPS Performance Standard

3.6.1.1 Assumptions

This SPS PS [3] is conditioned upon certain assumptions regarding use of the SPS SIS. The following assumptions have been extracted from [3]:

- SPS user: This SPS PS assumes a SPS user with a SPS receiver. This SPS PS assumes the GPS receiver complies with the technical requirements related to the interface between the Space Segment and SPS receivers as established by IS-GPS-200 [29].
- C/A code: This SPS PS assumes the GPS receiver is tracking, processing, and using the C/A code signals transmitted by the GPS satellites. Pseudorange

measurements are assumed to be made by C/A code tracking with an early-minus-late correlator at 1-chip spacing using an exact replica of the waveform within an ideal sharp-cutoff filter bandwidth at 24 MHz with linear phase centered at the L1 frequency. Carrier phase measurement processing is not assumed.

- Single-frequency operation: This SPS PS assumes a GPS receiver which only has the hardware capability to track and use the C/A code signals transmitted by the satellites on L1. The performance standards in Section 3 of [3] are independent of whether the GPS receiver uses the satellite-transmitted ionospheric parameters for single-frequency model-based ionospheric delay compensation purposes or not. This SPS PS assumes that a GPS receiver will apply the single-frequency group delay time correction (T_{GD}) term in accordance with IS-GPS-200 [29].

3.6.1.2 SPS SIS URE Accuracy

Table 3.4-1 from [3] contains the SPS SIS URE accuracy standards. The following are those excerpts from this table that are a function of Age of Data. (AOD is the time between fresh uploads of SV clock offset and ephemeris data from the Control Segment to the SV.)

- Overall AODs: <= 7.8m 95% global average URE during normal operations;
- At Zero AOD: <= 6.0m 95% global average URE during normal operations;
- At Any AOD: <= 12.8m 95% global average URE during normal operations.

Note: The reader is referred to [3] to obtain additional performance standards (e.g., availability and integrity) for other conditions and constraints.

3.6.1.3 GPS Constellation Geometry

The conditions and constraints are:

- Defined for a position/time solution meeting the representative user conditions and operating within the service volume over any 24-hour interval;
  - PDOP availability standards: ≥ 98% global PDOP of 6 or less and ≥ 88% worst site PDOP of 6 or less.

3.6.1.4 SPS Position/Time Accuracy Standards

The conditions and constraints are:

- Defined for a position/time solution meeting the representative user conditions and operating within the service volume over any 24-hour interval;

- Standards based on a measurement interval of 24 hours averaged over all points in the service volume;
  - Global average position domain accuracy: <= 9m 95% horizontal error and <= 15m 95% vertical error;
  - Worst site position domain accuracy: <= 17m 95% horizontal error and <= 37m 95% vertical error;
  - Time transfer domain accuracy: <= 40 ns time transfer error 95% of time (SIS only).

Measured SPS Data

SPS measured data is contained in the GPS SPS Performance Analysis Reports [83], which are published quarterly by the U.S. Federal Aviation Authority (FAA). This report contains measured data of the following performance categories stated in [3]: PDOP Availability Standard, Service Availability Standard, Service Reliability Standard, and Positioning, Ranging and Timing Accuracy Standard.

Measured data is collected at 28 Wide Area Augmentation System (WAAS) reference station locations: Bethel, Alaska; Billings, Montana; Fairbanks, Alaska; Cold Bay, Alaska; Kotzebue, Alaska; Juneau, Alaska; Albuquerque, New Mexico; Anchorage, Alaska; Boston, Massachusetts; Washington, D.C.; Honolulu, Hawaii; Houston, Texas; Kansas City, Kansas; Los Angeles, California; Salt Lake City, Utah; Miami, Florida; Minneapolis, Minnesota; Oakland, California; Cleveland, Ohio; Seattle, Washington; San Juan, Puerto Rico; Atlanta, Georgia; Barrow, Alaska; Merida, Mexico; Gander, Canada; Tapachula, Mexico; San Jose Del Cabo, Mexico; and Iqaluit, Canada.

Measured SPS URE Data

For the period from April 1, 2010, through March 31, 2016, and the constraints and conditions cited in [3], it can be observed from Figure 3.27 that the maximum measured URE varied from 3.13m to 5.96m. This data was obtained from quarterly FAA GPS Performance Analysis (PAN) Reports #70 through #93 [83]. These reports can be found at http://www.nstb.tc.faa.gov.

Measured SPS Position and Time Data

During the same time period that the URE was measured, the following position and data were measured and provided in the PAN reports for the conditions and constraints cited in [3]:

- Global average position domain accuracy, horizontal error;
- Global average position domain accuracy, vertical error;
- Worst site position domain accuracy, horizontal error;
- Worst site position domain accuracy, vertical error;
- Time transfer domain accuracy (SIS only).

These data sets are shown in Figure 3.28.

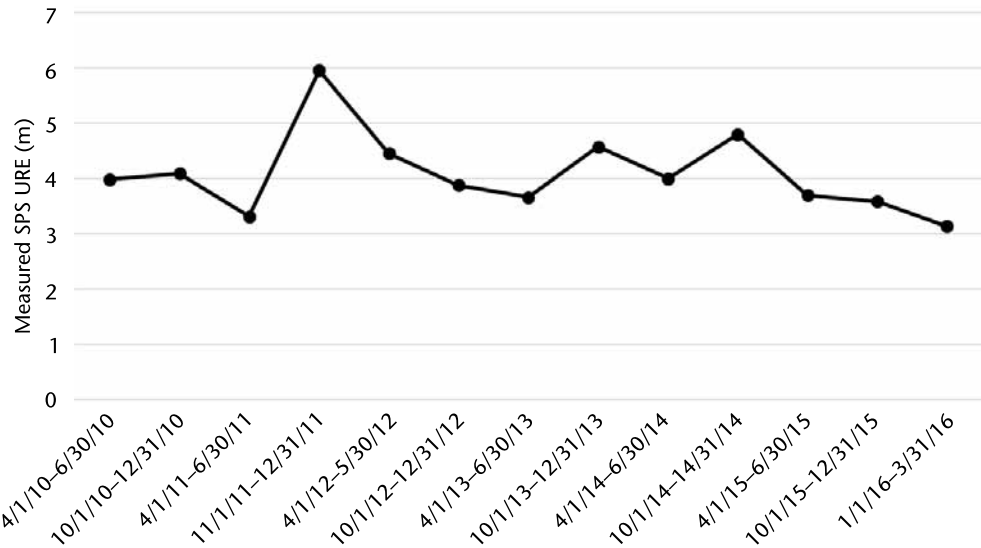


Figure 3.27 Measured maximum SPS URE [83].

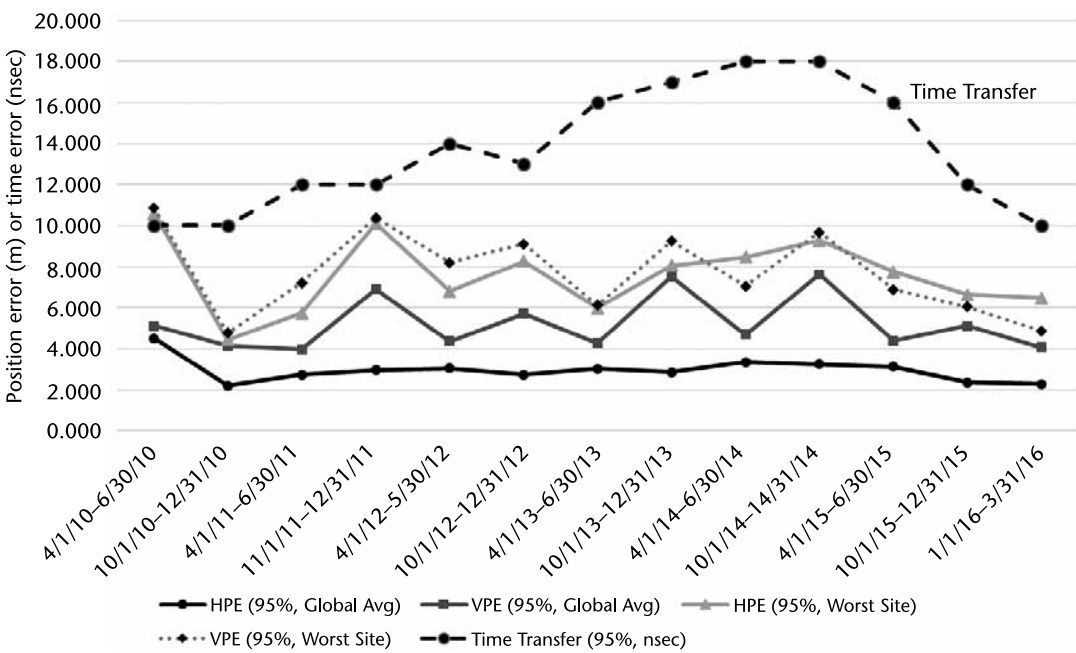


Figure 3.28 Measured SPS maximum position and time error [83].

3.6.2 PPS Performance Standard

The PPS PS defines the levels of SIS performance to be provided by the DoD to the authorized PPS user community. It is established to provide a basis for certification of PPS receivers for use in aviation Instrument Flight Rules (IFR) and to establish a minimum performance level that the GPS constellation must sustain [5].

3.6.2.1 Assumptions

There are numerous assumptions made for PPS user operation. These are contained in Section 2.4 of [5]. A few key assumptions are:

- An authorized user with a keyed GPS receiver. Specifically, the GPS receiver is assumed to contain current valid PPS keys and have the requisite hardware/software capabilities to be able to properly use those PPS keys.
- The GPS receiver is tracking and using the Y-code signals transmitted by the satellites for best PVT solution purposes.
- A keyed GPS receiver, which has the hardware capability to track and use the P(Y)-code signals transmitted by the satellites on L1 and on L2, will track and use both signals for dual-frequency, measurement-based ionospheric delay compensation purposes.
- A GPS receiver that only has the hardware capability to track and use the P(Y)-code signals transmitted by the satellites on L1 will track and use that signal for PVT solution purposes and the receiver will use the satellite-transmitted ionospheric parameters for single-frequency model-based ionospheric delay compensation purposes.
- The GPS receiver will track healthy SVs as defined in [5].
- This PPS PS does not take into consideration any error source that is not under direct control of the space segment or control segment. These excluded errors are listed in Section 2.4.5 of [5] and include receiver noise, multipath as well as receiver tropospheric delay compensation.

3.6.2.2 PPS Accuracy Standards

PPS SIS URE accuracy standards in Table 3.4-1 of [5]. A subset of these are:

- Dual-frequency operation conditions and constraints for any satellite marked as healthy in the NAV message: SIS Accuracy Standard, ≤ 5.9m 95% global average URE during normal operations over all AODs, ≤ 2.6m 95% global average URE during normal operations at zero AOD, and ≤11.8m 95% global average URE during normal operations at any AOD;
- Single frequency operation conditions and constraints for any satellite marked as healthy in the NAV message: neglecting single-frequency ionospheric delay model errors and including group delay time correction (T_{GD}) errors at L1;
  - SIS Accuracy Standard: ≤ 6.3m 95% global average URE during normal operations over all AODs, ≤ 5.4m 95% global average URE during normal operations at zero AOD, and ≤ 12.6m 95% global average URE during normal operations at any AOD;
  - Time transfer domain accuracy (dual- or single-frequency P(Y)-code). Note this is also defined as the UTC Offset Error (UTCOE) Accuracy. As stated in [5]: "The PPS SIS UTC(USNO) time accuracy is defined to be the statistical difference, at the 95th percentile, between the parameters

contained in the PPS SIS which relate GPS time to UTC as maintained by the USNO and the true value of the difference between GPS time and UTC(USNO). Also known as the UTC Offset Error (UTCOE)."

- Conditions and constraints for any satellite marked as healthy in the NAV message, PPS SIS UTCOE Accuracy Standard: <= 40 ns time transfer error 95% of time.

PPS Position and Time Accuracy Standards

There is no change to the PDOP definition as provided in the SPS PS mentioned above (i.e., the definition of PDOP is the same for both SPS and PPS users). However, the U.S. government does not commit to providing specific PPS position and time accuracies. It is up to the user to compute DOP based on his or her location and time of day and the appropriate User Equivalent Range Error (UERE) value that represents the UE configuration such that the user's predicated accuracy can be determined. Chapter 10 provides examples of UERE while user position and UTC(USNO) determination is given in Chapter 11.

Measured PPS URE Data

Figure 3.29 shows four PPS measured URE data sets for July 2016 [84]. It can be observed that all of these data sets are within the PPS PS accuracy standard of ≤ 5.9m 95% global average URE during normal operations over all AODs. It is important to note that the AF and NGA monitoring station measured PPS URE is subtracted from NGA precise ephemeris and SV clock offset data. As stated in Section 3.3.1.1, this NGA data serves as the reference or truth. From these curves, a decrease in URE can be observed, and is in part attributed to the L-AII and AEP upgrades and modeling improvements cited in Section 3.3.2, but also to launch replenishment of older, less stable satellites with newer, more stable clock technology. Of these four curves, two show the contributions from the worst SV in the constellation (i.e., greatest URE contributor) in terms of the 95% error as well as the RMS error. The other two curves depict the same error representations but for the entire constellation. Nonetheless, the steady decrease in URE is realized in greater user position and timing accuracy.

Figure 3.30 shows the URE contribution from each SV for July 2016 [84]. It can be observed that SVN 44 contributes the maximum URE while SVN 55 the least. The primary reason for the variable URE is that some SVs have better performing clocks. These variations can also be seasonal as current models tend to break down a bit during eclipse. Here the 95th percentile SIS URE across the entire constellation is 0.971m, with the RMS URE being 0.491m.

3.7 GPS Signals

This section describes the navigation signals transmitted by the GPS satellites including the legacy signals, which are the set of signals broadcast by every operational GPS satellite since the first satellite was launched in 1978. As described earlier in this chapter, the GPS constellation is being modernized. This section additionally

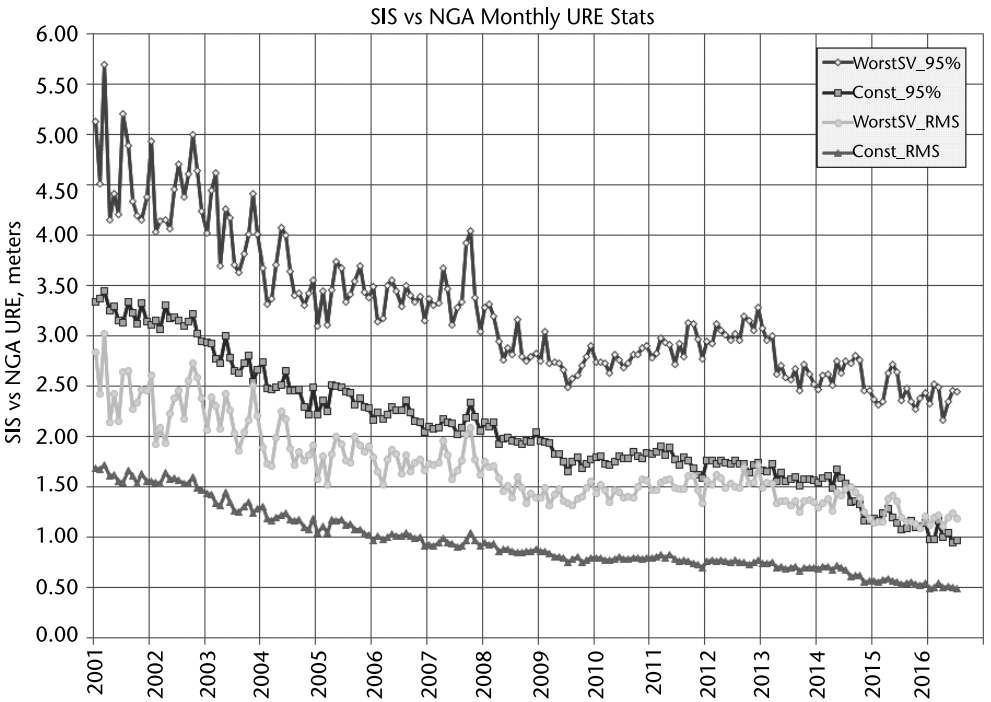


Figure 3.29 Measured PPS URE data as of July 2016 [84].

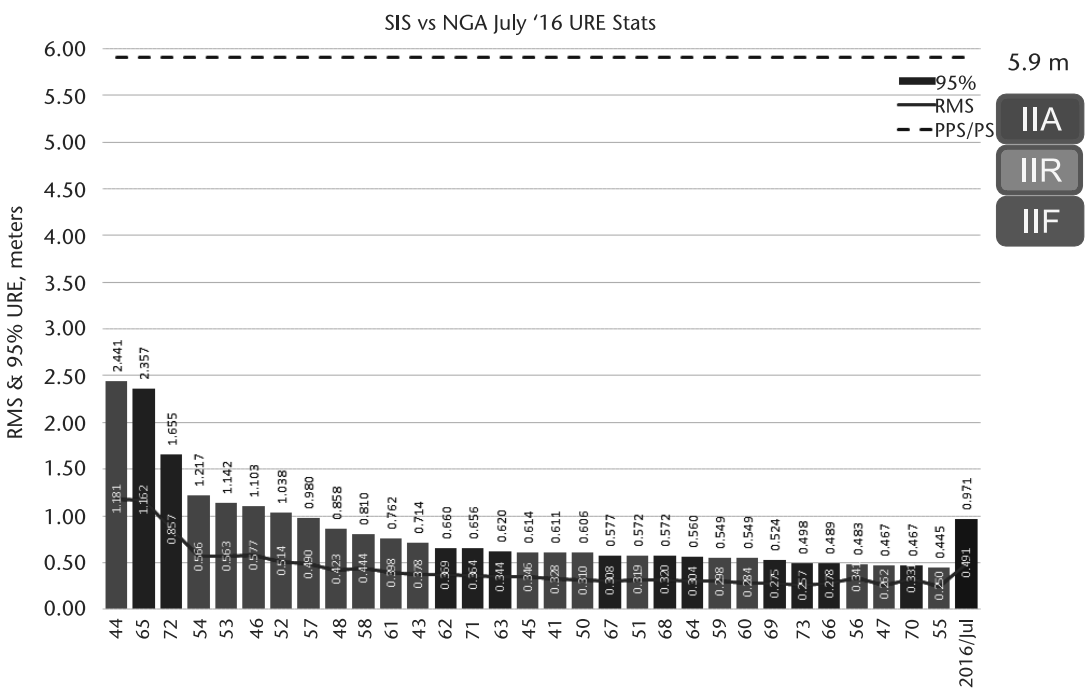


Figure 3.30 Measured PPS URE data as of July 2016: satellite ranking [84].

describes the modernized navigation signals that are being introduced into the constellation. The navigation data structures and contents for both the legacy and modernized signals are also addressed.

3.7.1 Legacy Signals

The legacy GPS SVs transmit navigation signals on two carrier frequencies called Link 1 (L1), the primary frequency, and Link 2 (L2), the secondary frequency. L1 is at 1,575.42 MHz and L2 is at 1,227.6 MHz. The two carrier frequencies were selected to be several hundred megahertz apart so that user equipment can estimate the delays experienced by the signals as they pass through the ionosphere. The carrier frequencies and modulation waveforms are all coherently generated onboard each GPS SV using a common frequency source that is driven by a rubidium or cesium AFS. The nominal reference frequency f_0 as it would appear to an observer on or near the ground is 10.23 MHz, but is set to run at a slightly lower frequency as would be seen by an observer moving with the SV to compensate for relativistic effects. The output of the SV's frequency standard (as it appears to an observer moving with the SV) is 10.23 MHz offset by a $\Delta f/f$ of 4.467×10^{-10}. This results in a Δf of 4.57×10^{-3} Hz and f_0 = 10.22999999543 MHz [29]. In the remainder of this section, all frequency values that are presented are with reference to how they would appear to the user on or near the ground. (Section 10.2.3 provides a detailed treatment of GNSS relativistic effects and associated compensation techniques.)

The carrier frequencies are modulated by spreading waveforms with a unique PRN sequence associated with each SV and also by navigation data. All GPS SVs transmit at the same carrier frequencies, but their signals do not interfere significantly with each other because of the PRN code modulation properties. Since each SV is assigned a unique PRN code and all PRN code sequences are nearly uncorrelated with respect to each other, the SV signals can be separated and detected. As discussed in Section 2.4.2.2, this technique of sharing a common carrier frequency amongst multiple transmitters (SVs) is referred to as code division multiple access (CDMA).

As shown in Figure 3.31, within the legacy GPS SVs the L1 frequency (154 f_0) is modulated by two PRN ranging codes: the coarse/acquisition code (C/A code) and the precision code (P code). The P code is encrypted when the GPS SV is in the antispoof (A-S) mode of operation, which is encountered almost always at the present time. The encrypted P code is referred to as the Y code, but it is common to refer to the P code in either mode of operation (A-S on or A-S off) as the P(Y) code. The C/A code has a chipping rate of 1.023 Mchips/s (= $f_0/10$) and the P code has a chipping rate of 10.23 Mchips/s (= f_0). Both the C/A code and P(Y) code on L1 are additionally modulated by 50 bps navigation data.

As shown in Figure 3.31, on legacy (e.g., Block II/IIA/IIR) SVs, the L2 frequency (120 f_0) can be modulated at any given time by: (1) P(Y) code with navigation data, (2) P(Y) code without navigation data, or (3) C/A code with navigation data. Of these choices, the P(Y) code with navigation data setting is most common. On the Block IIR-M and later GPS SVs, both legacy codes (C/A and P(Y)) can be broadcast with or without navigation data on L2 [29], but the most frequently encountered mode has only one legacy code, P(Y), with 50-bps navigation data, on L2 (as well as modernized signals to be described later in this section).

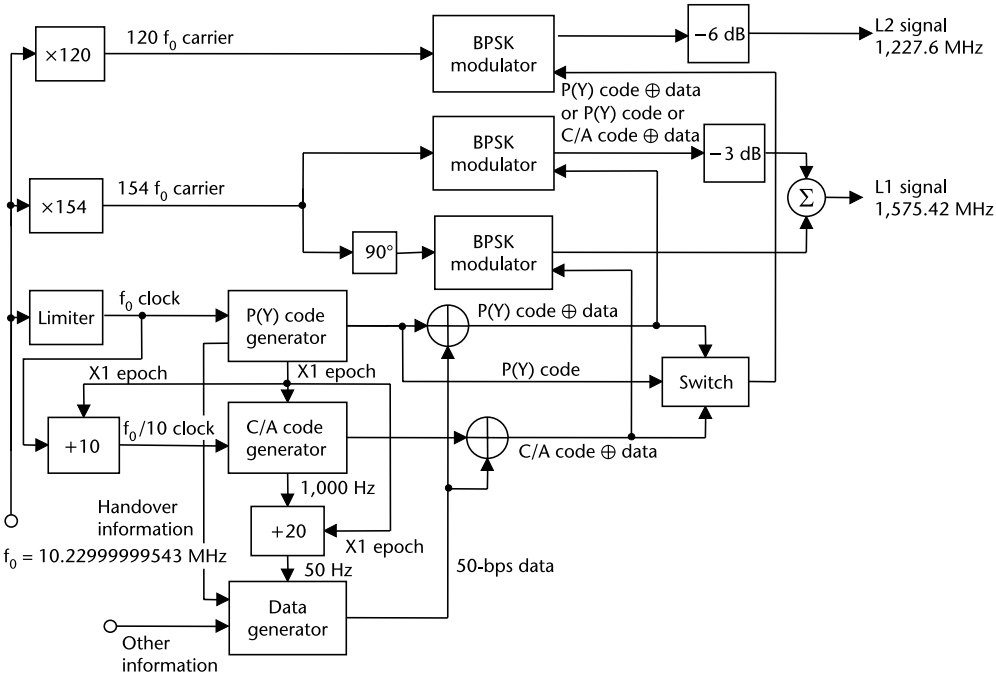
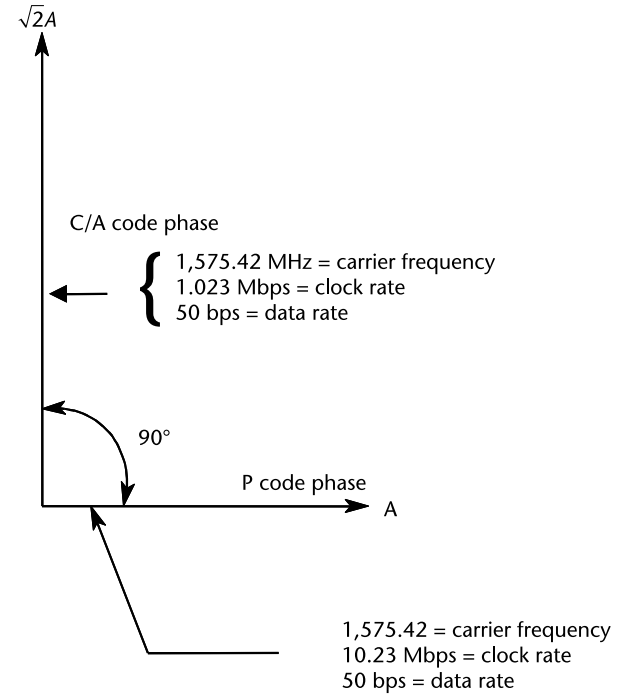


Figure 3.31 Legacy GPS (Block II/IIA/IIR) satellite signal synthesis.

The P(Y) code is ostensibly only available to PPS users when it is encrypted in the A-S mode of the SV. However, some civilian user equipment has been designed with the capability of tracking the encrypted P(Y) code. The techniques used in such equipment are referred to as either codeless or semi-codeless processing, and are discussed in Section 8.7.4.

In the past, both the C/A code and the P(Y) code as well as the L1 and L2 carrier frequencies were subjected to an encrypted, time-varying frequency offset (referred to as dither) plus an encrypted ephemeris and almanac offset error (referred to as epsilon) called selective availability (SA). SA denied the full accuracy of GPS to the standalone SPS users. However, SA has been deactivated on all GPS satellites since May 1, 2000. The United States has no intent to use SA again [1], so this subject will not be discussed further.

Note, as shown in Figure 3.31, that the same 50-bps navigation message data is modulo-2 summed to both the C/A code and the P(Y) code prior to modulation with the L1 carrier. An exclusive-or logic gate is used for this modulation process, denoted by ⊕. Since the C/A code ⊕ data and P(Y) code ⊕ data are both synchronous operations, the bit transition rate cannot exceed the chipping rate of the PRN ranging codes. Also note that BPSK (see Section 2.4.2.1) is used to modulate the carrier signals with the PRN ranging codes and navigation data. The P(Y) code ⊕ data is modulated in phase quadrature with the C/A code ⊕ data on L1. As shown in Figure 3.31, the L1 carrier is phase-shifted 90° before being BPSK modulated by the C/A code ⊕ data. Then this result is combined with the attenuated output of the BPSK modulation of L1 by the P(Y) code ⊕ data. The vector phase diagram in Figure 3.32 illustrates the 3-dB amplitude difference and phase relationship between P code and C/A code on L1. Figure 3.33 illustrates the result of P code ⊕ data and



$$L_i(\omega_1 t)=A[P_i(t)\oplus D_i(t)]\cos(\omega_1 t)+\sqrt{2}A[G_i(t)\oplus D_i(t)]\sin(\omega_1 t)$$

Figure 3.32 Legacy GPS signal structure for L1.

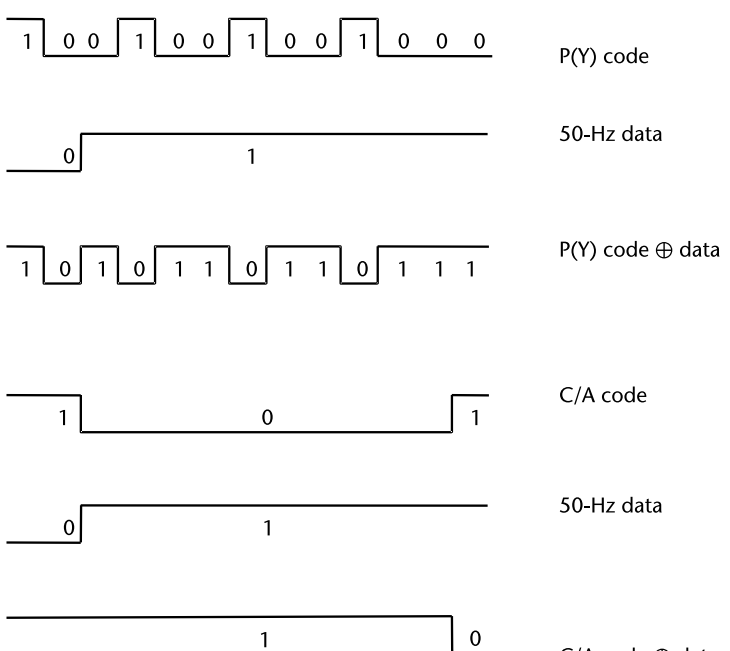


Figure 3.33 Modulo-2 summations of GPS ranging codes and navigation data.

C/A ⊕ data. As observed in Figure 3.33, the exclusive-or process is equivalent to binary multiplication of two 1-bit values yielding a one-bit product using the convention that logical 0 is plus and logical 1 is minus. There are 204,600 P(Y) code epochs between data epochs and 20,460 C/A code epochs between data epochs, so the number of times that the phase could change in the PRN code sequences due to data modulation is relatively infrequent, but the spectrum changes due to data modulation in the legacy GPS signals are very significant.

Figure 3.34 illustrates how the signal waveforms would appear before and after the BPSK modulation of one P(Y) code ⊕ data transition and one C/A code ⊕ data transition. There are 154 carrier cycles per P(Y) code chip and 1,540 carrier cycles per C/A code chip on L1, so the phase shifts on the L1 carrier are relatively infrequent. Although there are other SV modes (see [29]), the L2 frequency (1,227.60 MHz) is most commonly modulated only by one legacy signal, the P(Y) code ⊕ data. There are 120 carrier cycles per P(Y) code chip on L2, so the phase transitions on the L2 carrier are relatively infrequent. Table 3.12 summarizes the GPS signal structure on L1 and L2.

As mentioned in Section 3.2.3.2, there is more than one AFS in each SV for purposes of redundancy to improve reliability. For example, there are one cesium and two rubidium atomic standards on the Block IIF SVs. The CS selects only one atomic standard at a time to drive the reference frequency generator in the SV. Importantly, the reference frequency of 10.23 MHz has no relationship to the natural frequency of either a rubidium or cesium clock. Rather, this frequency was selected

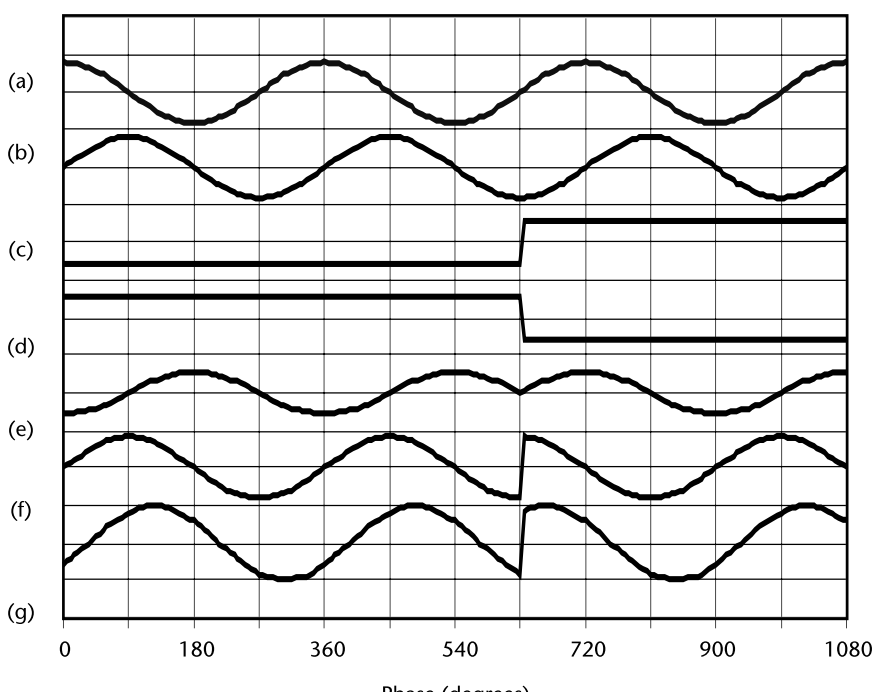


Figure 3.34 GPS L1 carrier modulation: (a) L1 carrier (0° phase), (b) L1 carrier (90° phase), (c) P(Y) code ⊕ data, (d) C/A code ⊕ data, (e) P(Y) code ⊕ data BPSK modulated on L1 carrier (0° phase) with 3-dB attenuation, (f) C/A code ⊕ data BPSK modulated on L1 carrier (90° phase), and (g) composite modulated L1 carrier signal.

Table 3.12 Legacy GPS Signal Structure

Signal designation	*L1*	*L2*
Carrier frequency (MHz)	1,575.42	1,227.60
PRN ranging codes (Mchips/s)	P(Y) = 10.23 and C/A = 1.023	P(Y) = 10.23 and/or C/A = 1.023*
Navigation message data nodulation (bps)	50	50**

*On the legacy SVs (Block II/IIA/IIR), only one legacy code, C/A or P(Y), can be modulated on L2. On newer SVs (Block IIR-M/IIF, GPS III), one or both of the legacy codes can be modulated on L2. However, for all GPS SVs, the most frequently encountered legacy signal configuration on L2 is P(Y) code. **The 50-Hz navigation data message is usually modulated on the legacy signal(s) on L2, but can be turned off in some available modes (see [29]).

so that the C/A ranging code of length 1,023 would repeat in a convenient interval of time (1 ms) when clocked at $f_0/10$.

3.7.1.1 PRN Ranging Code Generation

Figure 3.35 depicts a high-level block diagram of the direct sequence PRN ranging code generation used for GPS C/A code and P(Y) code generation to implement the CDMA technique. Each synthesized PRN code is derived from two other code generators. An exclusive-or circuit combines their outputs after the second code generator output is delayed with respect to the first. The amount of delay is variable. Associated with the amount of delay is the SV PRN number. In the case of P code, there were originally only 37 PRN codes with the integer delay in P-chips identical to the PRN number. In recent years, an expanded set of 26 P codes (PRNs 38–63) were added that may be generated by circularly shifting 26 of the original 37 sequences (over 1 week) by an amount corresponding to 1 day. For C/A code, the delay is unique to each SV, so there is only a table look-up relationship to the PRN number. These delays are summarized in Table 3.12 for PRNs 1 to 37. See [29] for the expanded sets of PRNs for P code and C/A code. The C/A code delay can be implemented by a simple but equivalent technique that eliminates the need for a delay register. This technique is explained in the following paragraphs.

The GPS C/A code is a Gold code [85] with a sequence length of 1,023 bits (chips). Since the chipping rate of the C/A code is 1.023 MHz, then the repetition period of the pseudorandom sequence is $1{,}023/(1.023 \times 10^6)$ or 1 ms. Figure 3.36 illustrates the design architecture of the GPS C/A code generator. Not included in this diagram are the controls necessary to set or read the phase states of the registers or the counters. There are two 10-bit shift registers, G1 and G2, which generate maximum length PRN codes with a length of $2^{10} - 1 = 1{,}023$ bits. (The one state that the shift register must not get into is the all-zero state.) It is common to describe the design of linear code generators by means of polynomials of the form $1 + \Sigma X^i$, where X^i means that the output of the ith cell of the shift register is used as the input to the modulo-2 adder (exclusive or) and the 1 means that the output of the adder is fed to the first cell [86]. The design specification for C/A code calls for the feedback taps of the G1 shift register to be connected to stages 3 and 10. These

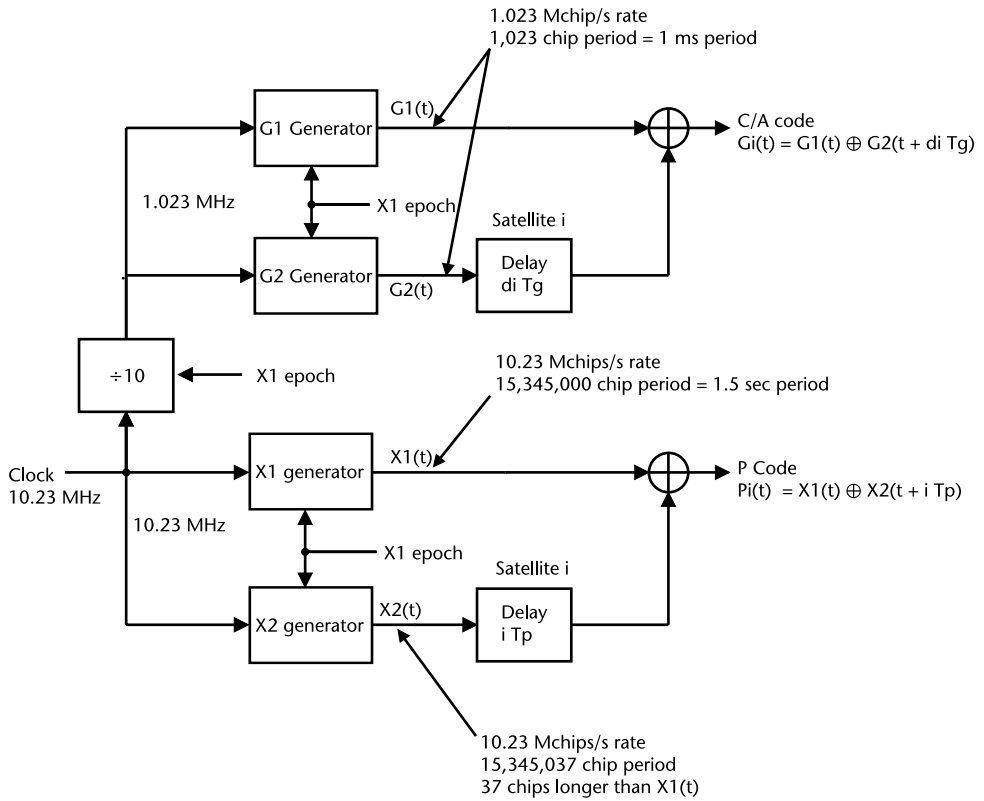


Figure 3.35 Legacy GPS signal ranging code generators.

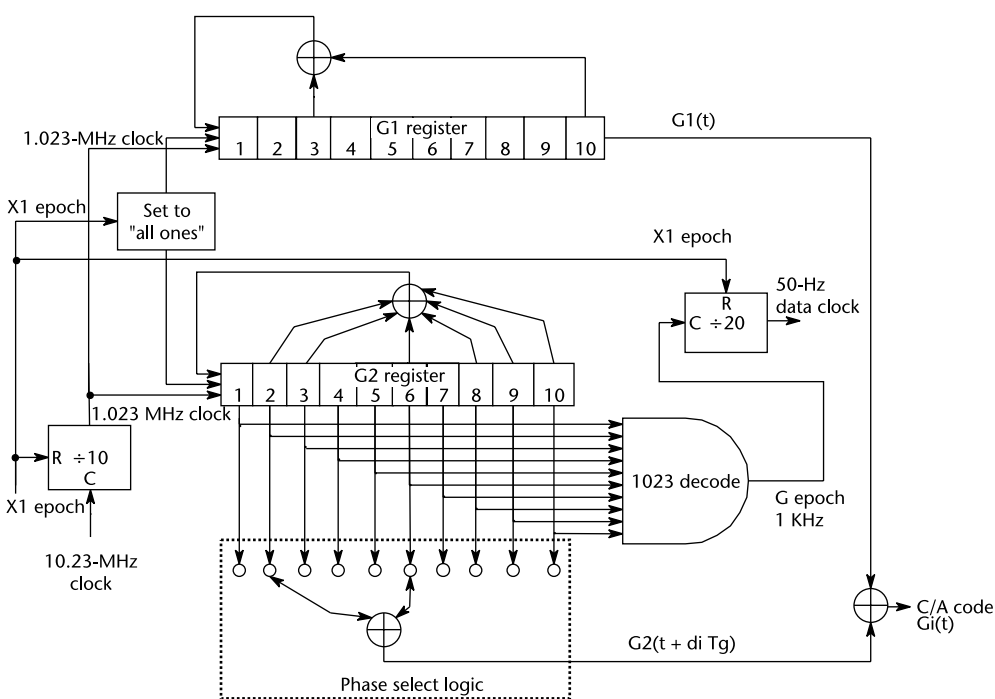


Figure 3.36 C/A code generator.

register states are combined with each other by an exclusive-or circuit and fed back to stage 1. The polynomial that describes this shift register architecture is: $G1 = 1 + X^3 + X^{10}$. The polynomials and initial states for both the C/A code and P code generator shift registers are summarized in Table 3.13. The G1 direct output sequence and the delayed version of the G2 direct output sequence are fed to an exclusive-or circuit that produces a unique C/A code for each SV. The equivalent delay effect in the G2 PRN code is obtained by the exclusive-or of the selected positions of the two taps whose output is called G21. This is because a maximum length PRN code sequence has the property that, added to a phase-shifted version of itself, it does not change but simply obtains another phase. The function of the two taps on the G2 shift register in Figure 3.36 is to shift the code phase in G2 with respect to the code phase in G1 without the need for an additional shift register to perform this delay. Each C/A code PRN number is associated with the two tap positions on G2. Table 3.13 describes these tap combinations for all defined GPS PRN numbers and also specifies the equivalent direct sequence delay in C/A code chips. The first 32 of these PRN numbers are reserved for the space segment. Five additional PRN numbers, PRN 33 to PRN 37, were originally reserved for other uses such as ground transmitters (also referred to as pseudo-satellites or pseudolites). At the time of this writing, only PRN 33 was reserved for such use. Pseudolites were used during Phase I (concept demonstration phase) of GPS to validate the operation and accuracy of the system before satellites were launched and in combination with the earliest satellites. C/A codes 34 and 37 are identical. The legacy C/A PRN codes have been expanded for GPS SV use and for numerous augmentation systems. See [29] for its modernized design details.

The GPS P code is a PRN sequence generated using four 12-bit shift registers designated X1A, X1B, X2A, and X2B. A detailed block diagram of this shift register architecture is shown in Figure 3.37 [29]. Not included in this diagram are the controls necessary to set or read the phase states of the registers and counters. Note that the X1A register output is combined by an exclusive-or circuit with the X1B register output to form the X1 code generator and that the X2A register output is combined by an exclusive-or circuit with the X2B register output to form the X2 code generator. The composite X2 result is fed to a shift register delay of the SV PRN number in chips and then combined by an exclusive-or circuit with the X1 composite result to generate the P code. With this shift register architecture, the P code sequence length would be more than 38 weeks in length, but is partitioned into 37 unique sequences that are truncated at the end of 1 week. Therefore, the sequence length of each PRN code is 6.1871×10^{12} chips and the repetition period is 7 days.

The design specification for the P code calls for each of the four shift registers to have a set of feedback taps that are combined by an exclusive-or circuit with each other and fed back to their respective input stages. The polynomials that describe the architecture of these feedback shift registers are shown in Table 3.14 and the logic diagram is shown in detail in Figure 3.37.

Referring to Figure 3.37, note that the natural cycles of all four feedback shift registers are truncated. For example, X1A and X2A are both reset after 4,092 chips, eliminating the last three chips of their natural 4,095 chip sequences. The registers X1B and X2B are both reset after 4,093 chips, eliminating the last two chips of their natural 4,095 chip sequences. This results in the phase of the X1B

Table 3.13 Code Phase Assignments and Initial Code Sequences for C/A Code and P Code

SV PRN Number	*C/A Code Tap Selection*	*C/A Code Delay (Chips)*	*P Code Delay (Chips)*	*First 10 C/A Chips (Octal)*[1]	*First 10 P Chips (Octal)*[1]
1	2 ⊕ 6	5	1	1,440	4,444
2	3 ⊕ 7	6	2	1,620	4,000
3	4 ⊕ 8	7	3	1,710	4,222
4	5 ⊕ 9	8	4	1,744	4,333
5	1 ⊕ 9	17	5	1,133	4,377
6	2 ⊕ 10	18	6	1,455	4,355
7	1 ⊕ 8	139	7	1,131	4,344
8	2 ⊕ 9	140	8	1,454	4,340
9	3 ⊕ 10	141	9	1,626	4,342
10	2 ⊕ 3	251	10	1,504	4,343
11	3 ⊕ 4	252	11	1,642	4,343
12	5 ⊕ 6	254	12	1,750	4,343
13	6 ⊕ 7	255	13	1,764	4,343
14	7 ⊕ 8	256	14	1,772	4,343
15	8 ⊕ 9	257	15	1,775	4,343
16	9 ⊕ 10	258	16	1,776	4,343
17	1 ⊕ 4	469	17	1,156	4,343
18	2 ⊕ 5	470	18	1,467	4,343
19	3 ⊕ 6	471	19	1,633	4,343
20	4 ⊕ 7	472	20	1,715	4,343
21	5 ⊕ 8	473	21	1,746	4,343
22	6 ⊕ 9	474	22	1,763	4,343
23	1 ⊕ 3	509	23	1,063	4,343
24	4 ⊕ 6	512	24	1,706	4,343
25	5 ⊕ 7	513	25	1,743	4,343
26	6 ⊕ 8	514	26	1,761	4,343
27	7 ⊕ 9	515	27	1,770	4,343
28	8 ⊕ 10	516	28	1,774	4,343
29	1 ⊕ 6	859	29	1,127	4,343
30	2 ⊕ 7	860	30	1,453	4,343
31	3 ⊕ 8	861	31	1,625	4,343
32	4 ⊕ 9	862	32	1,712	4,343
33[2]	5 ⊕ 10	863	33	1,745	4,343
34[2]	4 ⊕ 10[3]	950[3]	34	1,713[3]	4,343
35[2]	1 ⊕ 7	947	35	1,134	4,343
36[2]	2 ⊕ 8	948	36	1,456	4,343
37[2]	4 ⊕ 10[3]	950[3]	37	1,713[3]	4,343

1. In the octal notation for the first 10 chips of the C/A code as shown in this column, the first digit (1) represents a 1 for the first chip and the last three digits are the conventional octal representation of the remaining 9 chips. For example, the first 10 chips of the SV PRN number 1 C/A code are 1100100000.

2. PRN codes 33 through 37 are reserved for other uses (e.g., pseudolites).

3. C/A codes 34 and 37 are identical.

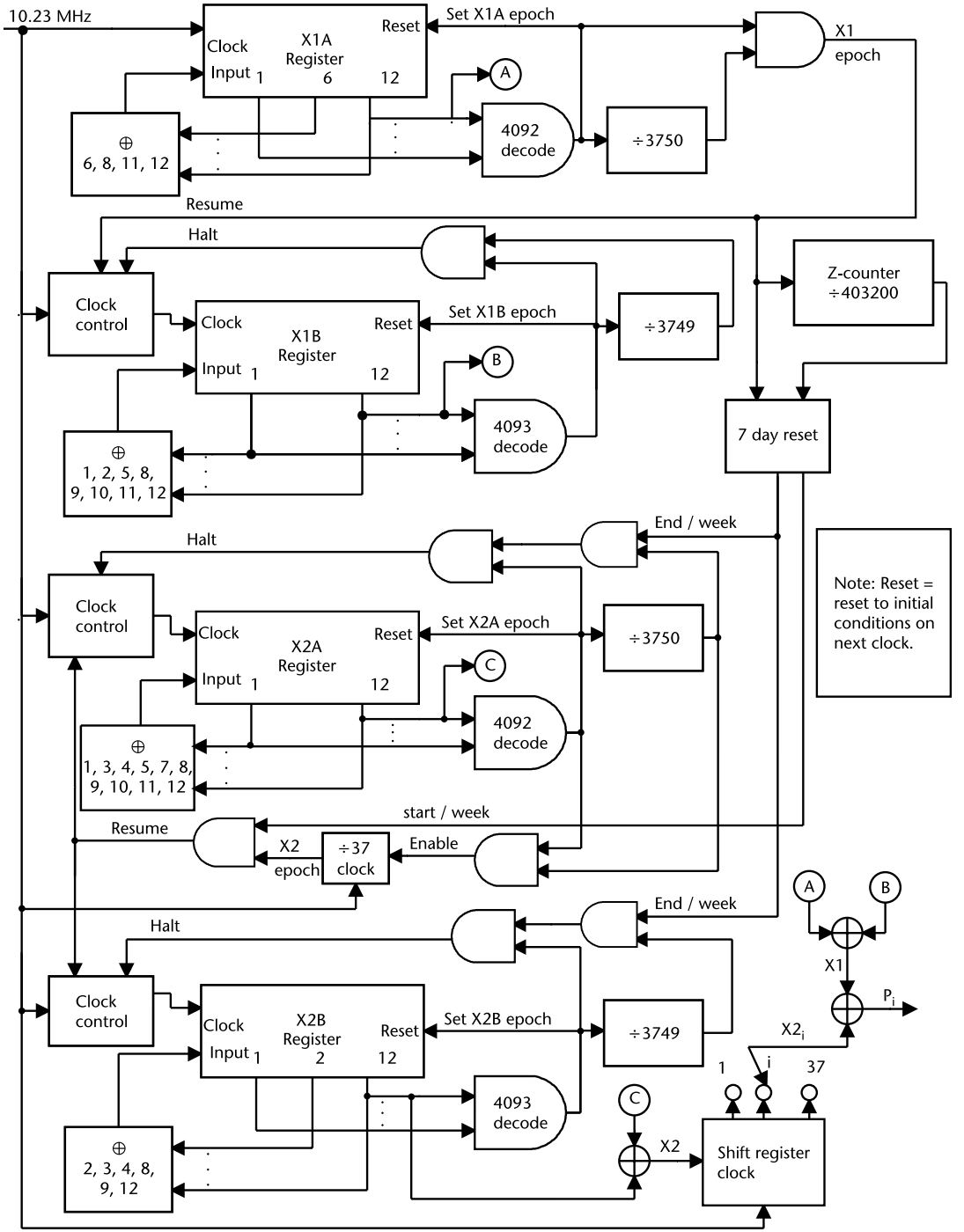


Figure 3.37 P code generator.

sequence lagging by one chip with respect to the X1A sequence for each X1A register cycle. As a result, there is a relative phase precession between the X1A and X1B registers. A similar phase precession takes place between X2A and X2B. At the beginning of the GPS week, all of the shift registers are set to their initial states

Table 3.14 GPS Code Generator Polynomials and Initial States

Register	*Polynomial*	*Initial State*
C/A code G1	$1 + X^3 + X^{10}$	1111111111
C/A code G2	$1 + X^2 + X^3 + X^6 + X^8 + X^9 + X^{10}$	1111111111
P code X1A	$1 + X^6 + X^8 + X^{11} + X^{12}$	001001001000
P code X1B	$1 + X^1 + X^2 + X^5 + X^8 + X^9 + X^{10} + X^{11} + X^{12}$	010101010100
P code X2A	$1 + X^1 + X^3 + X^4 + X^5 + X^7 + X^8 + X^9 + X^{10} + X^{11} + X^{12}$	100100100101
P code X2B	$1 + X^2 + X^3 + X^4 + X^8 + X^9 + X^{12}$	010101010100

simultaneously, as shown in Table 3.14. Also, at the end of each X1A epoch, the X1A shift register is reset to its initial state. At the end of each X1B epoch, the X1B shift register is reset to its initial state. At the end of each X2A epoch, the X2A shift register is reset to its initial state. At the end of each X2B epoch, the X2B shift register is reset to its initial state. The outputs (stage 12) of the A and B registers are combined by an exclusive-or circuit to form an X1 sequence derived from X1A $\oplus$ X1B, and an X2 sequence derived from X2A $\oplus$ X2B. The X2 sequence is delayed by i chips (corresponding to SVi) to form X2i. The P code for SVi is Pi = X1 $\oplus$ X2i.

There is also a phase precession between the X2A/X2B shift registers with respect to the X1A/X1B shift registers. This is manifested as a phase precession of 37 chips per X1 period between the X2 epochs and the X1 epochs. The divide-by-37 counter shown in Figure 3.37 causes the X2 period to be 37 chips longer than the X1 period. The details of this phase precession are as follows. The X1 epoch is defined as 3,750 X1A cycles. When X1A has cycled through 3,750 of these cycles or $3{,}750 \times 4{,}092 = 15{,}345{,}000$ chips, a 1.5-second X1 epoch occurs. When X1B has cycled through 3,749 cycles of 4,093 chips per cycle or 15,344,657 chips, it is kept stationary for an additional 343 chips to align it to X1A by halting its clock control until the 1.5-second X1 epoch resumes it. Therefore, the X1 registers have a combined period of 15,345,000 chips. X2A and X2B are controlled in the same way as X1A and X1B, respectively, but with one difference: when 15,345,000 chips have completed in exactly 1.5 seconds, both X2A and X2B are kept stationary for an additional 37 chips by halting their clock controls until the X2 epoch (the output of the divide by 37 counter) or the start of the week resumes it. Therefore, the X2 registers have a combined period of 15,345,037 chips, which is 37 chips longer than the X1 registers.

Note that if the P code were generated by X1 $\oplus$ X2, and if it were not reset at the end of the week, it would have the potential sequence length of $15{,}345{,}000 \times 15{,}345{,}037 = 2.3547 \times 10^{14}$ chips. With a chipping rate of 10.23×10^6, this sequence has a period of 266.41 days or 38.058 weeks. However, since the sequence is truncated at the end of the week, each SV uses only 1 week of the sequence and 38 unique one-week PRN sequences are available. As in the case of C/A code, the first 32 PRN sequences were originally reserved for the GPS space segment and PRN 33 to 37 were reserved for other uses (e.g., pseudolites). The PRN 38 P code was sometimes used as a test code in P(Y)-code GPS receivers as well as to generate a reference noise level (since, by the original interface specification, it could not correlate with any used SV PRN signals). In recent years, however, as noted earlier, an expanded set of P codes have been selected (PRNs 38–63) using a 1-day delay of

the original PRN 1–26 ranging codes, and now only PRN 33 is reserved for other uses.

The unique P code for each SV is the result of the different delay in the X2 output sequence. Table 3.12 shows this delay in P code chips for each SV PRN number. The P code delays (in P code chips) are identical to their respective PRN numbers for the SVs, but the C/A code delays (in C/A code chips) are different from their PRN numbers. The C/A code delays are typically much longer than their PRN numbers. The replica C/A codes for a conventional GPS receiver are usually synthesized by programming the tap selections on the G2 shift register.

Table 3.12 also shows the first 10 C/A code chips and the first 12 P code chips in octal format starting from the beginning of the week. For example, the binary sequence for the first 10 chips of PRN 5 C/A code is 1001011011, and for the first 12 chips of PRN 5 P code is 100011111111. Note that the first 12 P code chips of PRN 10 to PRN 37 are identical. This number of chips is insignificant for P code, so the differences in the sequence do not become apparent until later in the sequence.

3.7.1.2 Power Levels

Table 3.15 summarizes the minimum received power levels for the three legacy GPS signals, not including rarely used GPS satellite modes that broadcast C/A code on L2. The levels are specified in terms of decibels with respect to 1W (dBW). The specified received GPS signal power [29] is based on the signal received by a user antenna that is linearly polarized with a 3-dB gain, normally rotated to achieve the greatest polarization mismatch loss. This corresponds to an ideal RHCP antenna with unity gain that is expressed as 0 dBic (meaning 0-dB gain with respect to an isotropic circularly polarized antenna). A linear polarized antenna is used in the specification because: (1) it is impossible to build a perfect RHCP antenna; (2) it is possible to build and calibrate a linear polarized antenna with gain calibration traceable to the International Bureau of Weights and Measures through National Metrology Institutions; and (3) with such a reference user antenna, any imperfections in the satellite antenna polarization characteristics will not result in power loss for the user as it would have if the user antenna was defined to be 0-dBic RHCP.

Figure 3.38 illustrates that the minimum received power is met when the SV is at two elevation angles: 5° from the user's horizon and at the user's zenith. In between these two elevation angles, the minimum received signal power levels gradually increase up to 2-dB maximum for the L1 signals and up to 1-dB maximum for the L2 signal and then decrease back to the specified minimums. This characteristic occurs because the shaped beam pattern on the SV transmitting antenna arrays can only match the required minimum gain at the angles corresponding to the center of the Earth and to near the edge of the Earth, resulting in slightly increasing transmitting antenna array gain in between these nadir angles. The user's antenna gain

Table 3.15 Minimum Received Legacy GPS Signal Power Levels

Satellite Block	*L1 C/A Code*	*L1 P(Y) Code*	*L2 P(Y) Code*
IIA/IIR	–158.5	–161.5	–164.5
IIR-M/IIF/III	–158.5	–161.5	–161.5

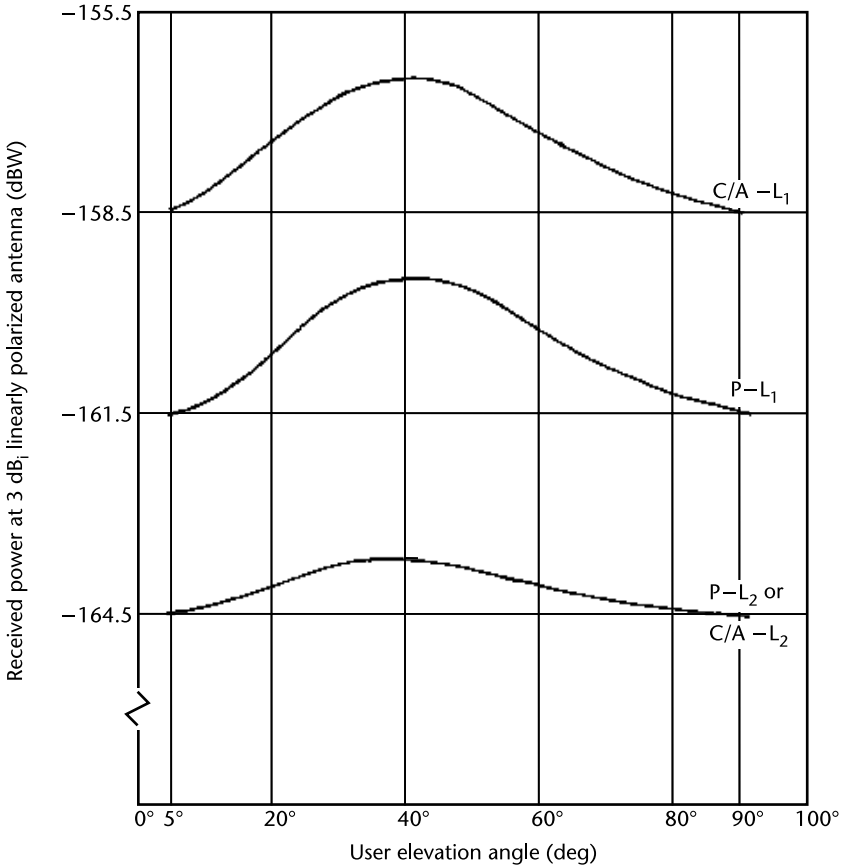


Figure 3.38 User received minimum signal power levels.

pattern is typically maximum at the zenith and minimum at 5° above the horizon and for lower elevation angles.

The received signal levels are not expected to exceed –153 dBW and –150.0 dBW, respectively, for the C/A code and P(Y) code components on the L1 channel and not expected to exceed –155.5 dBW for either signal on the L2 channel. In general, the signal powers for the SVs are at their maximum levels when the satellites are new and remain nearly constant until their end of life. The signal power variations above the guaranteed minimum power over the SV lifetime are therefore expected to be less than 5.5 dB, 11.5 dB, and 6 dB, respectively, for the L1 C/A code, L1 P(Y) code, and L2 P(Y) code (or L2 C/A code). Note that these are the Legacy maximum power limits. There are modernized flex power modes of operation that intentionally increase the P(Y) and M code powers in the newer SVs.

Table 3.16 tabulates the navigation satellite signal power budget for the Block II GPS satellites from [87] using the minimum user received power levels as the starting point. It shows the output power levels at the worst-case off-axis angle of 14.3° and for the assumed worst-case atmospheric loss of 0.5 dB. Referring to Table 3.16, the link budget for the L1 C/A code to provide the signal power with a unity gain transmitting antenna is: –158.5 – 3.0 +184.4 + 0.5 + 3.4 = 26.8 dBW. Since the satellite L1 antenna array has a minimum gain of 13.4 dB for C/A code at the worst case off-axis angle of 14.3°, the minimum L1 antenna transmitter power

Table 3.16 Block II SV L1 and L2 Signal Power Budget [87]

	L1 C/A Code	*L1 P Code*	*L2*
User minimum received power	–158.5 dBW	–161.5 dBW	–164.5 dBW
User linear antenna gain	3.0 dB	3.0 dB	3.0 dB
Free space propagation loss	184.4 dB	184.4 dB	182.3 dB
Total atmospheric loss	1.5 dB	1.5 dB	1.5 dB
SV polarization mismatch loss	3.4 dB	3.4 dB	4.4 dB
Required satellite EIRP	+26.8 dBW	+23.8 dBW	+19.7 dBW
SV antenna gain @ 14.3° worst-case off-axis angle	13.4 dB	13.5 dB	11.5 dB
Required minimum satellite antenna input power	+13.4 dBW, 21.88W	+10.3 dBW, 10.72W	+8.2 dBW, 6.61W

for C/A code is $\log^{-1}[(26.8 - 13.4)/10] = 21.9$W. Note that a minimum of 32.6W of L1 power and 6.6W of L2 power (for a total of 39.2W) must be delivered to the satellite antenna arrays to maintain the specification. The efficiency of the high-power amplifier (HPA) subassembly determines how much actual power must be provided in the satellite.

3.7.1.3 Legacy Navigation Data

As described earlier, both the C/A code and P(Y) code signals are modulated with the same 50-bps navigation data on both L1 and L2. This data provides the user with the information necessary to compute the precise locations of each visible satellite and time-of-transmission for each navigation signal. The data also includes a significant set of auxiliary information that may be used, for example, to assist the receiver in acquiring new satellites, to translate from GPS system time to UTC (see Section 3.5.2.2), and to correct for a number of errors that affect the range measurements. This section outlines the main features of the legacy GPS navigation (LNAV) message. For a more complete description, the interested reader is referred to [29].

The GPS LNAV navigation message is transmitted in five 300-bit subframes, as shown in Figure 3.39. Each subframe is itself composed of 10 30-bit words. The last 6 bits in each word of the navigation message are used for parity checking to provide the user equipment with a capability to detect bit errors during demodulation. A Hamming code [32, 26] is employed for error detection. The five subframes are transmitted in order beginning with subframe 1. Subframes 4 and 5 consist of 25 pages each, so that the first time through the five subframes, page 1 (of subframes 4 and 5) is broadcast. In the next cycle through the five subframes, page 2 is broadcast, and the cycling continues until page 25 is broadcast, and then the paging sequence of subframes 4 and 5 begins again. It requires 30 seconds to read all 5 pages and ($25 \times 30 = 750$ seconds) 12.5 minutes for the receiver to read all 25 pages, assuming no data dropouts.

Although there are provisions for a loss of ground contact, normally the CS uploads critical navigation data elements once or twice per day per satellite. In this nominal mode of operation, the same critical navigation data elements (e.g., satellite ephemeris and clock correction data) are broadcast repeatedly over 2-hour time

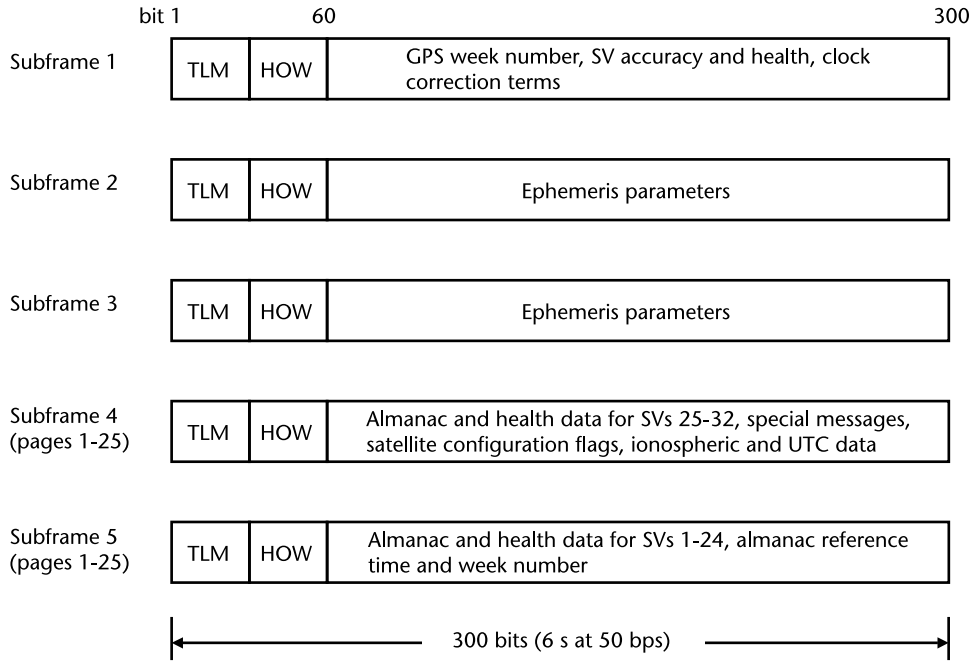


Figure 3.39 Legacy navigation message format.

spans (except if an upload occurs during this interval). On 2-hour boundaries, each satellite switches to broadcasting a different set of these critical elements, which are stored in tables in the satellite's random access memory. The CS generates these message elements based upon its current estimates of each satellite's position and clock error and prediction algorithms on how these parameters will change over time.

The first two words of each subframe (bits 1–60) contain telemetry (TLM) data and a handover word (HOW). The TLM word is the first of the 10 words in each subframe and includes a fixed preamble, a fixed 8-bit pattern 10001011 that never changes. This fixed-pattern preamble is included to assist the user equipment in locating the beginning of each subframe (called frame sync), but it must be tested for consistency in its location just in case the same bit pattern occurs elsewhere in the message. Each TLM word also includes 14 bits of data that are only meaningful to authorized users. The HOW, named because it allows the PPS user equipment to hand over from C/A code tracking to P(Y)-code tracking, provides the GPS time-of-week (TOW) modulo 6 s corresponding to the leading edge of the following subframe. The TOW provision in the HOW is also essential to the SPS user to remove the time ambiguity of the 1-ms C/A code period. The receiver must first determine where the data transition (20-ms) boundaries are located (called bit sync) to an accuracy much better than 1 ms before the HOW can be used reliably to establish the time to an ambiguity of 6 seconds. The HOW also provides two flag bits, one that indicates whether anti-spoofing is activated (see Section 3.7.1), and one that serves as an alert indicator. If the alert flag is set, it indicates that the signal accuracy may be poor and should be processed at the user's own risk. Lastly, the HOW provides the subframe number (1 to 5).

Subframe 1 provides the GPS transmission week number, which is the number of weeks modulo 1,024 that have elapsed since January 5, 1980. The first rollover of the GPS week number occurred on August 22, 1999. The next rollover will occur in April 2019. It is prudent that the GPS receiver designer keep track of these rare but inevitable rollover epochs in nonvolatile memory (see Section 3.5.2.1). Subframe 1 also provides the following satellite clock correction terms: a_{f0}, a_{f1}, a_{f2}, and time-of-clock, t_{oc}. These terms are extremely important for precise ranging since they account for the lack of perfect synchronization between the timing of the SV broadcast signals and GPS system time (see Section 10.2.1). A 10-bit number referred to as Issue of Data, Clock (IODC) is included in subframe 1 to uniquely identify the current set of navigation data. User equipment can monitor the IODC field to detect changes to the navigation data. The current IODC is different from IODCs used over the past 7 days. Subframe 1 also includes a group delay correction, T_{gd}, a user range accuracy (URA) indicator, a SV health indicator, an L2 code indicator, and an L2 P data flag. T_{gd} is needed by single-frequency (L1- or L2-only) users since the clock correction parameters refer to the timing of the P(Y) code on L1 and L2 as apparent to a user that is using a linear combination of dual-frequency L1/L2 P(Y) code measurements to mitigate ionospheric errors (see Sections 10.2.4.1 and 10.2.7.1). The URA indicator provides the user with an estimate of the 1-sigma range errors to the satellite due to satellite and CS errors (and is fully applicable only for L1/L2 P code users). The SV health indicator is a 6-bit field that indicates whether the satellite is operating normally or whether components of the signal or navigation data are suspected to be erroneous. The L2 code indicator field indicates whether the P(Y) code or C/A code is active on L2. Finally, the L2 P data flag indicates whether navigation data is being modulated onto the L2 P(Y) code.

Subframes 2 and 3 include the osculating Keplerian orbital elements-that allow the user equipment to precisely determine the location of the satellite.

Subframe 2 also includes a fit interval flag and an age of data offset (AODO) term. The fit interval flag indicates whether the orbital elements are based upon a nominal 4-hour curve fit (that corresponds to the 2-hour nominal data transmission interval described above) or a longer interval. The AODO term provides an indication of the age of the elements of a Navigation Message Correction Table (NMCT) that has been included in the GPS navigation data since 1995 [88]. Both subframes 2 and 3 also include an issue of data ephemeris (IODE) field. IODE consists of the 8 least significant bits (LSBs) of IODC and may be used by the user equipment to detect changes in the broadcast orbital elements.

Pages 2, 3, 4, 5, 7, 8, 9, and 10 of subframe 4 and pages 1 to 24 of subframe 5 contain almanac data (coarse orbital elements that allow the user equipment to determine approximate positions of other satellites to assist acquisition) for SVs 1 to 32 (see Table 20-VI of [29].). Page 13 of subframe 4 includes the NMCT range corrections. Page 18 of subframe 4 includes ionospheric correction parameters for single-frequency users (see Section 10.2.4.1) and parameters so that user equipment can relate UTC to GPS system time (see Section 3.5.2.2). Page 25 of subframes 4 and 5 provide configuration and health flags for SVs 1 to 32. The data payloads of the remaining pages of subframes 4 and 5 are currently reserved for military use. Historically, the data in these reserved subframes are all zeros when not activated,

but are encrypted and nonzero when activated for important military use. The public availability of the entire legacy navigation message on the L1 C/A signal has historically caused costly CS disruptions when these reserved subframes were activated for intended and important military purposes. This is because of some SPS receivers whose designers have disregarded the reserved warning by creating a receiver dependence on the dataless intervals in these subframes.

3.7.2 Modernized Signals

As illustrated in Figure 3.40, the modernized signals include three new civil signals, an L2 civil (L2C) signal [29, 89], a signal at 1,176.45 MHz (115 f_0) referred to as L5 [90, 91], and an additional signal at L1 called L1C [92]. The modernized military signal, called M code, is also added on L1 and L2 [93]. The L2C and M code signals were first implemented along with legacy GPS signals on the Block IIR-M satellites, so L5 is not included. The first Block IIR-M satellite was launched in 2005. Block IIR-M satellites were the first modernized versions of the Block IIR satellites that continued to support legacy GPS signals. The L5 modernized civil signal, often designated as the safety-of-life signal, was first included on the Block IIF satellites. The first Block IIF satellite was launched in 2010. The Block IIF satellites are the last of the Block II series and were intended to provide modernized signals until the advent of GPS III satellites (currently scheduled to be available for launch in 2018). The modernized L1 civil signal (L1C) will be available from GPS III and subsequent satellites.

Full monitoring, control, and navigation data for the modernized signals is not provided by the legacy CS. It must be either updated or replaced by the modernized CS OCX Block 2 becoming operational (not yet operational as of this edition). The L1C signal, like all modernized GPS signals, provides a pilot (dataless) component that improves carrier tracking. The L1C signal provides enhanced robustness and accuracy relative to the C/A signal, and enables improved interoperability with signals transmitted at the same carrier frequency by other satellite navigation systems.

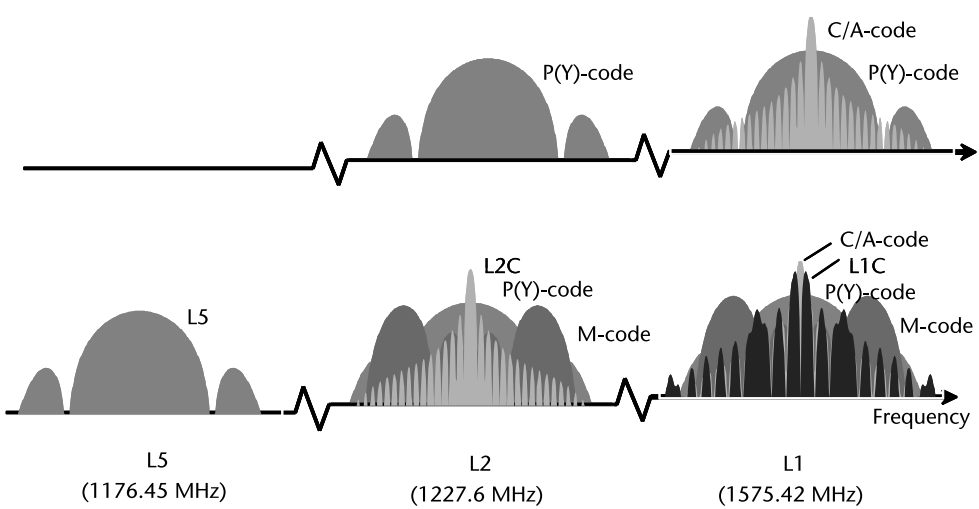


Figure 3.40 Legacy only (top) and legacy plus modernized (bottom) GPS signals.

3.7.2.1 L2C Signal

As shown in Figure 3.40, the L2 civil (L2C) signal uses the same BPSK-R(1) spreading modulation as the C/A signal. However, the L2C signal is very different from the C/A signal in many other ways. First, L2C uses two different PRN codes per signal. The first PRN code is referred to as the Civil Moderate (CM) code because it employs a PRN code that repeats every 10,230 chips, which is considered to be of moderate length. The second spreading code, the Civil Long (CL) code, is an extremely long code with a length of 767,250 chips. As shown in Figure 3.41, these two spreading codes are generated, each with a 511.5 kchip/s rate, and are used in the following manner to generate the overall L2C signal. First, a 25-bps navigation data stream modulates the CM code after the navigation data has been encoded into a 50-baud stream with a rate ½ forward error correction (FEC) code. The 25-bps data rate is one-half the rate of the navigation data on the C/A code and P(Y) code signals and was chosen so that the data on the L2C signal can be demodulated in challenged environments (e.g., indoors or under heavy foliage) where 50-bps data could not be. Next, the chip-by-chip multiplexing of the CM (with data) and CL codes forms the baseband L2C signal. The L2C signal has an overall chip rate of 2×511.5 kchip/s = 1.023 Mchip/s, needed for the BPSK-R(1) spreading modulation. There are important differences between the L2C and C/A code signal power spectra; however, since both CM and CL are much longer than the length-1,023 C/A code, the lines in the L2C power spectrum are spaced much more closely in frequency, and far lower in power, than the lines in the C/A code power spectrum. As will be discussed in Chapter 9, the lower lines in the L2C power spectrum lead to greatly increased robustness in the presence of narrowband interference.

The CM and CL codes are generated using the same 27-stage linear feedback shift register shown in Figure 3.42. A shorthand notation is used in the diagram. The number that appears in each block in the figure represents the number of stages (each holding 1 bit) between feedback taps. CM and CL codes for different

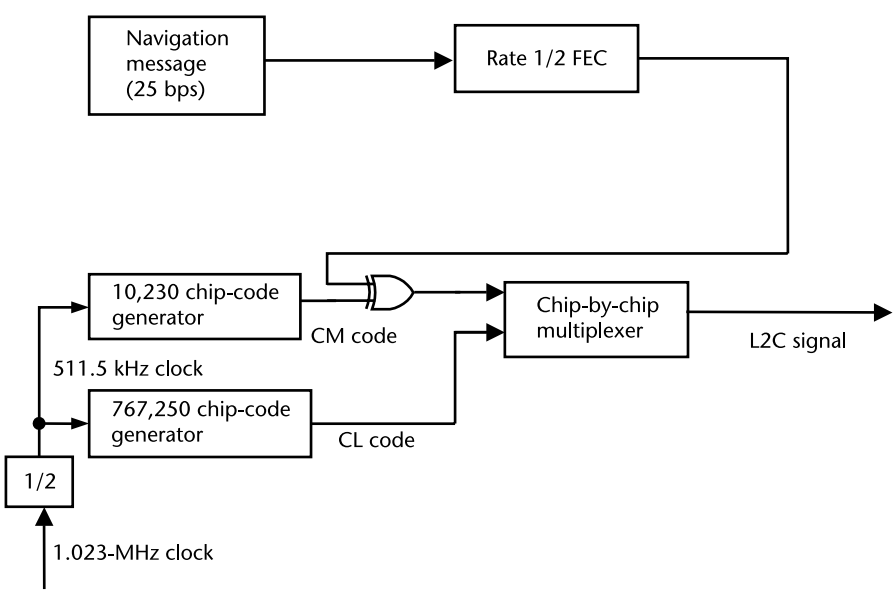


Figure 3.41 Baseband L2C signal generator.

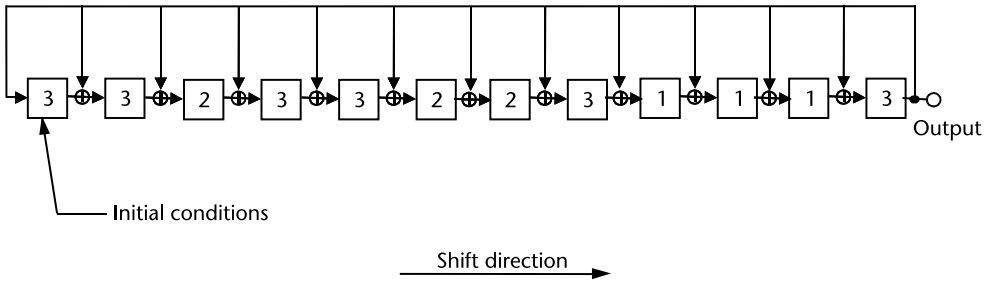


Figure 3.42 CM and CL PRN code generation.

satellites are generated by different initial loads of the register. The register is reset every 10,230 chips for CM and every 767,250 chips for CL. The CM code repeats 75 times for each repetition of the CL code. At the 511.5 kchip/s, the period of the CM code is 20 ms [one P(Y) code data bit period] and the period of the CL code is 1.5 seconds (one X1 epoch or Z-count).

The rate ½ constraint length 7 FEC scheme used to encode the 25-bps L2C navigation data into a 50-baud bit stream is shown in Figure 3.43.

The minimum specified received L2C signal power levels for signals broadcast from Block IIR-M and IIF satellites is –160 dBW and –158.5 dBW from GPS III satellites [29].

3.7.2.2 L5 Signal

The GPS L5 signal is generated as shown in Figure 3.44. Quadra-phase shift keying (QPSK) is used to combine an in-phase signal component (I5) and a quadra-phase signal component (Q5). Different PRN codes, each having length of 10,230 bits, are used for I5 and Q5. I5 is modulated by 50-bps navigation data that, after the addition of FEC using the same convolutional encoding as L2C, results in an overall symbol rate of 100 baud. A 10.23-MHz chipping rate is employed for both the I5 and Q5 PRN codes resulting in a 1-ms code repetition period.

Neuman-Hofman (NH) synchronization codes [94] are modulated upon I5 and Q5 at a 1-kHz rate. For I5, the 10-bit NH code 0000110101 is generated over a 10-ms interval and repeated. For Q5, the 20-bit NH code 00000100110101001110 is used. Every 1 ms, the current NH code bit is modulo-2 added to the PRN code chip. For example, on I5, the PRN code repeats 10 times over each 10-ms interval.

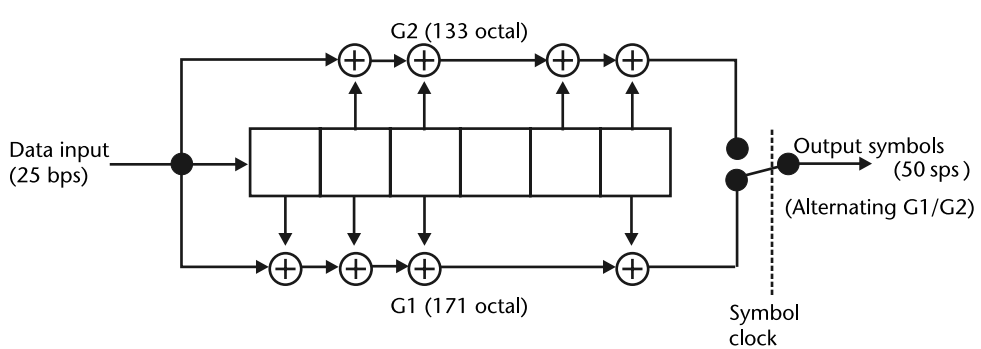


Figure 3.43 L2C data convolution encoder.

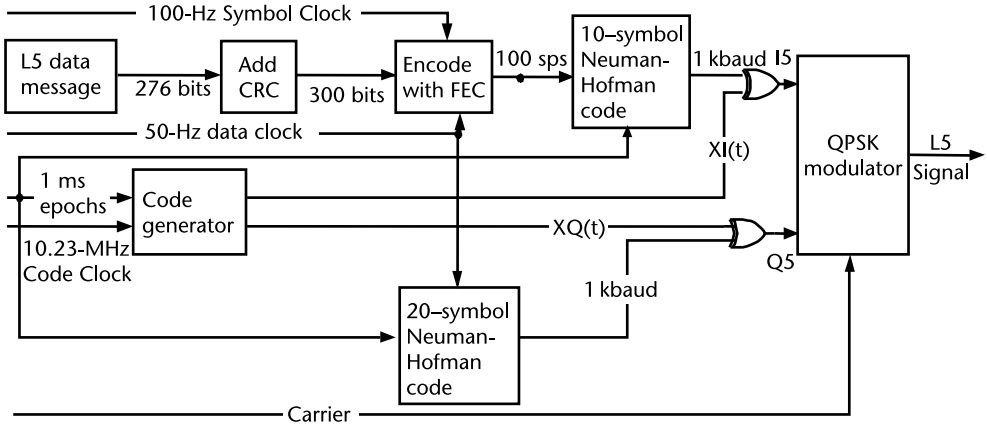


Figure 3.44 L5 signal generation.

During this interval, the PRN code is generated normally (upright) for repetitions 1 to 4, 7, and 9 (the zero bits in the I5 NH code 0000110101) and is inverted over repetitions 5, 6, 8, and 10 (corresponding to the set bits in the I5 NH code). The start of the I5 NH code is aligned with the start of each 10-ms data symbol that results from the FEC encoding. The Q5 NH code is synchronized with the 20-ms data bits.

The I5 and Q5 PRN codes are generated using the logic circuit shown in Figure 3.45, which is built around three 13-bit linear feedback shift registers. Every 1 ms, the XA coder is initialized to all 1s. Simultaneously, the XBI and XBQ coders are initialized to different values, specified in [91], to yield the I5 and Q5 PRN codes. The L5 minimum received power levels are shown in Table 3.17.

3.7.2.3 M Code Signal

The modernized military signal (M code) is designed exclusively for military use and is intended to become the primary signal for military use. During the transition period of replacing the GPS constellation with modernized SVs, the military user equipment has combined P(Y) code, M code and C/A code operation in the YMCA receiver. The primary military benefits that M code provides are improved security plus spectral isolation from the civil signals to reduce interference from higher power M code modes that enhance jamming resistance. Other benefits include enhanced tracking and data demodulation performance, robust acquisition, and com-

Table 3.17 L5 Minimum Received Signal Power Levels [91]

	Signal	
SV	*I5*	*Q5*
Block IIF (dBW)	–157.9	–157.9
GPS III (dBW)	–157.0	–157.0

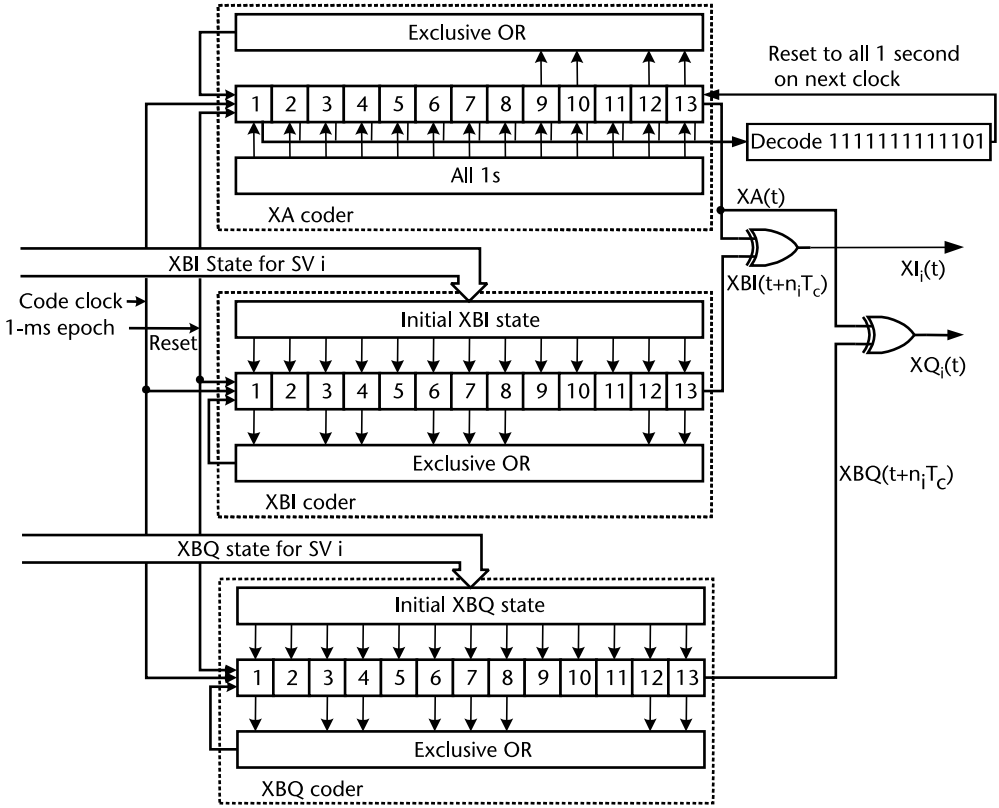


Figure 3.45 I5 and Q5 PRN code generation.

patibility with the C/A code and P(Y) code. It accomplishes these objectives within the existing GPS L1 (1,575.42 MHz) and L2 (1,227.60 MHz) frequency bands.

To accomplish the spectral separation shown in Figure 3.40, M code employs binary offset carrier (BOC) modulation. Specifically, M code uses a BOC(10,5) spreading modulation. The first parameter denotes the fundamental frequency of an underlying square wave subcarrier, which is 10×1.023 MHz, and the second parameter denotes the underlying M code generator code chipping rate, which is 5×1.023 mega chips per second (Mcps). Figure 3.46 depicts a high level block diagram of the M code generator. It illustrates the 10.23-MHz BOC square wave modulation of the underlying 5.115 Mcps M code generator that results in the split spectrum signals of Figure 3.40.

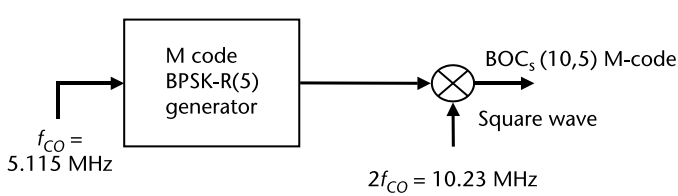


Figure 3.46 M code signal generation.

3.7.2.4 L1C Signal

The L1C signal, described in [32, 92, 95], has very different characteristics from those of other GPS signals. While it comprises pilot and data components like other modernized GPS signals, 75% of the signal power is allocated to the pilot component and only 25% to the data component, rather than the 50%/50% power allocation in other modernized GPS signals. Further, the two components are added in-phase using CDMA, rather than the time-division or phase quadrature division used by the other modernized GPS signals. Also, the pilot component and data component use different spreading modulations, with the spreading modulation for the pilot component selected to enhance tracking performance. Finally, the forward error encoding of the data messages uses a modern powerful encoding approach, low-density parity check (LDPC) encoding along with block interleaving, rather than the weaker convolutional encoding used on other modernized GPS signals.

Both components of the L1C signal are modulated onto the same L1 carrier as the C/A signal and L1P(Y) signal, in the same phase with each other and L1 P(Y), and in-phase quadrature with the C/A signal. There is no specified phase relationship between L1C and the L1M signal.

A BOC(1,1) spreading modulation is used for the data component. A time-multiplexed BOC (TMBOC) spreading modulation is used for the pilot component, with each of the 10,230 spreading symbols consisting of 310 repetitions of a specific pattern of 33 spreading symbols. Each of these 33 spreading symbols has four BOC(6,1) symbols in the first, fifth, seventh, and thirtieth locations, and BOC(1,1) symbols in the other 29 locations. Figure 3.47 [95] illustrates this configuration of spreading symbols in both components, including the relative amplitudes needed to provide the uneven division of power between the components.

The power spectral densities and autocorrelation functions of each L1C component, assuming ideal long spreading codes that contribute no additional structure, are shown in Figures 3.48 and 3.49. The normalized (unit-power) power spectral density of the data component, for ideal long spreading codes, is given by

$$\Phi_{\mathrm{L1C_D}}(f) = \frac{1}{1.023\times10^6}\operatorname{sinc}^2\left(\frac{\pi f}{1.023\times10^6}\right)\tan^2\left(\frac{\pi f}{2\times1.023\times10^6}\right)$$

while for the pilot component the corresponding power spectral density is

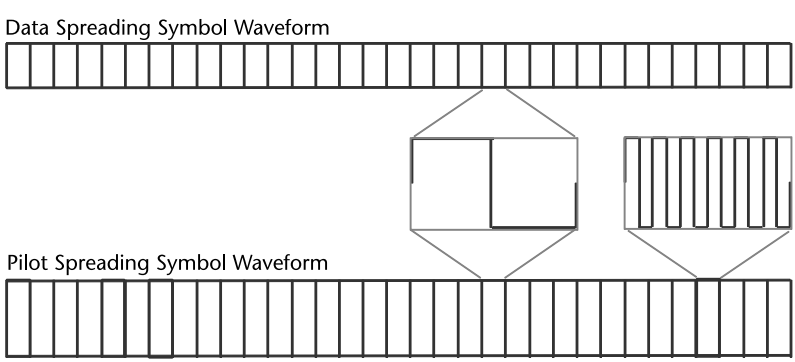


Figure 3.47 Segments of spreading waveforms for L1C components [95].

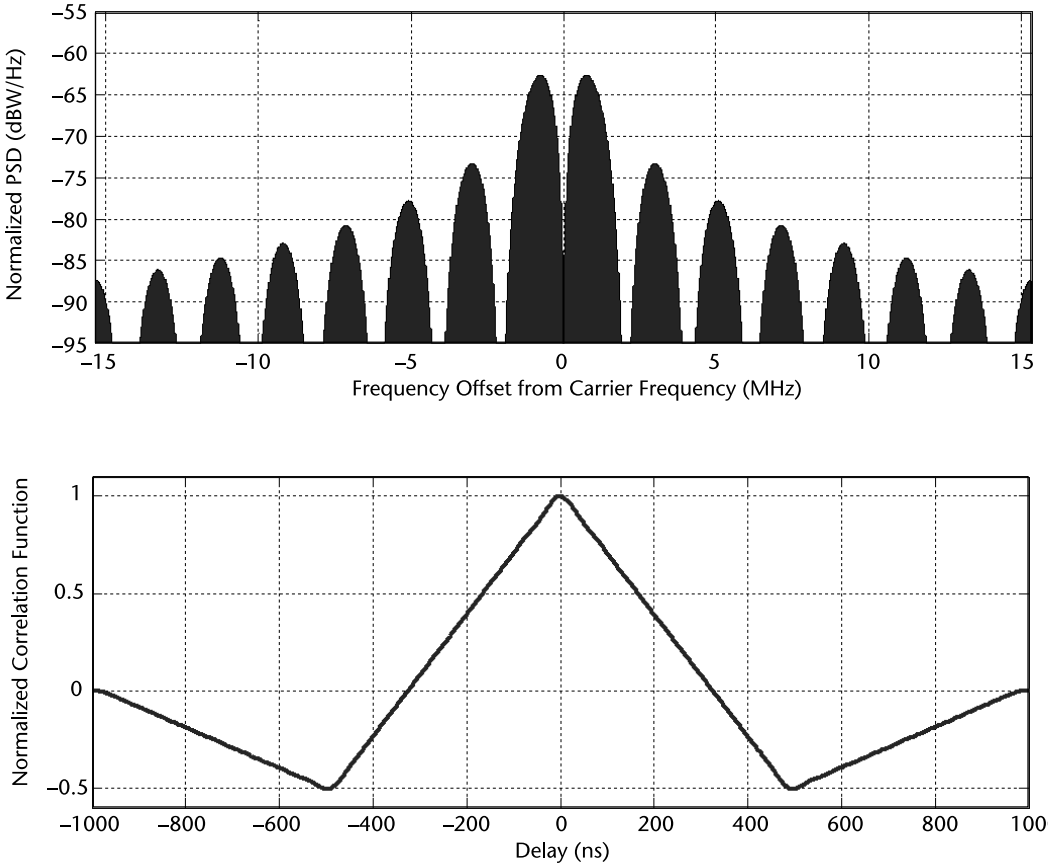


Figure 3.48 Normalized (unit-power) power spectral density and autocorrelation of L1C data component [95].

$$\Phi_{\mathrm{L1C_P}}(f) = \frac{1}{1.023\times10^6}\operatorname{sinc}^2\left(\frac{\pi f}{1.023\times10^6}\right)\left[\begin{array}{l}\dfrac{29}{33}\tan^2\left(\dfrac{\pi f}{2\times1.023\times10^6}\right)\\ +\dfrac{4}{33}\tan^2\left(\dfrac{\pi f}{12\times1.023\times10^6}\right)\end{array}\right]$$

The normalized power spectral density of the L1C signal, assuming ideal long spreading codes, is then

$$\begin{aligned}\Phi_{\mathrm{L1C}}(f) &= \frac{1}{4}\Phi_{\mathrm{L1C_D}}(f) + \frac{3}{4}\Phi_{\mathrm{L1C_P}}(f)\\ &= \frac{1}{1.023\times10^6}\operatorname{sinc}^2\left(\frac{\pi f}{1.023\times10^6}\right)\left[\begin{array}{l}\dfrac{10}{11}\tan^2\left(\dfrac{\pi f}{2\times1.023\times10^6}\right)\\ +\dfrac{1}{11}\tan^2\left(\dfrac{\pi f}{12\times1.023\times10^6}\right)\end{array}\right]\end{aligned}$$

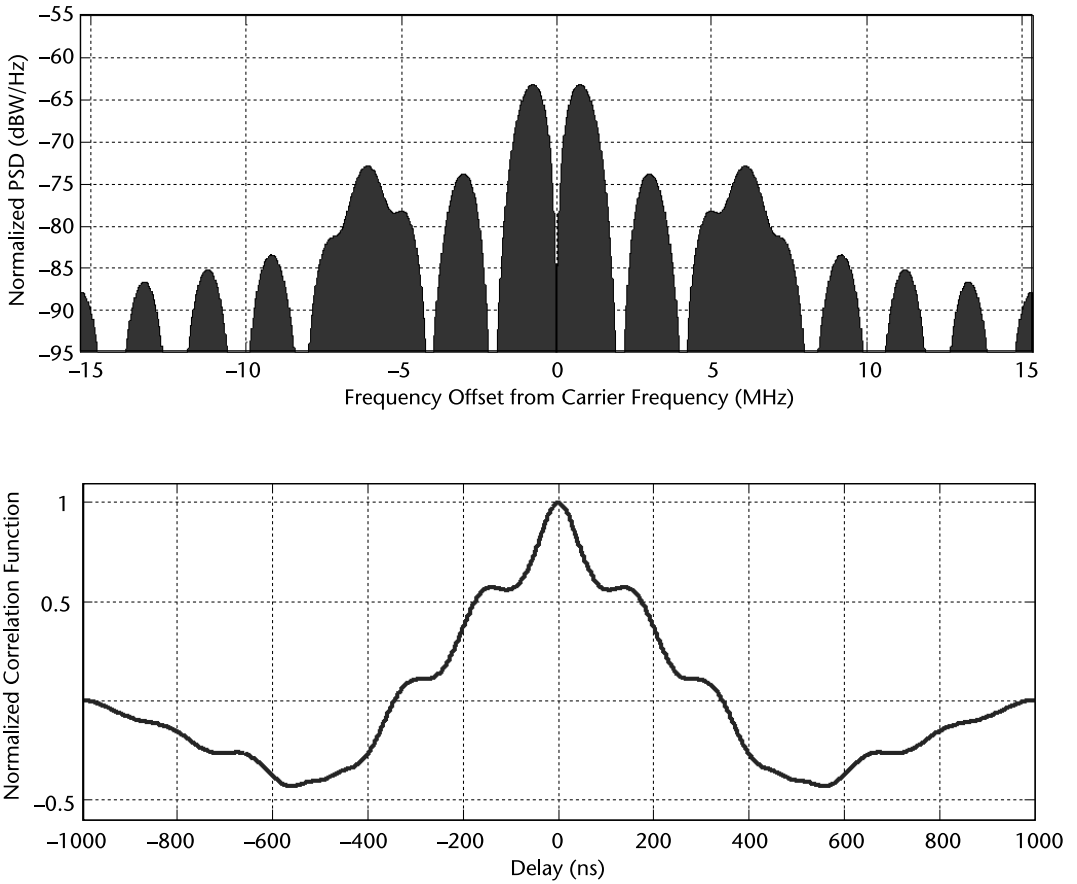
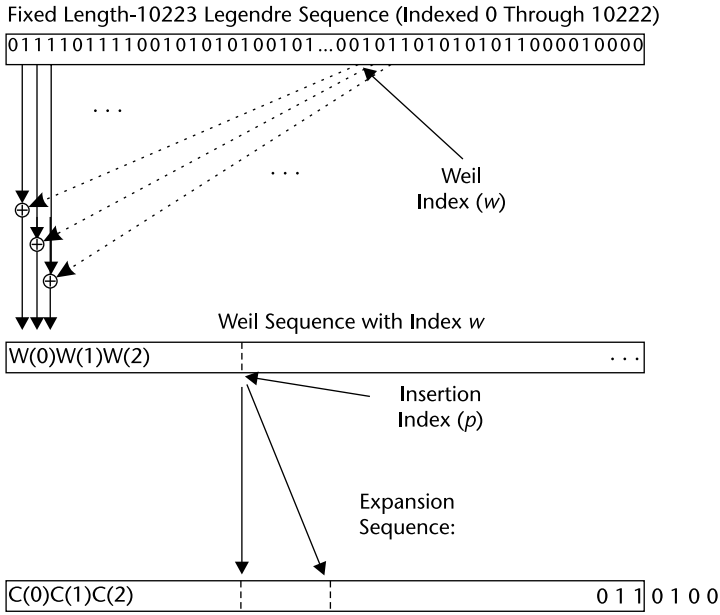


Figure 3.49 Normalized (unit-power) power spectral density and autocorrelation of L1C pilot component [95].

Spreading codes for the L1C components are generated using modified Weil sequences, as described in [95]. As described in [92], these Weil-based codes are generated from a 10,223-length Legendre sequence that can be generated using a simple algorithm, or stored permanently, as shown in Figure 3.50 [92, 95]. The 10,223-bit Weil sequences are constructed from the exclusive OR-ing of the Legendre sequence and a circularly shifted Legendre sequence. A 7-bit expansion sequence is then inserted to produce a 10,230-bit spreading code. Choosing the number of circular shifts in constructing the Weil sequence, along with the insertion point of the expansion, yields the different spreading codes.

The L1C data message, described in Section 3.7.3.3, is modulated onto the signal at 100 symbols/second, meaning that each symbol has 10-ms duration. The duration of each data message is 18 seconds, or 1,800 symbols. Since the spreading code duration is the same as the duration of a data message symbol, there is no need for an overlay code on the data component, unlike the L5 signal. However, as described in [92], an 1,800-bit-long overlay code is used on the pilot component at a 100-bps rate, making the duration of the overlay code 18 seconds. Consequently, when a receiver is aligned to the spreading code, it is also aligned to a data message symbol. Also, when a receiver is aligned to the overlay code, it is aligned to a data message.



Note: Weil Indices and Insertion Indices given in Table 3.2-2

Figure 3.50 Generation of L1C spreading codes [92, 95].

As described in [33], the L1C overlay codes are different for each signal. Families of spreading codes and overlay codes sufficient for 210 signals are defined in [92], enabling sharing from these families with other satnav systems if desired, since GPS is not expected to need more than 63 of these. The first 63 L1C overlay codes are segments of different m-sequences that are generated using an 11-stage shift register. The remaining overlay codes are Gold codes generated using a combination of two 11-stage shift registers. Figure 3.51 [95] shows these two registers. The coefficients m_k differ for each PRN, yielding a different polynomial, and are given in [92] along with the initial values for each register.

3.7.3 Civil Navigation (CNAV) and CNAV-2 Navigation Data

All of the modernized data messages for L2C, L5, and L1C differ from LNAV in several important ways:

- They are modulated onto a data component of the signal that is distinct from the pilot component primarily intended for tracking.
- They employ flexible data messages, where different message types containing different information can be transmitted in a variety of different sequences, rather than the fixed message structure employed in LNAV.
- They use forward error control that enables not only detection, but also correction, of some errors made in receiver processing to interpret the data message.

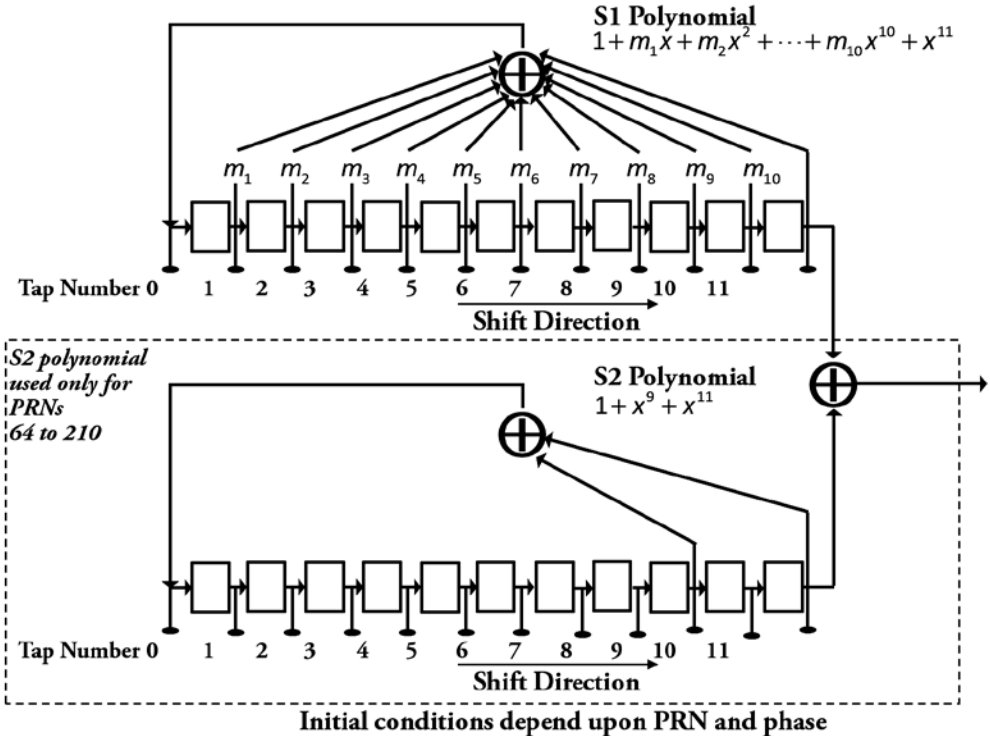


Figure 3.51 Shift registers used to generate L1C overlay codes [95].

- They use much more powerful techniques that allow receivers to detect, with very high probability, random errors in reading data message bits.
- They use higher precision representation of satellite ephemeris.

3.7.3.1 L2C CNAV Navigation Data

Section 3.3.3 and Appendix III of [29] describe the CNAV navigation data for L2C. The Block IIR-M/IIF, GPS III, and future SVs provide continuous L2C with CNAV navigation data on the L2 CM code, subject to control segment capability. Like the LNAV data message used on C/A and P(Y) signals, each satellite's L2C CNAV message provides the information necessary to compute the precise locations of that satellite, along with time of transmission for that L2C signal. The data also includes a significant set of auxiliary information that may be used [e.g., to assist the receiver in acquiring new satellites, to translate from GPS system time to UTC, and to correct for a number of errors that affect the range measurements]. This section outlines the main features of the L2C CNAV message. For a more complete description, the interested reader is referred to [29].

Each L2C CNAV navigation message consists of 300 bits, having duration of 12 seconds at the data rate of 25 bps. As shown in Figure 3.52 [29, 95], each message starts with an 8-bit preamble, followed by a 6-bit PRN number of the transmitting satellite, a 6-bit message type, a 17-bit message time of week (TOW) count, and a single-bit alert flag that indicating when the signal accuracy may be worse than

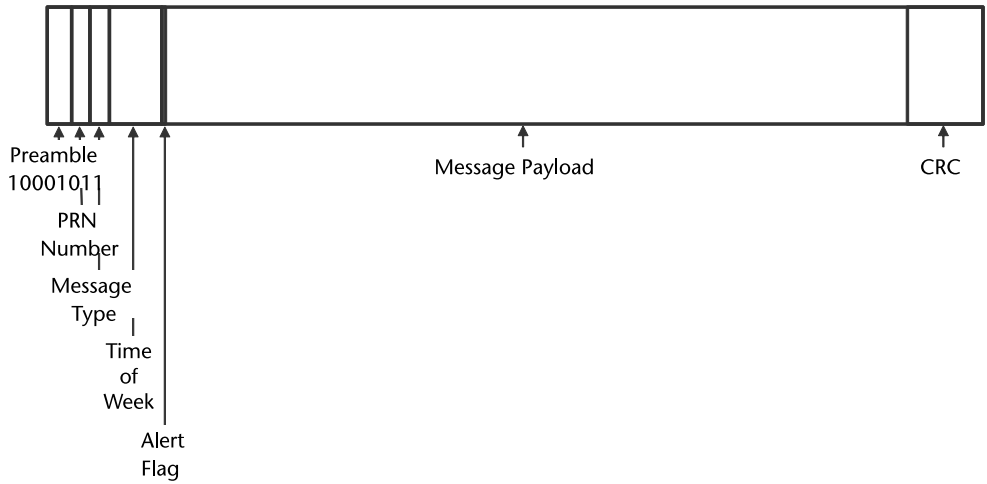


Figure 3.52 L2C CNAV data message structure [29, 95].

indicated in other messages. Following these first 38 bits of the message are 238 bits of message payload, followed by a 24-bit CRC covering the message contents. The entire message contents are encoded using half-rate constraint length-seven convolutional coding, producing 50 symbols per second. Messages are continuously encoded independent of message boundaries so that, at the beginning of each new message, the encode registers contain the last 6 bits of the previous message.

Different message types have different structures and contents of their payloads, as defined in [29]. Currently defined message types are listed in Table 3.18. The CS can direct that different message types be transmitted in different sequences, under certain constraints. Messages containing clock corrections and ephemeris are broadcast at least every 48 seconds. When the entire constellation is transmitting L2C with a fully functional control segment, the reduced almanac will be broadcast at least every 20 minutes, and the full almanac will be broadcast at least every 2 hours. Time offsets to other SATNAV systems will be broadcast every 288 seconds, or more often.

3.7.3.2 L5 CNAV Navigation Data

Section 3.3.3 and Appendix II of [91] describe the CNAV navigation data for L5. The Block IIF, GPS III and future SVs provide continuous L5 with CNAV navigation data on the I5 component of the L5 signal, with L5 Q5 a pilot component. The Block IIF, GPS III, and future SVs provide continuous L5 with CNAV navigation data, subject to control segment capability. The L5 data message is similar to that for the L2C signal in many ways. This section outlines the main features of the L5 CNAV message. For a more complete description, the interested reader is referred to [91].

Each L5 CNAV navigation message consists of 300 bits, having duration of 6 seconds at the data rate of 50 bps. The message structure is identical to that for L2C CNAV, shown in Figure 3.52. The entire message contents are encoded using half-rate constraint length-seven convolutional coding, producing 100 symbols per second. Like L2C CNAV, L5 CNAV messages are continuously encoded

Table 3.18 Currently Defined L2C Message Types [29, 95]

Message Type Number	*Message Contents*
0	Default
10	Ephemeris 1 and Health
11	Ephemeris 2 and Health
12	Reduced Almanac
13	Differential Correction Parameters
14	Differential Correction Parameters
15	Text Message
30	Clock Correction, Ionosphere Correction, Group Delay
31	Clock Correction, Reduced Almanac
32	Clock Correction, Earth Orientation Parameters
33	Coordinated Universal Time Parameters
34	Differential Correction Parameters
35	GPS-to-GNSS Time Offsets (GGTO)
36	Clock Corrections and Text Message
37	Almanac

independent of message boundaries so that, at the beginning of each new message, the encode registers contain the last 6 bits of the previous message.

Different message types have different structures and contents of their payloads, as defined in [91]. Currently defined message types are listed in Table 3.19. The CS can direct that different message types be transmitted in different sequences, under certain constraints. Messages containing clock corrections and ephemeris are broadcast at least every 24 seconds. When the entire constellation is transmitting L5 with a fully functional control segment, the reduced almanac will be broadcast at least every 10 minutes, and the full almanac will be broadcast at least every one hour. Time offsets to other SATNAV systems will be broadcast every 144 seconds or more often.

3.7.3.3 L1C CNAV-2 Navigation Data

Section 3.2.3.1 of [92] describes the L1C Message Structure and Section 3.5 of that same reference describes the CNAV-2 navigation data for L1C. The GPS III and future SVs will provide continuous L1C with CNAV-2 navigation data, subject to control segment capability. Like all of the previously described data messages, each satellite's L1C CNAV-2 message provides the information necessary to compute the precise locations of that satellite, along with time of transmission for that L1C signal. The data also includes a significant set of auxiliary information that may be used, for example, to assist the receiver in acquiring new satellites, to translate from GPS system time to UTC, and to correct for a number of errors that affect the range measurements. This section outlines the main features of the L1C CNAV-2 message. For a more complete description, the interested reader is referred to [92].

While it is a flexible data message like the other CNAV messages, the L1C CNAV-2 message structure is entirely different from that of other GPS signals. Each L1C CNAV navigation message consists of 900 bits, having duration of 18 seconds

Table 3.19 Currently Defined L5 Message Types

Message Type Number	*Message Contents*
0	Default
10	Ephemeris 1
11	Ephemeris 2
12	Reduced Almanac
13	Clock Differential Correction
14	Ephemeris Differential Correction
15	Text
30	Clock Correction, Ionosphere Correction, Group Delay
31	Clock Correction, Reduced Almanac
32	Clock Correction, Earth Orientation Parameters
33	Clock and Coordinated Universal Time Parameters
34	Clock and Differential Correction Parameters
35	Clock and GPS-to-GNSS Time Offsets (GGTO)
36	Clock Corrections and Text Message
37	Clock and Midi Almanac

at the data rate of 50 bps. Since the data message is aligned with the overlay code on the pilot component, no preamble is needed for the receiver to recognize the start of a message.

As shown in Figure 3.53 [95], each CNAV-2 message comprises three subframes, each separately encoded. Subframe 1 is the time of interval (TOI) denoting system time at the start (leading edge) of the message. These 9 bits are encoded with a Bose, Chaudhuri, and Hocquenghem (BCH) code into 54 symbols, TOI denotes the number of 18-second messages that have occurred since the beginning of the current interval. Subframe 2 contains the clock corrections and ephemeris, along with 8 bits representing the interval time of the week (ITOW), all protected with a CRC. ITOW denotes the number of 2-hour intervals since the beginning of the

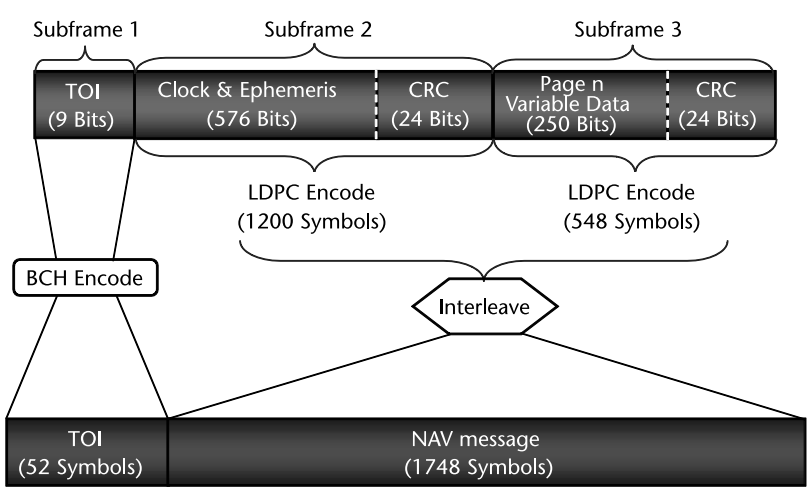


Figure 3.53 CNAV-2 data message structure [29, 95].

GPS week. The 600 bits of subframe 2 are encoded with a half-rate, low-density parity check (LDPC) code into 1,200 symbols. Subframe 3 in sequential messages provides all of the other data message information in 250-bit pages, protected with a CRC. The 274 bits of subframe 3 are encoded with a half-rate LDPC code into 574 symbols. The encoded symbols of subframes 2 and 3 are block interleaved before being modulated onto the data component of L1C.

Since subframe 2 bits are separately encoded, the symbols change only when the subframe 2 bits change, which is typically every two hours. When the bits have not changed, the receiver can use data symbol combining as explained in detail in [95]. Basically, if the receiver detects an uncorrected error after decoding subframe 2, it can coherently add the soft decisions from the next subframe 2, increasing the signal-to-noise ratio and thus the likelihood that the resulting combined symbols can be successfully decoded without uncorrected errors. However, subframe 3 data typically changes with each message so data symbol combining cannot be used for subframe 3 data.

Since TOI is the same from all SVs, it only needs to be read from one signal. The receiver can select the signal having highest received power for this purpose, or can use data symbol combining across messages for improved performance.

Different pages in subframe 3 have different structures and contents of their payloads, as defined in [92]. Currently defined message types are listed in Table 3.20. The control segment can direct that different pages be transmitted in different sequence.

3.8 GPS Ephemeris Parameters and Satellite Position Computation

We close the chapter with a discussion of the GPS ephemeris parameters and the computation of an SV's position in ECEF coordinates. GPS almanac data and ephemeris data transmitted by the satellites include the Keplerian orbital elements, described in Section 2.3.1, plus additional parameters. Also, note that in the GPS ephemeris data, the time of perigee passage is converted to mean anomaly at epoch by (2.11). The orbital elements include a reference time, known as the time of epoch or time of ephemeris, at which the orbital elements were valid. Only at epoch are the orbital elements exactly as described by the given values, known as osculat-

Table 3.20 Currently Defined L1C Pages [95]

Page Number	*Message Contents*
1	UTC and Ionospheric Corrections
2	GPS/GNSS Time Offset and Earth Orientation Parameters
3	Reduced Almanac Parameters
4	Almanac Parameters
5	Differential Correction Parameters
6	Text Message
7	Signal Phase for Each SV

ing orbital elements. At all later times, the true orbital elements deviate from the osculating values.

Because it is necessary for the GPS ephemeris message to contain very accurate information about the satellite's position and velocity, it is insufficient to use only the osculating Keplerian orbital elements for computing the position of a GPS satellite, except very near the epoch of those elements. One solution to this problem would be to update the GPS ephemeris messages very frequently. Another solution would be for the GPS receiver to integrate the fully perturbed equation of motion, (2.7), which would include a detailed force model, from epoch to the desired time. Because these solutions are complex and computationally intensive, they are impractical for real-time operations. Therefore, the osculating Keplerian orbital elements in the GPS ephemeris message are augmented by correction parameters that allow the user to estimate the Keplerian elements fairly accurately during the periods of time between updates of the satellite's ephemeris message. (Particulars on ephemeris message updating are provided in Section 3.3.1.4.) Any time after the epoch of a particular ephemeris message, the GPS receiver uses the correction parameters to estimate the true orbital elements at the desired time.

In this section, we present the ephemeris data transmitted by the GPS satellites, and we show how the ephemeris data are used to compute the satellite position in ECEF coordinates. We show this for the legacy GPS ephemeris message in Section 3.8.1, and for the civil navigation ephemeris message in Section 3.8.2.

3.8.1 Legacy Ephemeris Parameters

Table 3.21 summarizes the parameters contained in the GPS Legacy ephemeris message. These parameters are found in Table 20-III of IS-GPS-200 [29]. As can be seen, the first seven parameters of the GPS ephemeris message are time of epoch and, essentially, the osculating Keplerian orbital elements at the time of epoch, with the exceptions that the semimajor axis is reported as its square root and that mean

Table 3.21 Legacy GPS Ephemeris Data Definitions [29]

t_{0e}	Reference time of ephemeris
$\sqrt{a}$	Square root of semimajor axis
e	Eccentricity
i_0	Inclination angle (at time t_{0e})
Ω_0	Longitude of the ascending node (at weekly epoch)
ω	Argument of perigee (at time t_{0e})
M_0	Mean anomaly (at time t_{0e})
di/dt	Rate of change of inclination angle
$\dot{\Omega}$	Rate of change of longitude of the ascending node
Δ_n	Mean motion correction
C_{uc}	Amplitude of cosine correction to argument of latitude
C_{us}	Amplitude of sine correction to argument of latitude
C_{rc}	Amplitude of cosine correction to orbital radius
C_{rs}	Amplitude of sine correction to orbital radius
C_{ic}	Amplitude of cosine correction to inclination angle
C_{is}	Amplitude of sine correction to inclination angle

anomaly is used instead of time of perigee passage. The next nine parameters allow for corrections to the Keplerian elements as functions of time after epoch. (The osculating elements and associated particulars are described in detail in Section 2.3.1.

Table 3.22 provides the algorithm by which a GPS receiver computes the position vector of a GPS satellite (x_s, y_s, z_s) in the ECEF coordinate system from the parameters in Table 3.21 using the ephemeris parameters from the legacy navigation message. The computation produces the ECEF coordinates of the antenna phase center of the satellite. For computation (3) in Table 3.22, t represents the GPS system time at which the GPS signal was transmitted. In the notation of Table 3.22, the subscript k appearing in computation (3) and below means that the subscripted variable is measured at time t_k, the time (in seconds) from epoch to the GPS system time of signal transmission.

There are a few subtleties worth noting in the computations described in Table 3.22. First, computation (5), which is Kepler's equation, (2.9), is transcendental in the desired parameter, E_k. Therefore, the solution must be carried out numerically. Kepler's equation is readily solved either by iteration or Newton's method.

Table 3.22 Computation of a Satellite's ECEF Position Vector Using Legacy GPS Navigation Message Data [29]

(1)	$a = \left(\sqrt{a}\right)^2$	Semimajor axis
(2)	$n = \sqrt{\dfrac{\mu}{a^3}} + \Delta n$	Corrected mean motion, $\mu = 398{,}600.5 \times 10^8$ m^3/s^2
(3)	$t_k = t - t_{0e}$	Time from ephemeris epoch
(4)	$M_k = M_0 + n(t_k)$	Mean anomaly
(5)	$M_k = E_k - e\sin E_k$	Eccentric anomaly (must be solved iteratively for E_k)
(6)	$\sin\nu_k = \dfrac{\sqrt{1-e^2}\sin E_k}{1-e\cos E_k}$ $\cos\nu_k = \dfrac{\cos E_k - e}{1-e\cos E_k}$	True anomaly
(7)	$\phi_k = \nu_k + \omega$	Argument of latitude
(8)	$\delta\phi_k = C_{us}\sin(2\phi_k) + C_{uc}\cos(2\phi_k)$	Argument of latitude correction
(9)	$\delta r_k = C_{rs}\sin(2\phi_k) + C_{rc}\cos(2\phi_k)$	Radius correction
(10)	$\delta i_k = C_{is}\sin(2\phi_k) + C_{ic}\cos(2\phi_k)$	Inclination correction
(11)	$u_k = \phi_k + \delta\phi_k$	Corrected argument of latitude
(12)	$r_k = a(1 - e\cos E_k) + \delta r_k$	Corrected radius
(13)	$i_k = i_0 + (di/dt)t_k + \delta i_k$	Corrected inclination
(14)	$\Omega_k = \Omega_0 + (\dot{\Omega} - \dot{\Omega}_e)(t_k) - \dot{\Omega}_e t_{0e}$	Corrected longitude of the ascending node
(15)	$x_p = r_k\cos u_k$	In-plane x position
(16)	$y_p = r_k\sin u_k$	In-plane y position
(17)	$x_s = x_p\cos\Omega_k - y_p\cos i_k\sin\Omega_k$	ECEF x-coordinate
(18)	$y_s = x_p\sin\Omega_k + y_p\cos i_k\cos\Omega_k$	ECEF y-coordinate
(19)	$z_s = y_p\sin i_k$	ECEF z-coordinate

A second subtlety is that computation (6) must produce the true anomaly in the correct quadrant. Therefore, it is necessary either to use both the sine and the cosine or to use a "smart" arcsine function. Also, to carry out computation (14), it is necessary to know the rotation rate of the Earth. According to IS-GPS-200 [29], this rotation rate is $\dot{\Omega}_e = 7.2921151467 \times 10^{-5}$ rad/s, which is consistent with the WGS 84 value to be used for navigation, although WGS 84 also provides a slightly different value in defining the ellipsoid. Finally, IS-GPS-200 [29] defines the value of π to be used by GPS user equipment as exactly 3.1415926535898.

As can be seen from the computations in Table 3.22, the variations in time of the orbital parameters are modeled differently for particular parameters. For example, mean motion is given a constant correction in computation (2), which effectively corrects the mean anomaly computed in (4). However, argument of latitude, radius, and inclination are corrected by truncated harmonic series in computations (8), (9), and (10), respectively. Eccentricity is given no correction. Finally, longitude of the ascending node is corrected linearly in time in computation (14). It is a misnomer of GPS system terminology, as in Table 3.21, that the longitude of the ascending node, Ω_0, is given at a weekly epoch. In reality, Ω_0 is given at the reference time of ephemeris, t_{0e}, the same as the other GPS parameters. This can be verified by inspection of computation (14) from Table 3.22. Reference [96] provides an excellent description of the trade-offs that resulted in the use of ephemeris message parameters and computations described in Tables 3.21 and 3.22.

3.8.2 CNAV and CNAV-2 Ephemeris Parameters

We conclude with a discussion on computation of an SV's position in ECEF coordinates from CNAV and CNAV-2 ephemeris message data contained in CNAV Message Types 10 and 11 and CNAV-2 subframe 2. The CNAV and CNAV-2 ephemeris

Table 3.23 CNAV/CNAV-2 Ephemeris Parameters

ΔA	Semimajor axis difference at reference time
$\dot{A}$	Change rate in semimajor axis
Δn_0	Mean motion difference from computed value at reference time
$\Delta \dot{n}_0$	Rate of mean motion difference from computed value
M_0	Mean anomaly at reference time
e_{n}	Eccentricity
ω_{n}	Argument of perigee
t_{0e}	Ephemeris data reference time of week
Ω_{0n}	Longitude of ascending node of orbit plane at weekly epoch
$\Delta\dot{\Omega}$	Rate of right ascension difference
$i_{0\mathrm{n}}$	Inclination angle at reference time
$i_{0\mathrm{n\text{-}DOT}}$	Rate of inclination angle
$C_{\mathrm{is\text{-}n}}$	Amplitude of sine correction to inclination angle
$C_{\mathrm{ic\text{-}n}}$	Amplitude of cosine correction to inclination angle
$C_{\mathrm{rs\text{-}n}}$	Amplitude of sine correction to orbital radius
$C_{\mathrm{rc\text{-}n}}$	Amplitude of cosine correction to orbital radius
$C_{\mathrm{us\text{-}n}}$	Amplitude of sine correction to argument of latitude
$C_{\mathrm{uc\text{-}n}}$	Amplitude of cosine correction to argument of latitude

Table 3.24 Computation of a Satellite's ECEF Position Vector Using CNAV/CNAV-2 Navigation Message Data

	Element/Equation	*Description*
(1)	$\mu = 3.986005 \times 10^{14}$ m^3/s^2	WGS 84 value of the Earth's gravitational constant
(2)	$\Omega_e = 7.2921151467 \times 10^{-5}$ rad/s	WGS 84 value of the Earth's rotation rate
(3)	$A_0 = A_{\text{REF}} + \Delta A$	Semimajor axis at reference time, A_{REF} = 26,559,710m
(4)	$t_k = t - t_{0e}$	Time from ephemeris reference time
(5)	$A_k = A_0 + \left(\dot{A}\right)t_k$	Semimajor axis
(6)	$n_0 = \sqrt{\dfrac{\mu}{A_0^3}}$	Computed mean motion (rad/s)
(7)	$\Delta n_a = \Delta n_0 + \dfrac{1}{2}\Delta \dot{n}_0 t_k$	Mean motion difference from computed value
(8)	$n_a = n_0 + \Delta n_a$	Corrected mean motion
(9)	$M_k = M_0 + n_a\, t_k$	Mean anomaly
(10)	$M_k = E_k - e_{\text{n}}\, E_k$	Kepler's equation for eccentric anomaly (E_k), rad; solve by iteration for E_k, same as for legacy
(11)	$\sin \nu_k = \dfrac{\sqrt{1-e_{\text{n}}^2}\sin E_k}{1 - e_{\text{n}}\cos E_k}$ $\cos \nu_k = \dfrac{\cos E_k - e_{\text{n}}}{1 - e_{\text{n}}\cos E_k}$ $\nu_k = \tan^{-1}\dfrac{\sin \nu_k}{\cos \nu_k}$	True anomaly calculation; first compute $\sin\nu_k$ and $\cos\nu_k$. Then compute ν_k with the smart, or 4-quadrant, arctan function (atan2 in many computer languages).
(12)	$\Phi_k = \nu_k + \omega_{\text{n}}$	Argument of latitude
(13)	$\delta u_k = C_{\text{us-n}}\sin(2\Phi_k) + C_{\text{uc-n}}\cos(2\Phi_k)$	Argument of latitude correction
(14)	$\delta r_k = C_{\text{rs-n}}\sin(2\Phi_k) + C_{\text{rc-n}}\cos(2\Phi_k)$	Radial correction
(15)	$\delta i_k = C_{\text{is-n}}\sin(2\Phi_k) + C_{\text{ic-n}}\cos(2\Phi_k)$	Inclination correction
(16)	$u_k = \Phi_k + \delta u_k$	Corrected argument of latitude
(17)	$r_k = A_k\,(1 - e_{\text{n}}\cos E_k) + \delta r_k$	Corrected radius
(18)	$i_k = i_{0\text{n}} + (i_{0\text{n-DOT}})t_k + \delta i_k$	Corrected inclination
(19)	$x_k' = r_k \cos u_k$	x-position in orbital plane
(20)	$y_k' = r_k \sin u_k$	y-position in orbital plane
(21)	$\dot{\Omega} = \dot{\Omega}_{\text{REF}} + \Delta\dot{\Omega}$	Rate of right ascension; $\dot{\Omega}_{\text{REF}} = -2.6 \times 10^{-9}$ semicircles/s
(22)	$\Omega_k = \Omega_{0\text{n}} + \left(\dot{\Omega} - \dot{\Omega}_e\right)t_k - \dot{\Omega}_e t_{0e}$	Corrected longitude of the ascending node
(23)	$x_k = x_k' \cos\Omega_k - y_k' \cos i_k \sin\Omega_k$ $y_k = x_k' \sin\Omega_k + y_k' \cos i_k \cos\Omega_k$ $z_k = y_k' \sin i_k$	ECEF coordinates of space vehicle antenna phase center

parameters follow the same principles as the legacy GPS ephemeris parameters: there are Keplerian elements augmented by correction parameters. However, there are two main differences between the CNAV/CNAV-2 and legacy ephemeris parameters: (1) CNAV/CNAV-2 ephemeris messages have additional parameters, and (2) some of the CNAV/CNAV-2 parameters are expressed as differences from specified reference values, as opposed to absolute values. The additional parameters added to

CNAV/CNAV-2 are rate of change of the following: semimajor axis and mean motion. The parameters expressed as differences instead of absolute values in CNAV/CNAV-2 are the following: semimajor axis and rate of longitude of the ascending node. Finally, note that with CNAV/CNAV-2, semimajor axis is used, instead of square root of semimajor axis as in the legacy ephemeris message.

Table 3.23 summarizes the ephemeris parameters contained in the CNAV/CNAV-2 ephemeris messages. These parameters are found in Table 30-I of IS-GPS-200H [29] for CNAV and Table 3.5-1 in IS-GPS-800D [92] for CNAV-2. Table 3.23 summarizes only the parameters required to compute the position of a GPS SV. The CNAV and CNAV-2 ephemeris messages contain additional parameters pertaining to signal health and user range error (URA) elevation-dependent accuracy. For CNAV-2, there are additional parameters beyond those provided with CNAV, specifically for clock corrections to improve PNT accuracy. In addition, CNAV-2 has more parameters on URA elevation dependent accuracy and inter-signal corrections. Nonetheless, the basic ephemeris parameters and method for computing space vehicle position are the same for CNAV and CNAV-2.

Table 3.24 provides the algorithm by which a GPS receiver computes the position vector of a satellite antenna phase center (x_k, y_k, z_k) in the ECEF coordinate system from the parameters in Table 3.23. As with the legacy ephemeris computations, the value of π to be used is 3.1415926535898, and the WGS-84 rotation rate of the earth is $\dot{\Omega}_e = 7.2921151467 \times 10^{-5}$ rad/s. In computation (4) in Table 3.24, t represents the GPS system time at which the GPS signal was transmitted. In the notation of Table 3.24, the subscript k appearing in computation (4) and below means that the subscripted variable is measured at time t_k, the time (in seconds) from epoch to the GPS system time of signal transmission.

References

[1] U.S. government information about the Global Positioning System (GPS) and related topics, www.gps.gov.

[2] Bates, R., et al., *Fundamentals of Astrodynamics*, New York: Dover Publications, 1971.

[3] U.S. Department of Defense, *Global Positioning System Standard Positioning Service Performance Standard*, September 2008.

[4] http://www.defense.gov/Contracts/, September 21, 2016.

[5] U.S. Government, Department of Defense, *Global Positioning System Precise Positioning Service Performance Standard*, February 2007.

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

[7] http://www.navcen.uscg.gov.

[8] United States Air Force Report to Congressional Committees, *Global Positioning System Constellation Replenishment,* February 2015.

[9] "GPS Almanac," *GPS World Magazine*, August 2016, http://gpsworld.com/the-almanac/.

[10] Global Positioning Systems Directorate "GPS Status & Modernization Progress: Service, Satellites, Control Segment, and Military GPS User Equipment," *Presentation to the National Space-Based PNT Advisory Board*, May 18–19, 2016.

[11] Marquis, W., "Increased Navigation Performance from GPS Block IIR," *NAVIGATION: Journal of the Institute of Navigation*, Vol. 50, No. 4, Winter 2003-2004.

[12] Marquis, W., and D. Riggs, "Impact of GPS Block IIR Space Vehicle Lifetime on Constellation Sustainment," *ION-GNSS-2010*, September 2010.

[13] Marquis, W., "Recent Developments in GPS Performance and Operations," AAS 11-014, February 2011.

[14] Riley, W. J., "Rubidium Atomic Frequency Standards for GPS Block IIR," *ION-GPS-92*, Albuquerque, NM, September 1992.

[15] Taylor, J., and E. Barnes, "GPS Current Signal-in-Space Navigation Performance," *ION 2005 Annual Technical Meeting*, San Diego, CA, January 24–26, 2005.

[16] Marquis, W., and C. Krier, "Examination of the GPS Block IIR Solar Pressure Model," *ION-GPS-2000*, Salt Lake City, UT, September 2000.

[17] Swift, E. R., *GPS REPORTS: Radiation Pressure Scale and Y-axis Acceleration Estimates for 1998-1999*, Naval Surface Warfare Center, report #3900 T10/006, March 9, 2000.

[18] Marquis, W., and D. Reigh, "The GPS Block IIR and IIR-M Broadcast L-Band Antenna Panel -- Its Pattern and Performance," *Navigation – The Journal of the Institute of Navigation*, December 2015.

[19] Hartman, T., et al., "Modernizing the GPS Block IIR Spacecraft," *ION-GPS-2000*, Salt Lake City, UT, September 2000.

[20] Marquis, W., "M Is for Modernization: Block IIR-M Satellites Improve on a Classic," *GPS World Magazine*, Vol. 12, No. 9, September 2001, pp. 38–44.

[21] Madden, D. Col. USAF, "GPS Program Update to 49th CGSIC Meeting," *ION-GNSS-2009*, September 2009.

[22] Barbour, B., "GPS Constellation Health and Modernization,"*CGSIC*, October 2009.

[23] "U.S. Air Force/Lockheed Martin Team Complete GPS III Design Phase Ahead of Schedule," Lockheed Martin Press Release, August 20, 2010.

[24] Marquis, W., and M. Shaw, "Design of the GPS III Space Vehicle," *ION-GNSS-2011*, September 2011.

[25] Marquis, W., and M. Shaw, "GPS III -- Bringing New Capabilities to the Global Community," *Inside GNSS*, September 2011.

[26] Kaplan, E., and C. Hegarty, (eds.), *Understanding GPS: Principles and Applications*, 2nd ed., Norwood, MA: Artech House, 2006, Section 3.2.3.6.

[27] Dass, T., et al., "Analysis of On Orbit Behavior of GPS Block IIR Time Keeping System," *30th Annual Precise Time and Time Interval (PTTI) Meeting*, December 1998.

[28] Marquis, W., and D. Reigh, "On-Orbit Performance of the Improved GPS Block IIR Antenna Panel," *ION-GNSS-2005*, September 2005.

[29] GPS Directorate, IS-GPS-200H, NAVSTAR, *GPS Space Segment/Navigation User Interfaces*, United States Air Force GPS Directorate, El Segundo, CA, September 24, 2013, www.gps.gov.

[30] Shaw, S., and A. Katronick, "GPS III Signal Integrity Improvements," *ION-GNSS-2013*, September 2013.

[31] Affens D., et al., "The Distress Alerting Satellite System," *GPS World*, January 2011.

[32] Betz, J. W., et al., "Description of the L1C Signal," *Proceedings of the Institute of Navigation Conference on Global Navigation Satellite Systems 2006, ION-GNSS-2006*, Institute of Navigation, September 2006.

[33] Rushanan, J. J., "The Spreading and Overlay Codes for the L1C Signal," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 54, No. 1, Spring 2007, pp. 43–51.

[34] "Lockheed Martin Team Completes Design Milestone for GPS III Program," Lockheed Martin Press Release, July 5, 2011.

[35] "National Space Policy of the United States of America," Office of the President of the United States, June 2010.

[36] Parkinson, B., et al., *Global Positioning System: Theory and Applications*, Vol. I, Washington, D.C.: American Institute of Aeronautics and Astronautics, 1996.

[37] Creel, T., et al., "Accuracy and Monitoring Improvements from the GPS Legacy Accuracy Improvement Initiative," *ION-NTM-2006*, January 2006.

[38] Creel, T., et al., "New, Improved GPS -- The Legacy Accuracy Improvement Initiative," *GPS World*, March 2006.

[39] Haerr, D., L. Harmon, and A. Bokelman, "Transitioning the GPS OCS to a Modern Architecture," *NAVIGATION: Journal of the Institute of Navigation*, Vol. 44, No. 2, 1997, Summer 1997.

[40] Brown, K., et al., "L-Band Anomaly Detection in GPS," *Proc. of the 51st Annual Meeting, Inst. of Navigation*, Washington, D.C., 1995.

[41] Wiley, B., et al., "NGA's Role in GPS," *Proceedings of the 19th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2006)*, Fort Worth, TX, September 2006, pp. 2111–2119.

[42] Mendicki, P., Private Communication, Aerospace Corporation, March 2016.

[43] Hay, C., and J. Wong, "Improved Tropospheric Delay Estimation at the Master Control Station," *GPS World*, July 2000, pp. 56–62.

[44] "GPS OCS Mathematical Algorithms, Volume GOMA-S," DOC-MATH-650, Operational Control System of the NAVSTAR Global Positioning System, December 2011.

[45] Cappelleri, J., C. Velez, and A. Fucha, *Mathematical Theory of the Goddard Trajectory Determination System*, Goddard Space Flight Center, April 1976.

[46] Springer, T., "Modeling and Validating Orbits and Clocks Using the Global Positioning System," Ph.D. Dissertation, Astronomical Institute, University of Bern, November 1999.

[47] Maybeck, P. S., *Stochastic Models, Estimation and Control, Vol. 1*, New York: Academic Press, 1979.

[48] Bierman, G. J., *Factorization Methods for Discrete Sequential Estimation*, Orlando, FL: Academic Press, 1977.

[49] "GPS OCS Mathematical Algorithms, Volume GOMA-E," DOC-MATH-650, Operational Control System of the NAVSTAR Global Positioning System, December 2011.

[50] "Global Positioning System Clock Analysis Quarterly Report, 2016-1," U.S. Naval Research Laboratory, Washington, D.C., February 10, 2016.

[51] Van Dierendonck, A., and R. Brown, "Relationship Between Allan Variances and Kalman Filter Parameters," *Proc. of 16th Annual PTTI Meeting*, Greenbelt, MD, 1984.

[52] Saastamoinen, J., "Contributions to the Theory of Atmospheric Refraction," *Bulletin Géodésique*, 1973, No. 105, pp. 270–298; No. 106, pp. 383–397; No. 107, pp. 13–34.

[53] Niell, A., "Global Mapping Functions for the Atmosphere Delay at Radio Wavelengths," *Journal of Geophysical Research*, Vol. 101, No. B2, 1996, pp. 3227–3246.

[54] Brown, K., "The Theory of the GPS Composite Clock," *Proc. of ION GPS-91*, Institute of Navigation, Washington, D.C., 1991.

[55] Senior, K., et al., "Developing an IGS Time Scale," *IEEE Trans. on Ferroelectronics and Frequency Control*, June 2003, pp. 585–593.

[56] Mendicki, P., Private Communication, Aerospace Corporation, August 2016.

[57] Taylor, J., et al., "GPS Control Segment Upgrade Goes Operational - Enhanced Phased Operations Transition Details," *ION-NTM-2008*, January 2008.

[58] Taylor, J., et al., "AEP Goes Operational: GPS Control Segment Upgrade Details," *GPS World*, June 2008.

[59] Weiss, M. A., and A. Masarie, "GPS Changes Before and After Implementation of the Architecture Evolution Plan," NIST, October 2008.

[60] Weiss, M., "GPS Changes Before and After Implementation of the Architecture Evolution Plan," *PTTI*, December 2008.

[61] Malys, L., et al., "The GPS Accuracy Improvement Initiative," *ION-GPS-1997*, September 1997.

[62] Hay, C., "The GPS Accuracy Improvement Initiative," *GPS World*, June 2000.

[63] Yinger, C., et al., "GPS Accuracy Versus Number of NIMA Stations," *Proc. of ION GPS 03*, Institute of Navigation, Washington, D.C., 2003.

[64] McCarthy, D., (ed.), *IERS Technical Note, 21*, U.S. Naval Observatory, July 1996.

[65] Marquis, Krier, "Examination of the GPS Block IIR Solar Pressure Model," *ION-GPS-2000*, 2000.

[66] Bar-Sever, Y., and D. Kuang, "New Empirically Derived Solar Radiation Pressure Model for GPS Satellites," *JPL Interplanetary Network Progress Report*, Vol. 24-159, November 2004; addendum: "New Empirically Derived Solar Radiation Pressure Model for Global Positioning System Satellites During Eclipse Seasons," *JPL Interplanetary Network Progress Report*, Vol. 42-160, February 2005.

[67] Hopfield, H., "Tropospheric Effects on Electromagnetically Measured Range, Prediction from Surface Weather Data," *Radio Science,* Vol. 6, No. 3, March 1971, pp. 356–367.

[68] Black, H., "An Easily Implemented Algorithm for the Tropospheric Range Correction," *Journal of Geophysical Research*, Vol. 83, April 1978, pp. 1825–1828.

[69] Dieter, G. L., G. E. Hatten, and J. Taylor, "MCS Zero Age of Data Measurement Techniques," *35th PTTI*, 2003.

[70] Creel, T., et al., "Summary of Accuracy Improvements from the Legacy Accuracy Improvement Initiative (L-AII)," *ION-GNSS-2007*, 2007.

[71] Brown, K., et al., "Dynamic Uploading for GPS Accuracy," *ION-GPS-1997*, 1997.

[72] Pullen, E., F. Shaw, and S. Frye, "GPS III Accuracy and Integrity Improvements Using ARAIM with Shorter Age of Data," *ION-GNSS-2014*, 2014.

[73] GPS World Staff, "Lockheed Martin Advances Threat Protection on GPS Control Segment," *GPS World Magazine*, Vol. 26, No. 12, December 2015.

[74] Petit, G., and B. Luzum,, *IERS Technical Note 36*, 2010.

[75] Host, P.,"Air Force Awards Lockheed Martin for GPS III Temporary Contingency Operations," *Defense Daily Network*, February 2016.

[76] *GNSS Market Report, Issue 4* copyright © European GNSS Agency, 2015

[77] http://www2.st.com/content/st_com/en/about/innovation---technology/BiCMOS.html.

[78] "2015 GPS World Receiver Survey," *GPS World Magazine*, January 2015.

[79] National Geospatial-Intelligence Agency, *Department of Defense World Geodetic System 1984 (WGS 84): Its Definition and Relationships with Local Geodetic Systems*, NGA. STND.0036_1.0.0_WGS84, Version 1.0.0, National Geospatial-Intelligence Agency, Office of Geomatics, July 8, 2014.

[80] Schwarz, C. R., "Relation of NAD 83 to WGS 84," in *North American Datum of 1983*, Ed. C. R. Schwarz, NOAA Professional Paper NOS 2, National Geodetic Survey, NOAA, Silver Spring, MD, December 1989, pp. 249–252.

[81] Powers, E., *GPS Timing Services,* United States Naval Observatory, October 22, 2015.

[82] *Global Positioning System (GPS) Standard Positioning Service (SPS) Performance Analysis Report,* William J. Hughes Technical Center, WAAS Test and Evaluation Team, http://www.nstb.tc.faa.gov/.

[83] Mendicki, P. J., "GPS Signal in Space Weekly Status," Aerospace Corporation, Navigation Division/GPS Operations, August 3, 2016.

[84] Gold, R., "Optimal Binary Sequences for Spread Spectrum Multiplexing," *IEEE Transactions on Information Technology*, Vol. 33, No. 3, October 1967.

[85] Forssell, B., *Radionavigation Systems*, Upper Saddle River, NJ: Prentice Hall International, 1991, pp. 250–271.

[86] Czopek, F. M., "Description and Performance of the GPS Block I and II L-Band Antenna and Link Budget," *Proc. 6th International Technical Meeting of The Satellite Division of The Institute of Navigation*, Salt Lake City, UT, September 22–24, 1993, Vol. I, pp. 37–43.

[87] Shank, C., B. Brottlund, and C. Harris, "Navigation Message Correction Tables: On Orbit Results," *Proc. of the Institute of Navigation Annual Meeting*, Colorado Springs, CO, June 1995.

[88] Fontana, R. D., W. Cheung, and T. Stansell, "The New L2 Civil Signal," *GPS World*, September 2001.

[89] Van Dierendonck, A. J., and C. Hegarty, "The New Civil GPS L5 Signal," *GPS World*, September 2000.

[90] IS-GPS-705D, *NAVSTAR GPS Space Segment/Navigation User Interfaces*, September 24, 2013, GPS.gov.

[91] IS-GPS-800D, *NAVSTAR GPS Space Segment/Navigation User Interfaces*, September 24, 2013, GPS.gov.

[92] Barker, B. C., et al., "Overview of the GPS M Code Signal," *Proceedings of Institute of Navigation National Technical Meeting* 2000, ION-NTM-2000, Institute of Navigation, January 2000.

[93] Spilker, J. J., Jr., *Digital Communications by Satellite*, Upper Saddle River, NJ: Prentice Hall, 1977.

[94] Betz, J. W., *Engineering Satellite-Based Navigation and Timing: Global Navigation Satellite Systems, Signals, and Receivers*, New York: Wiley-IEEE Press, 2016.

[95] Van Dierendonck, A. J., et al., "The GPS Navigation Message," *GPS Papers Published in Navigation*, Vol. I, Washington, DC: Institute of Navigation.

CHAPTER 4

GLONASS

Nadejda Stoyanova, Scott Feairheller, and Brian Terrill

4.1 Introduction

The Global Navigation Satellite System (GLONASS) is the Russian Federation counterpart to the U.S. GPS. [GLONASS as a program is capitalized, whereas, when addressing the actual space vehicles, only the first letter of the satellite name is capitalized (e.g., Glonass, Glonass-M, Glonass-K).] GLONASS provides military and civil multifrequency L-band navigation services for positioning, navigation, and timing solutions for maritime, air, land, and space applications both inside Russia and internationally.

The history of the GLONASS program is similar to GPS. Like GPS, the (then) Soviet military initiated the program in the mid-1970s to support military requirements. The first GLONASS satellite launched on October 12, 1982. An initial test constellation of four SVs was deployed by January 1984. Originally, GLONASS was funded to support naval demands for navigation and time dissemination. Early system testing convincingly demonstrated that GLONASS could also support civilian use while concurrently meeting Soviet defense needs. Thus, the mission was broadened to include civilian users [1].

At a meeting of the Special Committee on Future Air Navigation Systems (FANS) of the International Civil Aviation Organization (ICAO) in 1988, the USSR offered the world community free use of GLONASS navigation signals for air safety. A similar offer was made at the 35th Session of the International Maritime Organization (IMO) Subcommittee of Navigation Safety in the same year [1, 2].

After the collapse of the Soviet Union in 1991, the Russians established a test constellation of 10 to 12 satellites. Extensive testing of the system followed. As a result, in September 1993, Russian President Boris Yeltsin officially proclaimed GLONASS to be an operational system, part of the Russian Armory, and the basis for the Russian Radio-navigation Plan [3].

Between April 1994 and December 1995, Russia conducted seven more launches, thus completing a 24-satellite constellation. In February 1996, these satellites were declared operational and the constellation was fully populated for the first time. However, a number of older satellites soon thereafter failed, and

the constellation quickly degraded. From 1996 through 2001, the Russians only launched two sets of three satellites, which eventually left the constellation with only 6 to 8 working spacecraft. It was not until 2011 that Russia was able to restore its constellation back to full global service.

During the buildup, the Government of Russia issued Decree 237 on March 7, 1995, that opened the GLONASS C/A-code signals for civil use and guaranteed they would be available free of charge, affirming the Soviet 1988 statement. Russia also published and made publicly available an Interface Control Document (ICD), which detailed the structure of the open service GLONASS signals and navigation message. The latest version of the ICD was published in 2008 [4, 5].

Later, on February 18, 1999, the Russian President issued decree 38-RP, which declared GLONASS a dual-use system. This was followed by a decree on March 29, 1999, opening GLONASS up for international cooperation [6, 7].

In August 2001, the Russians created the first federal targeted program for GLONASS for the years 2001-2011, stabilizing the program, securing funding, and developing the associated infrastructure. The program was further reinforced on May 17, 2007, by the President of the Russian Federation, Vladimir Putin, when issued Decree 638, declaring the GLONASS open service available to all national and international users without any limitations [8]. Today maintenance and modernization of GLONASS is financed through a federal targeted program, "Maintenance, Development, and Use of GLONASS 2012–2020," which covers upgrades to the space, ground, and user segments, as well as transportation and geodetic applications [9].

By 2016, there had been 49 successful launches (plus 4 launch vehicle failures) in the program, placing in orbit a total of 86 Glonass satellites, 45 Glonass-M satellites, 2 Glonass-K1 satellites, and 2 Etalon passive geodetic satellites. Details of the present constellation and each of these spacecraft types are provided next.

4.2 Space Segment

4.2.1 Constellation

The GLONASS constellation nominally consists of 24 active satellites plus 6 on-orbit spares. (As of 2017, this number of on-orbit spares had not yet been achieved.) They are positioned in a 19,100-km orbit with a 64.8° inclination, and a period of revolution of 11 hours and 15 minutes. The 24 satellites are uniformly located in three orbital planes, 120° apart in right ascension. Each plane contains eight satellites, equally spaced with 45° displacement in argument of latitude, and a 15° argument of latitude difference between satellites in the same slot in two different planes. The ground track repeat cycle for GLONASS is 8 days (Figure 4.1) [5]. The current orbital configuration and overall system design (including satellite nominal L-band antenna beamwidths of 35° to 40°) provide navigation service to users up to 2,000 km above the Earth's surface [1].

Each GLONASS satellite is assigned an orbital slot number from 1 to 24, which is relative to the satellite's position within the constellation (see Figure 4.2). A 24-satellite constellation provides continuous 4-satellite visibility from more than 99% of the Earth's surface. Under the 24-satellite concept, the performance of all

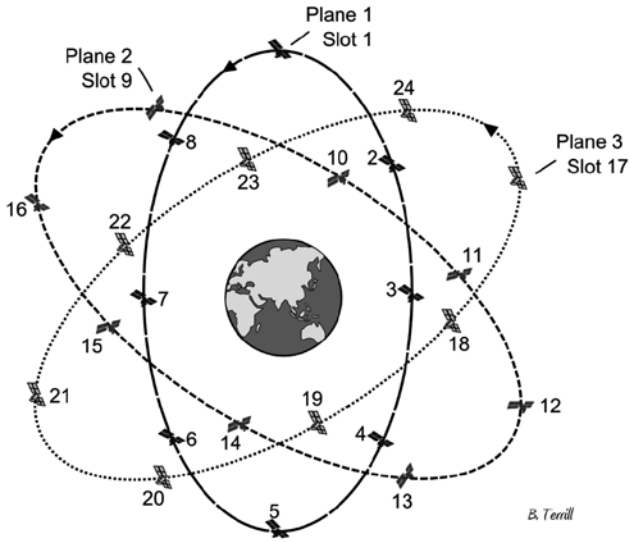
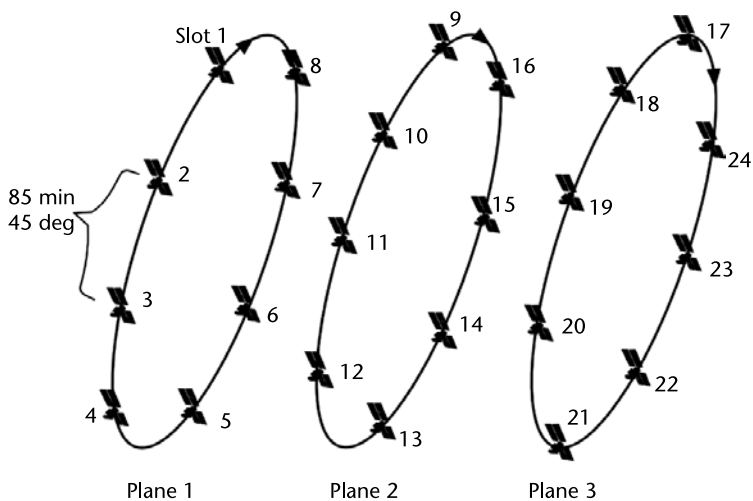


Figure 4.1 GLONASS constellation structure.



Slot 17 (Plane 3) reaches maximum norther latitude 28 min before Slot 9 (Plane 2) and 56 min before Slot 1 (Plane 1)

Figure 4.2 GLONASS constellation orbital arrangement.

in-orbit satellites (nominally 30) will be determined by GLONASS controllers and the best 24 will be activated. The remaining satellites (nominally 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 [1, 3, 10, 11].

The constellation also includes two Etalon passive geodetic satellites in a slightly elliptical medium earth orbit. They were launched on January 10, 1989, and May 31, 1989, each along with a pair of Glonass satellites. Each Etalon SV is a 1.294-m diameter, 1,415-kg sphere covered with a retro-reflector array. The mission of the satellites was to establish a highly accurate terrestrial reference frame.

Table 4.1 GLONASS Constellation as of August 2016

Orbital Slot	*Plane*	*Satellite Name*	*Launch Date*
1	1	Glonass-M/Kosmos-2456	December 14, 2009
2	1	Glonass-M/Kosmos-2485	April 26, 2013
3	1	Glonass-M/Kosmos-2476	November 4, 2011
4	1	Glonass-M/Kosmos-2474	October 2, 2011
5	1	Glonass-M/Kosmos-2458	December 14, 2009
6	1	Glonass-M/Kosmos-2457	December 14, 2009
7	1	Glonass-M/Kosmos-2477	November 4, 2011
8	1	Glonass-M/Kosmos-2475	November 4, 2011
9	2	Glonass-K1/Kosmos-2501	December 1, 2014
10	2	Glonass-M/Kosmos-2426	December 25, 2006
11	2	Glonass-M/Kosmos-2516	May 29, 2016
12	2	Glonass-M/Kosmos-2436	December 25, 2007
13	2	Glonass-M/Kosmos-2434	December 25, 2007
14	2	Glonass-M/Kosmos-2424	December 25, 2006
15	2	Glonass-M/Kosmos-2425	December 25, 2006
16	2	Glonass-M/Kosmos-2466	September 2, 2010
17	3	Glonass-M/Kosmos-2514	February 7, 2016
18	3	Glonass-M/Kosmos-2494	March 24, 2014
19	3	Glonass-M/Kosmos-2433	October 26, 2007
20	3	Glonass-M/Kosmos-2432	October 26, 2007
20	3	Glonass-K1/Kosmos-2471	February 26, 2011
21	3	Glonass-M/Kosmos-2500	June 14, 2014
22	3	Glonass-M/Kosmos-2459	March 2, 2010
23	3	Glonass-M/Kosmos-2466	March 2, 2010
24	3	Glonass-M/Kosmos-2461	March 2, 2010

Today, the Etalons are used by Russia, as well as the international space community to calibrate ground laser ranging equipment.

4.2.2 Spacecraft

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-K spacecraft design, first launched in 2011. Russia plans to introduce the next generation of spacecraft, Glonass-K2, starting in 2018.

4.2.2.1 Glonass Spacecraft

From 1982 through 2005, Russia launched Glonass series satellites (see Figure 4.3). These satellites were a traditional Russian design consisting of a pressurized, hermetically sealed cylinder that is three-axis stabilized (i.e., oriented in all three axes of motion, usually measured as in-track, cross-track, and radial from the satellite's point of view). Circulation of gas inside the pressurized vessel allows for cooling of the satellite electronics. Attached on the bottom of the spacecraft was the payload

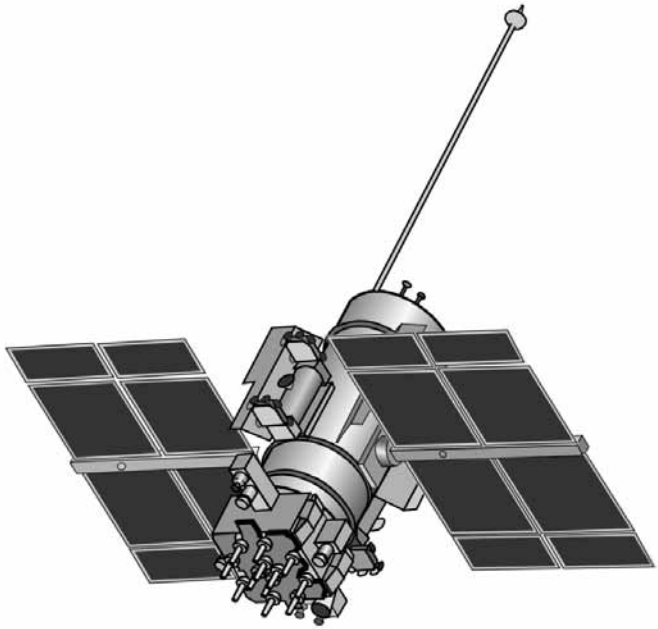


Figure 4.3 Glonass spacecraft.

assembly. This assembly consists of the horizon sensor, laser retro-reflectors, a 12-element navigation signal antenna, and various command and control antennas. Attached to the sides of the pressurized cylinder are the solar panels, orbital correction engines, a portion of the attitude control system, and the thermal control louvers [1]. The original Glonass I satellite series carried two rubidium clocks with stability of 5×10^{-12} (at 1 day), whereas Glonass II spacecraft switched to three Cesium AFSs, thus bumping up the AFS stability to 5×10^{-13} (at 1 day) [12]. These Glonass satellites transmitted an L1 FDMA signal. (For signal descriptions, see Section 4.7.)

4.2.2.2 Glonass-M Spacecraft

Beginning in 2003, Russia began launching Glonass-M spacecraft (see Figure 4.4), where M stands for modified. The Glonass-M is a modernized version of the Glonass spacecraft using upgraded electronics and supporting a number of new features. The spacecraft carries three more-accurate cesium AFSs (1×10^{-13} at 1 day), a better attitude control system, and intersatellite navigation links (incorporated after the second Glonass-M satellite). These features reduced errors in measurements of time and ephemeris calculation. Glonass-M also carries increased propellant, improved the onboard batteries, and modernized spacecraft electronics, which increased the satellite design lifetime to 7 years. An improved navigation message transmits corrections between GPS and GLONASS time to facilitate joint use and navigation data authentication information every 4 seconds and navigation age-of-data information. Glonass-M adds a second civil modulation on L2 signal. Since 2014, newly launched Glonass-M spacecraft have transmitted an additional open CDMA service signal in L3. (For signal descriptions, see Section 4.7.)

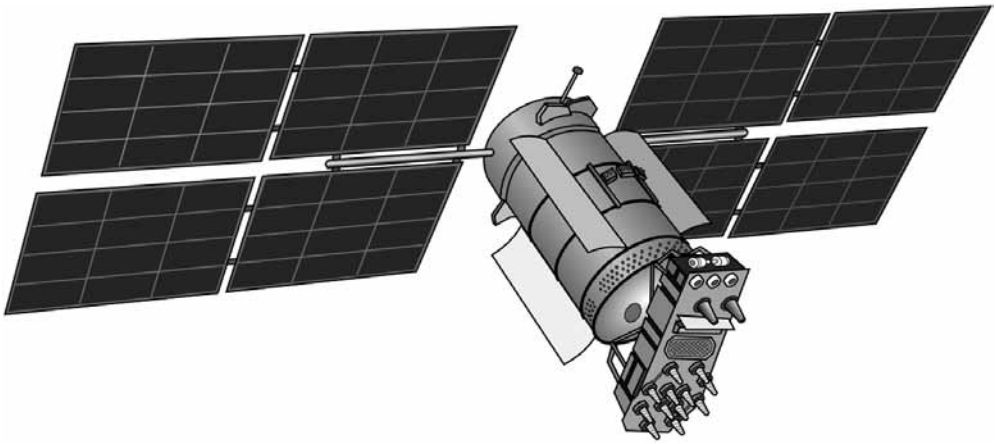


Figure 4.4 Glonass-M spacecraft.

Like the original Glonass SVs, the Glonass-M spacecraft consists of a pressurized, hermetically sealed cylinder that is three-axis stabilized. In contrast, the solar panels are attached to the top of the cylinder and the payload assembly (attached on the bottom of the spacecraft), and are much larger in one dimension. The spacecraft mass is approximately 1,415 kg. This assembly consists of the horizon sensor, laser retro-reflectors, a 12-element navigation signal antenna, a cross-link antenna, and various command and control antennas. The longer assembly allows the navigation payload and laser retro-reflector arrays to be mounted separately. Attached to the sides of the pressurized cylinder are the orbital correction engines, a portion of the attitude control system, and the thermal control louvers [1, 13–19].

As of the beginning of 2017, Russia had 7 Glonass-M satellites left for launch before completely retiring the series [9].

4.2.2.3 Glonass-K1 Spacecraft

Beginning in 2011, Russia began testing a new generation spacecraft, which represents a departure from legacy Soviet systems. The Glonass-K1 satellite (see Figure 4.5) uses an Express-1000K unpressurized bus. The new bus offers several new features: light honeycomb panel structure, heat pipe thermal control, radiation-hardened electronics, and 17 m^2 GaAs solar panels. Glonass-K1 satellites weigh only 935 kg and currently launch on the Soyuz-2 space launch vehicle out of Plesetsk.

Just like its predecessors, Glonass-K1 carries a 12-element navigation signal antenna, laser retroreflectors, and an RF satellite-to-satellite crosslink. The first two Glonass-K1 satellites transmit the legacy FDMA signals for backwards compatibility, and the CDMA L3 open service signal already introduced on the latest Glonass-M satellites. (For signal descriptions, see Section 4.7.)

The Glonass-K1 satellites carry two cesium and two rubidium AFSs, which give the satellites AFS stability on the order of 0.5 to 5×10^{-13} (at 1 day), with that number expected to increase to 1×10^{-14} (at 1 day), starting with the third spacecraft in this series [13–20].

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

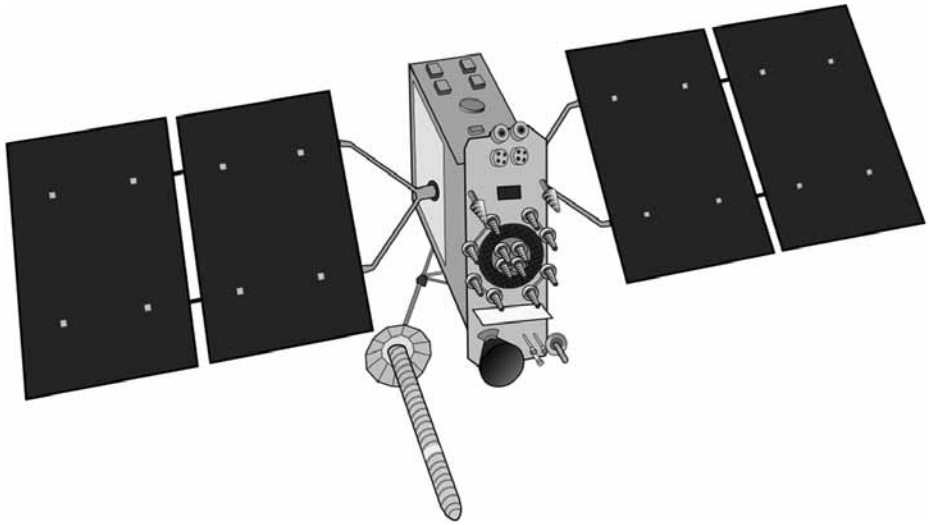


Figure 4.5 Glonass-K1 spacecraft.

the currently deployed COSPAS-SARSAT system. This payload is similar in design and concept to the payload on the European Galileo satellite navigation system [13–19]. Glonass-K also carries a payload for confirmation of nuclear detonation (NUDET) and treaty verification [21].

Russia plans on launching nine more Glonass-K1 satellites before switching to the next generation [22].

4.2.2.4 Glonass-K2 Spacecraft

Starting in 2018, Russia will begin launching Glonass-K2 satellites (see Figure 4.6). The new spacecraft will be based on a modified Express-1000A bus, thus using triple junction GaAs solar cells and a lithium-ion battery. The satellite has a 10-year design life. Its estimated weight is around 1,645 kg. While a launch vehicle for this satellite has not been specified yet, it is likely that a single Glonass-K2 launch will utilize a Soyuz-2 SLV, whereas a Proton-M SLV with a Briz-M upper stage will be able to deliver a pair of satellites to orbit. Glonass-K2 is expected to carry additional payloads, such as COSPAS-SARSAT and NUDET support, which were already introduced on the Glonass-K1 SV. The Glonass-K2 SVs are also expected to carry two cesium and two rubidium AFSs, which give the satellites AFS stability on the order of $0.5–1\times10^{-14}$ (at 1 day).

Glonass-K2 will continue to carry the legacy FDMA signals for backwards compatibility. In addition to the L3 CDMA signal introduced on the latest Glonass-M and Glonass-K1 satellites, the Glonass-K2 will also transmit CDMA signals in L1 and L2 [23] (see Section 4.7.7.9).

4.2.2.5 Glonass-KM Spacecraft

While the satellite design is unknown at this point, this satellite generation will likely add the L5 frequency as a standard part of its payload. It will still transmit the legacy FDMA signals, in addition to the CDMA signals on L1, L2, and L3.

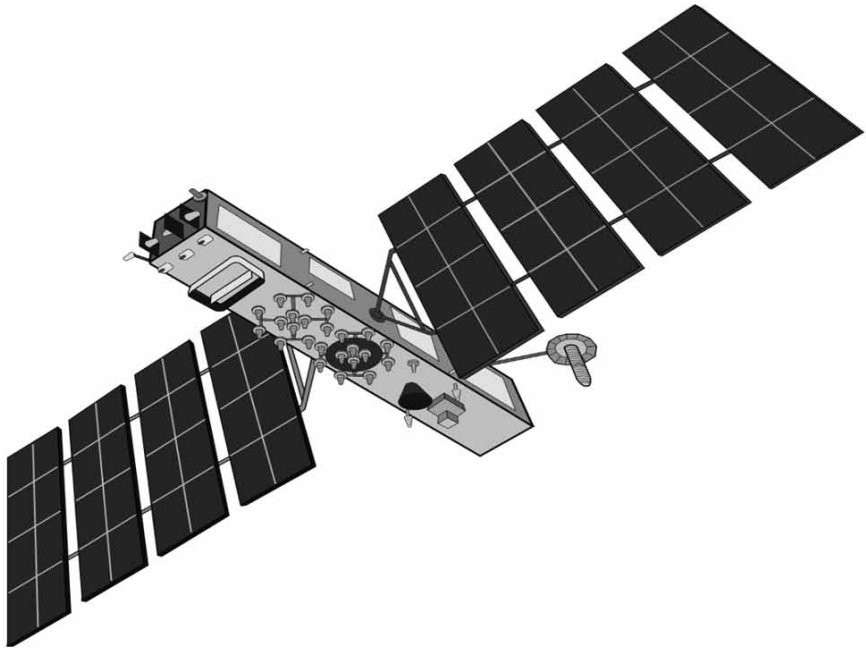


Figure 4.6 Glonass-K2 spacecraft.

4.3 Ground Segment

GLONASS is supported by a network of ground sites mainly located within the borders of Russia and augmented by monitor sites located throughout the rest of the world (see Figure 4.7). The ground-based control complex (GBCC) is responsible for the following functions:

- Measurement and prediction of individual satellite ephemeris;
- Uploading of predicted ephemeris, clock corrections, and almanac information into each GLONASS satellite for later incorporation into the navigation message;
- Synchronization of the satellite clocks with GLONASS system time;
- Calculation of the offset between GLONASS system time and UTC(SU);
- Spacecraft command, control, housekeeping, and tracking [1].

4.3.1 System Control Center (SCC)

The SCC, formerly known as Golitsino-2, a military complex run by the Russian Space Forces, is located in Krasnoznamensk, about 40-km southwest of Moscow. The SCC schedules and coordinates all functions for GLONASS [1].

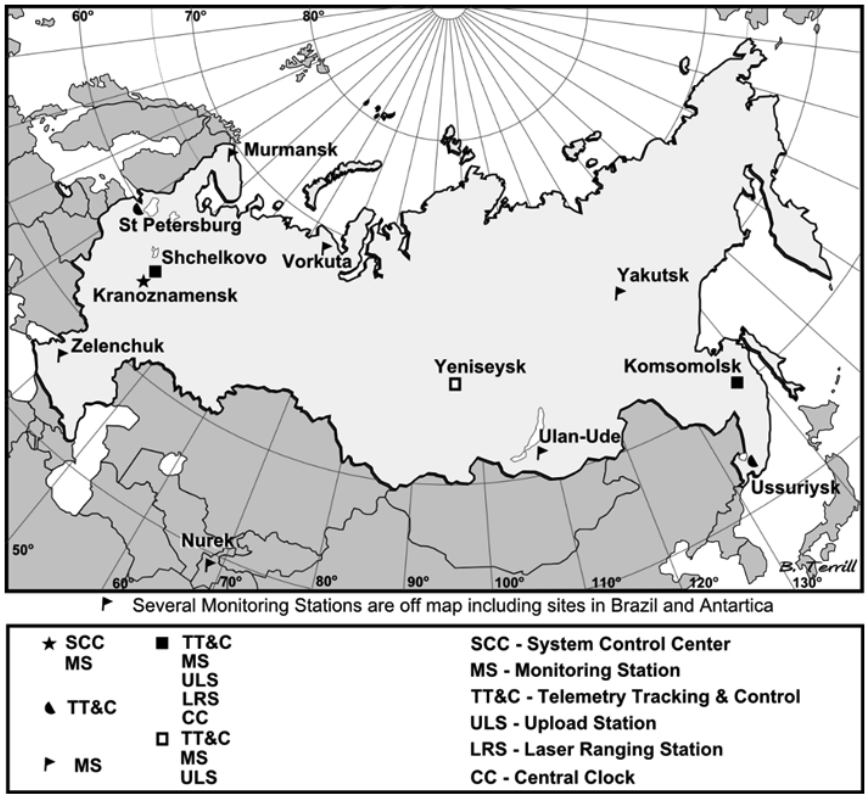


Figure 4.7 GLONASS ground segment.

4.3.2 Central Synchronizer (CS)

The CS, or the system clock, is located near Schelkovo, about 20-km northeast of Moscow, and forms the GLONASS system time. GLONASS system time is also synchronized to the Universal Time Coordinated of Russia, UTC (SU), which is maintained by the National Metrology Institute of the Russian Federation (VNIIFTRI) in Mendeleevo near Moscow. Signals from the central synchronizer are relayed to the phase control system (PCS), which monitors satellite clock time/phase as transmitted by the navigation signals. The PCS performs two types of measurements in order to determine the satellite time/phase offsets. The PCS directly measures the range to the satellites by use of radar techniques. The PCS also simultaneously compares the satellite transmitted navigation signals to a reference time/phase generated by a highly stable frequency standard (relative error approximately 10^{-13}) at the ground site. These two measurements are then differenced to determine the satellite clock time/phase offsets. Measurements from the PCS are used to predict the satellite clock time/phase corrections, which are uploaded by the ground station into the satellite. This comparison of each satellite's time/phases errors is carried out at least on a daily basis [1, 24].

4.3.3 Telemetry, Tracking, and Command (TT&C)

TT&C stations measure individual satellite trajectories and uplink required control and payload information to the satellite's onboard processor. Tracking involves between three and five measurement sessions, each lasting 10 to 15 minutes. Range to the satellite is measured by radar techniques with a maximum error of between 2m and 3m. These radio-frequency ranges are periodically calibrated using a laser ranging device at the laser tracking stations. Each satellite carries laser retro-reflectors specifically for this purpose. Ephemeris is predicted 24 hours in advance and uploaded once per day. The spacecraft clock correction parameters are renewed twice a day. Any interruption in the normal operation of the ground segment interrupts the accuracy of GLONASS signals. Tests have shown that a spacecraft clock can maintain acceptable accuracy for no more than two to three days of autonomous operations. Although the satellite's central processor is capable of 30 days of autonomous operations, this variability in the time standard is the limiting component for autonomous GLONASS operations [1].

4.3.4 Laser Ranging Stations (SLR)

SLR stations calibrate radio-frequency tracking measurements and provide optical measurement for orbit determination for GLONASS [1, 3]. SLR stations are generally colocated with monitoring stations. The laser ranging network is also supported by an experimental multifunctional optical and laser complex located near Kitab in southern Uzbekistan on Mt. Maidanak. Cameras located on Mt. Maidanak are capable of measuring ranges to an object up to an altitude of 40,000 km and down to a visible stellar magnitude of 16. The maximum error of satellite angular coordinate determination does not exceed 1 to 2 arc-seconds under normal operating conditions and 0.5 arc-second under special experimental conditions. The maximum ranging error is not more than 1.5 to 1.8 cm, and the error of the fix to the UTC (SU) scale is not more than $\pm 1\ \mu$s. GLONASS measurements are relayed via secure radio link to the system control center once per hour. Mt. Maidanak provides unique climatic characteristics with more than 220 clear days annually, thus making it a reliable source of correction data to the system control center [1, 3].

4.4 GLONASS User Equipment

GLONASS is designed to support a wide variety of civil, commercial, and military PNT applications in Russia, and throughout the rest of the world. Note that time is user time scale to the National Reference of Coordinated Universal Time UTC (SU) [5].

The first Russian automotive navigation receiver to utilize GLONASS along with GPS was the Glospace-SGK70, released on December 27, 2007. At the time of this writing, GLONASS had been incorporated into Russian and Western GPS-GLONASS or GNSS chipsets and incorporated in many consumer items such as phones since the 2011 timeframe. The first smart phone with GLONASS (and GPS) was the ZTE MTS 945 powered by Qualcomm's Snapdragon MSM7x30 chipset [25]. The proposed civil applications include: terrestrial, air (aviation) and marine

navigation, disaster management, vehicle tracking and fleet management, integration with mobile phones, precise timing, collection of mapping and geodetic data, and visual and voice navigation for drivers. At the time of this writing, no specific information was available on planned military applications.

4.5 Geodesy and Time Systems

4.5.1 Geodetic Reference System

Since August 1993, geodetic support for GLONASS has been provided by the national coordinate system of the Russian Federation, Parametry Zemly, or the Earth Parameter System 1990 (PZ-90) (see Figure 4.8). The PZ-90 system was established by the Russian Ministry of Defense to replace the previously used Soviet Geodetic System 1985 (SGS-85). PZ-90 is similar in quality to the Earth model employed in WGS-84, which is used by GPS [26]. The basic characteristics of PZ-90 are provided in Table 4.2 [27–29].

Since its inception, the PZ-90 coordinate system has had two revisions in order to improve consistency of broadcast orbits with WGS-84. The first revision was completed in 2002 (PZ-90.02) with the help of extensive data collects from geodetic satellites. PZ-90.02 was officially implemented with Decree 797, dated June 20, 2007. The latest enhancement, PZ-90.11, was introduced by Decree 1463 on December 28, 2012, and was implemented on December 31, 2013, at epoch 2010.0. Official support for orbital missions began on January 15, 2014 [30]. At one time, it was common practice to transform PZ-90 to or from other coordinate systems such as WGS-84 and ITRF (see Table 4.3). The latest realizations of PZ-90, ITRF, and WGS 84 are coincident at the 1-cm level, which obviates the need for such transformations for most applications (see Table 4.3 and Section 3.5.1.1.)

Figure 4.8 PZ-90 terrestrial network.

Table 4.2 PZ-90 Characteristics

Name and Designation of the Constant	*Unit of Measurement*	*Value for PZ-90.11*
Fundamental Geodetic Constants		
Angular rate of rotation of Earth (ω)	rad/s	$7.292\ 115 \times 10^{-5}$
Geocentric gravitational constant, including atmosphere (GM)	m^3/s^2	$398{,}600.44 \times 10^9$
Geocentric gravitational constant of atmosphere (GM_A)	m^3/s^2	0.35×10^9
Speed of light (c)	m/s	299,792,458
Parameters of the Common Terrestrial Ellipsoid		
Semimajor axis (α_ε)	M	6,378,136
Denominator of compression ($1/\alpha$)	unit of denominator	298.25784
Acceleration of gravity at the equator ($\gamma\varepsilon$)	Mgal	978,032.8
Correction in the acceleration of gravity, g, due to the attraction of atmosphere at sea level ($\delta\gamma_\alpha$)	mgal	–0.9
Other Constants		
Second harmonic coefficient ($J^0{}_2$)	—	$1{,}082{,}625.7 \times 10^{-9}$
Fourth harmonic coefficient ($J^0{}_4$)	—	$-2{,}370.9 \times 10^{-9}$
Normal potential on the surface of the common terrestrial ellipsoid (U_0)	m^2/s	62,636,861

Table 4.3 Transformation Parameters for PZ-90, PZ-90.02, PZ-90.11, WGS 84 (G1150), and ITRF2008

No.	*From*	*To*	*ΔX(m)*	*ΔY (m)*	*ΔZ (m)*	*ωX (mas)*	*ωY (mas)*	*ωZ (mas)*	*M (10⁻⁶)*	Epoch
1	PZ-90	PZ-90.02	–1.07 ±0.10	–0.03 ±0.10	+0.02 ±0.10	0	0	–130 ±10	–0.220 ±0.020	2002.0
2	WGS 84 (G1150)	PZ-90.02	+0.36 ±0.10	–0.08 ±0.10	–0.18 ±0.10	0	0	0	0	2002.0
3	PZ-90.11	ITRF2008	–0.003 ±0.002	–0.001 ±0.002	+0.000 ±0.002	+0.019 ±0.072	–0.042 ±0.073	+0.002 ±0.090	–0.000 ±0.0003	2010.0

ΔX, ΔY, ΔZ: linear elements of reference system transformation for transforming system 1 to system 2, *m*. ωX, ωY, ωZ: angular elements of reference system transformation for transforming system 1 to system 2, rad. *m*: scale element of reference system transformation for transforming system 1 to system 2.

Currently, Russia uses two National Reference Frames for surveying and mapping: the National Geodetic Reference Frame 1942 (SK-42) or the Krasovsky ellipsoid, and the Geodetic Reference Frame 1995 (SK-95). In 2017, Russia will transition all geodetic services from SK-42/SK-95 systems to a new national geodetic reference system called Geodetic Reference System 2011 (GRS-2011). The reference frame for GRS-2011 is provided by a network of around 50 astronomical geodetic stations (see Figure 4.9). Just like PZ-90.11, this system is aligned to ITRF at epoch 2011.0 [30, 31].

4.5.2 GLONASS Time

All GLONASS satellites are synchronized to a GLONASS Central Synchronizer (CS) time, which is kept in Moscow. The daily instability of the Central Synchronizer hydrogen clocks is not worse than 2×10^{-15}. GLONASS system time is also

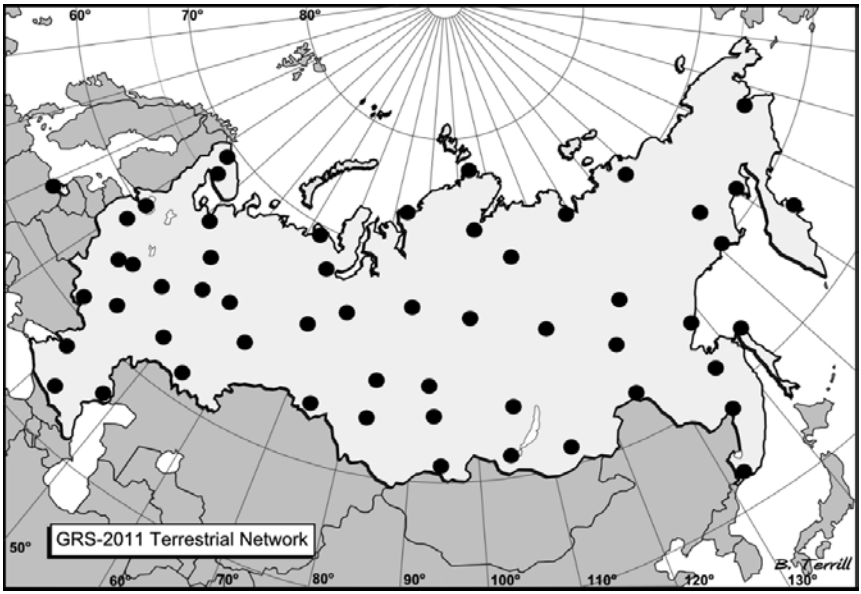


Figure 4.9 GRS-2011 terrestrial network.

synchronized to the Universal Time Coordinated of Russia, UTC (SU), which is maintained by the National Metrology Institute of the Russian Federation (VNIIFTRI) in Mendeleevo near Moscow [5, 7].

Periodically, the time scales of GLONASS satellites are compared with the CS time scale. Corrections are computed and uploaded by the ground control segment to the satellites twice a day. The GLONASS time scale is also periodically corrected to account for leap second adjustments. Typically, this correction is performed once a year or 1.5 years at midnight (00:00:00 on January 1, April 1, July 1, or October 1) by all UTC users. Users are generally notified at least 3 months in advance [5].

Given that the GLONASS time scale is periodically corrected to account for leap second adjustments, it is recommended that receivers simultaneously utilize the old and the corrected UTC (SU) (prior to and after the correction) in order to generate smooth and valid series of pseudorange measurements, and be able to resynchronize the data string time mark without loss of signal tracking.

4.6 Navigation Services

GLONASS provides an authorized (military) navigation and a civil navigation service similar to GPS. Both services are transmitted on both the L1 and L2 radio frequency bands. A new civil service in L3 has been added to newer Glonass-M and Glonass-K1 satellites. (The L3 signal is described in Section 4.7.9.

The high-accuracy (authorized) service is designated by VT (vysokaya tochnost, or high accuracy) by the Russians, and designed as the P-code in this chapter. The P-code is retained exclusively for Russian military use while the less accurate (open) service is for civil use [5]. The high-accuracy service is not encrypted; however, it has an anti-spoofing capability [32].

The open service is designated as ST by the Russians and designated as C/A code in this chapter. The C/A code is for military, civil and commercial use. By 2016, open service user positioning accuracy was estimated around 1.4m (horizontal), with that number eventually reaching 0.6m (horizontal) by 2020 [33].

Russia has developed several types of GLONASS differential services. (Differential services are described in Chapter 12.) They have deployed a coastal differential service for GLONASS and GPS using maritime radio beacons, similar to other services operating around the world. The Russians actively participated in RTCM Special Committee SC-104 that developed the series of standards that permit the seamless use of DGPS, differential GLONASS, and differential GPS/GLONASS services [1].

4.7 Navigation Signals

At the time of this writing, older Glonass-M satellites transmitted military and civil FDMA navigation signals at the L1 and L2 bands. The new Glonass-K1 and newer Glonass-M satellites transmit identical FDMA signals at L1 and L2 (see Section 4.7.1) and a new civil code division multiple access (CDMA) signal in the L3 band (see Section 4.7.9). At the time of this writing, the most recent ICD was Version 5.1 dated 2008 and only provides details on the FDMA L1 and L2 signals [5].

4.7.1 FDMA Navigation Signals

Unlike GPS, where each satellite transmits a unique PRN for each signal [e.g., one for the C/A modulation and one for the P(Y) modulation] on the same radio frequency (i.e., CDMA), each visible GLONASS satellite transmits the same PRN on a different radio frequency (i.e., FDMA) to distinguish between satellites in the constellation. Historically, GLONASS is the only SATNAV system to use FDMA modulation [5].

FDMA can result in larger, more expensive receivers because of the extra front-end components required to process multiple frequencies in some designs. In contrast, a CDMA signal can more easily be processed with the same set of front-end components. Section 8.3.10 provides details on design guidelines to adapt the receiver front end to process all of the SV signals centered at the GLONASS L1 frequency.

FDMA does have some redeeming qualities in terms of interference rejection. A narrowband interference source that disrupts only one FDMA signal would disrupt all CDMA signals simultaneously. Furthermore, FDMA eliminates the need to consider the interference effect between multiple signal codes (cross-correlation). Thus, GLONASS offers more frequency-based interference rejection options than GPS and also has a more simplified code selection criterion. GLONASS satellites transmit signals centered on two discrete L-band carrier frequencies. Each carrier frequency is modulated by the modulo-2 summation of either a 511-kHz or 5.11-MHz PRN ranging code sequence and a 50-bps data signal. This 50-bps data signal contains the navigation frames and is denoted as the navigation message. Figure 4.10 shows a simplified block diagram of the signal generator. Details of the fre-

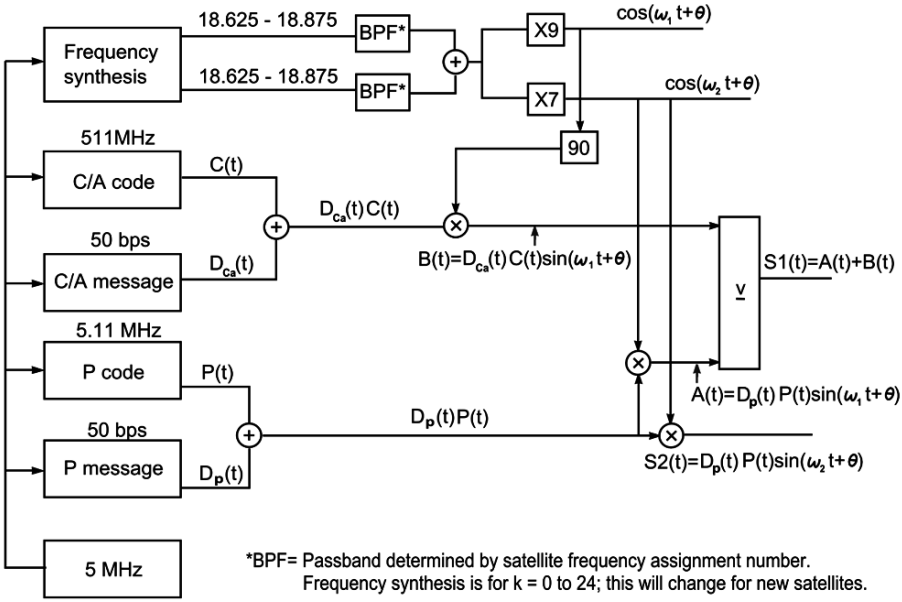


Figure 4.10 GLONASS signal generator. (Courtesy of Brian Terrill.)

quencies, modulation, PRN code properties, and navigation message are covered next [1, 34, 35].

4.7.2 Frequencies

Each GLONASS satellite is allocated a distinct pair of carrier frequencies, referred to as L1 and L2. Each of those distinct carrier frequencies is defined by the following equations:

$$f_{L1}(k) = (1602.0 + k0.5625)\ \text{MHz}$$
$$f_{L2}(k) = (1246.0 + k0.4375)\ \text{MHz}$$

where K is an integer value between –7 and + 6 that defines each distinctly allocated carrier frequency [5]. The spacing between adjacent frequencies on L1 is 0.5625 MHz and 0.4375 MHz for L2. Originally, K was a unique integer for each satellite and varied from 0 to 24. However, it was discovered that L1 signal transmissions interfered with radio astronomy measurements of the hydroxyl (OH) radical near 1,612 MHz. In accordance with recommendation by the International Telecommunications Union (ITU), in 1998, the initial frequency allocation was modified to k = 0, …, 12, and in 2005, negative channels were introduced, thus changing the channel range to k = –7, …,+6. This channel limitation is addressed by assigning the same K number to satellites on opposite sides of the Earth (antipodal). This center frequency modification has little effect on terrestrial users who cannot see antipodal satellites simultaneously [5].

The values of *K* listed above are the proposed values for satellites operating under normal conditions. Other values of *K* may be assigned for certain com-

Table 4.4 GLONASS Nominal L1 and L2 FDMA Frequencies

No. of Channel	*Nominal Value of Frequency in L1 Subband (MHz)*	*Nominal Value of Frequency in L2 Subband (MHz)*
06	1,605.375	1,248.625
05	1,604.8125	1,248.1875
04	1,604.2500	1,247.7500
03	1,603.6875	1,247.3125
02	1,603.125	1,246.4375
01	1,602.5625	1,246.000
00	1,602.0000	1,245.5625
–01	1,601.4375	1,245.5625
–02	1,600.8750	1,245.1250
–03	1,600.3125	1,244.6875
–04	1,599.7500	1,244.2500
–05	1,599.1875	1,243.8125
–06	1,598.6250	1,243.3750
–07	1,598.0625	1,242.9375

mand and control processing or under exceptional circumstances according to the Russians [1].

4.7.3 Modulation

In a similar manner to the legacy GPS signals, each satellite modulates its L1 carrier frequency with two PRN ranging sequences. One sequence, called the P-code, is reserved for military purposes. The other sequence, called the C/A-code, is for civil use and aids acquisition of the P-code. Each satellite modulates its L2 carrier frequency solely with the Modulo-2 summation of P-code and navigation data. The P-code and C/A-code sequences are the same for all satellites [1, 34, 35].

4.7.4 Code Properties

Both GLONASS and GPS use pseudorandom codes that facilitate satellite-to-user ranging and have inherent interference rejection. GLONASS C/A-code and P-code sequences are described next [1, 34, 35].

GLONASS C/A-code has the following characteristics:

- Code type: Maximum length 9-bit shift register;
- Code rate: 0.511 Mchips/s;
- Code length: 511 chips;
- Repeat rate: 1 ms.

A maximum-length code-sequence exhibits predictable and desirable autocorrelation properties (see Section 2.4). The 511-bit C/A-code is clocked at 0.511

Mchips/s, thus the code repeats every millisecond. This use of a relatively short code clocked at a high rate produces undesirable frequency components at 1-kHz intervals that can result in cross-correlation between interference sources, reducing the interference rejection benefit of the spread frequency spectrum. On the plus side, the FDMA nature of the GLONASS signal significantly reduces any cross-correlation between satellite signals because of the frequency separation. The reason for the short code is to allow quick acquisition, requiring a receiver to search a maximum of 511 code phase shifts. The fast code rate is necessary for range discrimination, with each code phase representing approximately 587m. Figure 4.11 shows the structure for shift registers used to generate the C/A-spreading code [5].

4.7.5 GLONASS P-Code

Because the P-code is strictly a military signal, there is very little Russian information available on the GLONASS P-code. Most P-code information is derived from analysis of the code performed by various independent individuals or organizations such as Dr. Peter Daly and his graduate students at the University of Leeds, United Kingdom. Based on such analysis, the P-code characteristics are [35–38]:

- Code type: Maximum length 25-bit shift register;
- Code rate: 5.11 Mchips/s;
- Code length: 33,554,432 chips;
- Repeat rate: 1 second (repeat rate is actually at 6.57-second intervals, but chipped sequence is truncated such that it repeats every 1 second) [35–38].

As with the C/A code, the maximum length code has exceptional, predictable, auto-correlation properties. The significant difference between the P-code and the C/A code is that the P-code is much longer compared to its clock rate, thus

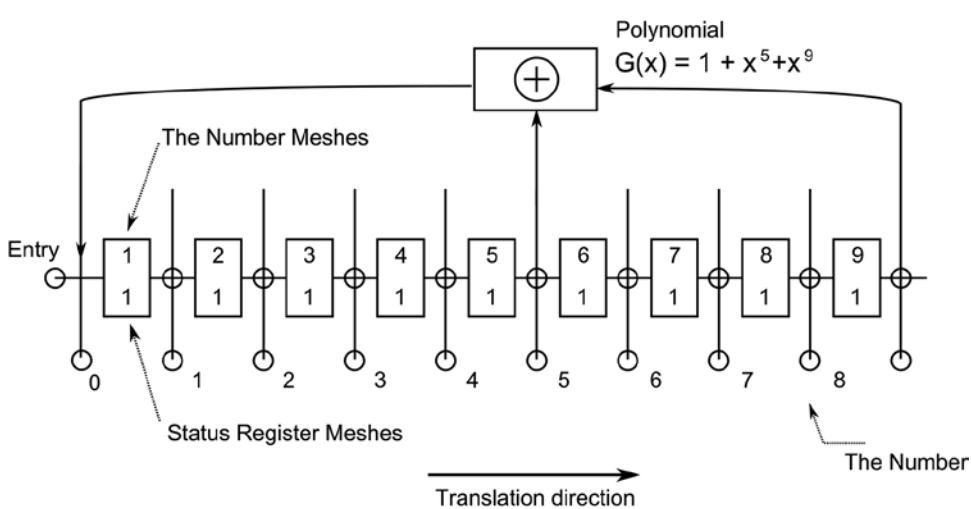


Figure 4.11 GLONASS C/A-code shift register [5].

repeating only once every second. Although this produces undesirable frequency components at 1-Hz intervals, the cross-correlation problem is not as severe as with the C/A-code. As with the C/A-code, FDMA virtually eliminates any problems involving cross-correlation between GLONASS satellite signals. While the P-code gains in terms of correlation properties, it sacrifices in terms of acquisition. The P-code contains 511 million code phase-shift possibilities. Thus, a receiver typically acquires C/A code first, and then uses the C/A code to help narrow the number of P-code phase shifts to search. Each P-code phase, clocked at 10 times the C/A code, represents 58.7m in range. A handover word (HOW) like the one used in GPS to facilitate handover to P(Y) code is not necessarily needed. The GLONASS P-code repeats once every second, making it possible to use the timing of the C/A code sequence to assist in the handover process. This is an example of one more design trade-off between the desired security and correlation properties of a long sequence and the desire for a faster acquisition scheme. GPS employs the former implementation while GLONASS employs the latter [35]. Figure 4.12 shows the structure for shift registers used to generate the P-code [36].

4.7.6 Navigation Message

Unlike GPS, GLONASS has two types of navigation messages. The C/A code navigation message is modulo-2 added to the C/A code at the satellite, whereas a P-code unique navigation message is modulo-2 added to the P-code. Both navigation messages are 50-bps data streams. The primary purpose of these messages is to provide information on satellite ephemeris and channel allocations. The ephemeris information allows the GLONASS receiver to accurately compute where each GLONASS satellite is located at any point in time. Although ephemeris is the predominant navigation information, there is an assortment of other items provided such as:

- Epoch timing;
- Synchronization bits;
- Error correction bits;
- Satellite health;
- Age of data;
- Spare bits.

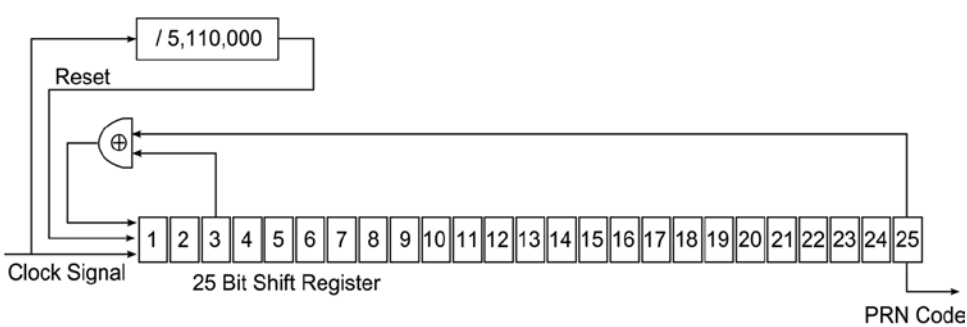


Figure 4.12 GLONASS P-code shift register [37]. (Courtesy of Brian Terrill.)

In addition, the Russians plan on providing data that will facilitate the combined use of GPS and GLONASS, particularly the difference between GLONASS system time and GPS system time. An overview of the C/A-code and P-code navigation messages is provided next [34, 35].

4.7.7 C/A Navigation Message

Each GLONASS satellite broadcasts a C/A-code navigation message that contains a superframe consisting of 5 frames. Each frame contains 15 lines with each line containing 100 bits of information. Each frame takes 30 seconds to broadcast, so the entire superframe is broadcast once every 2.5 minutes [5, 34].

The first 3 lines of each frame contain the detailed ephemeris for the satellite being tracked. Since each frame repeats every 30 seconds, a receiver will receive a satellite's ephemeris within 30 seconds once data reception begins [34].

The other lines of each frame consist primarily of approximate ephemeris (i.e., almanac) information for all the other satellites in the constellation. Each frame can hold the ephemeris for 5 satellites. Since the constellation will have 24 satellites, all 5 frames must be read in order to get the approximate ephemeris for all the satellites. This takes approximately 2.5 minutes [1, 34].

The approximate ephemeris information is not as accurate as the detailed ephemeris and is not used for the actual ranging measurement. Nonetheless, the approximate ephemeris is sufficient to allow the receiver to quickly align its code phase and acquire the desired satellite. Once acquired, the satellite's detailed ephemeris is used for the ranging measurement. As with GPS, the ephemeris information is often valid for hours. Therefore, a receiver does not need to continually read the data message in order to compute accurate position. Figure 4.13 shows the structure of C/A-code navigation message [5].

4.7.8 P-Code Navigation Message

The Russian military has not publicly published any specifics on their P-code. Nonetheless, a number of independent organizations and individuals have investigated the P-code waveform and published their results [34].

The following information is extracted from the published information. The important thing to remember is that the Russians publicly provided the detailed information on their C/A-code data message and have given certain guarantees regarding its continuity. No such information or guarantees exist regarding the P-code data. Thus, the P-code data structure described next may change at any time without notice.

Each GLONASS satellite broadcasts a P-code navigation message consisting of a superframe, consisting of 72 frames. Each frame contains 5 lines with each line containing 100 bits of information. Each frame takes 10 seconds to broadcast, so the entire superframe is broadcast once every 12 minutes [34].

The first 3 lines of each frame contain the detailed ephemeris for the satellite being tracked. Since each frame repeats every 10 seconds, a receiver will receive a satellite's ephemeris within 10 seconds once data reception occurs. The other lines of each frame consist primarily of approximate ephemeris information (i.e., almanac) for the other satellites in the constellation. All 72 frames must be read to get all

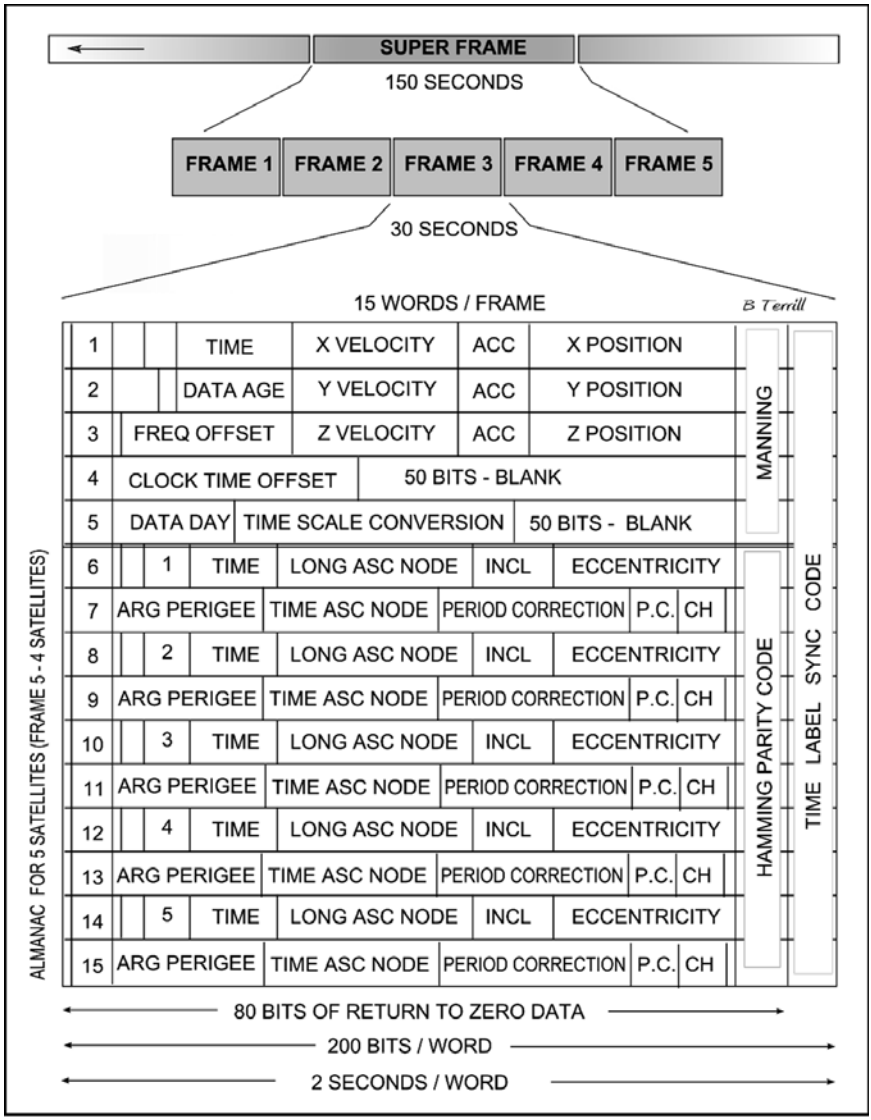


Figure 4.13 GLONASS C/A-code frame and message structure [5].

the ephemeris, taking 12 minutes [34]. Figure 4.14 shows the structure of P-code navigation message [35].

4.7.9 CDMA Navigation Signals

In 2012, as part of the GLONASS modernization efforts, Russia signed a new federal targeted program, Maintenance, Development, and Use of GLONASS 2012-2020, according to which the next-generation spacecraft would carry both legacy, as well as new signals with code division [9]. The new signals provide better accuracy, improved multipath resistance, and greater interoperability with other GNSS systems because of their CDMA modulation. The Glonass-K1 and new Glonass-M

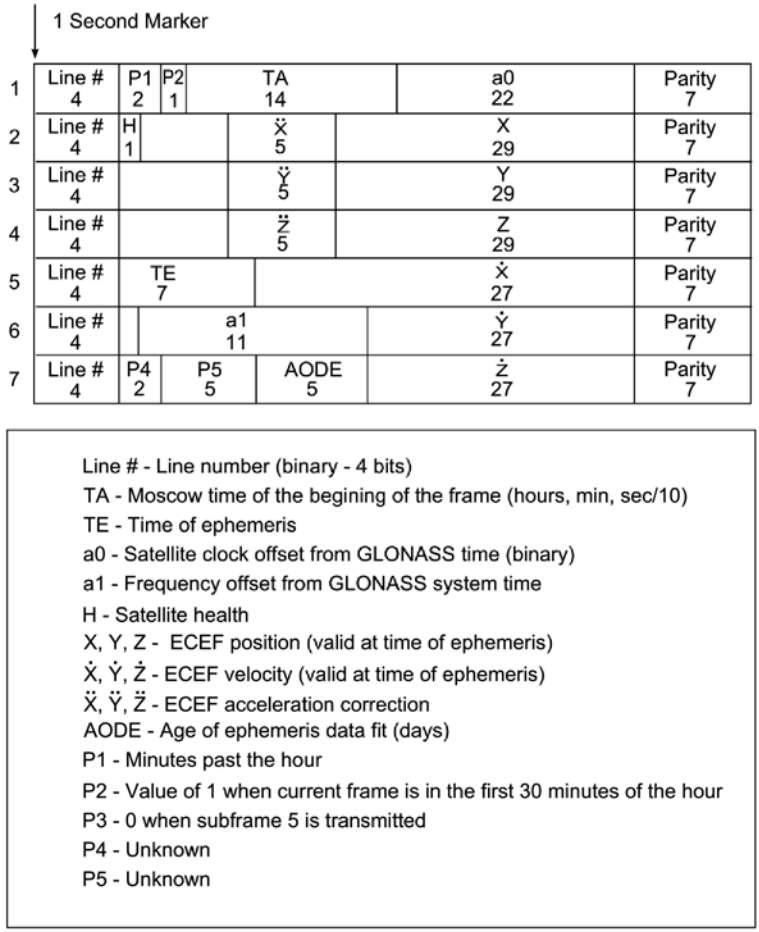
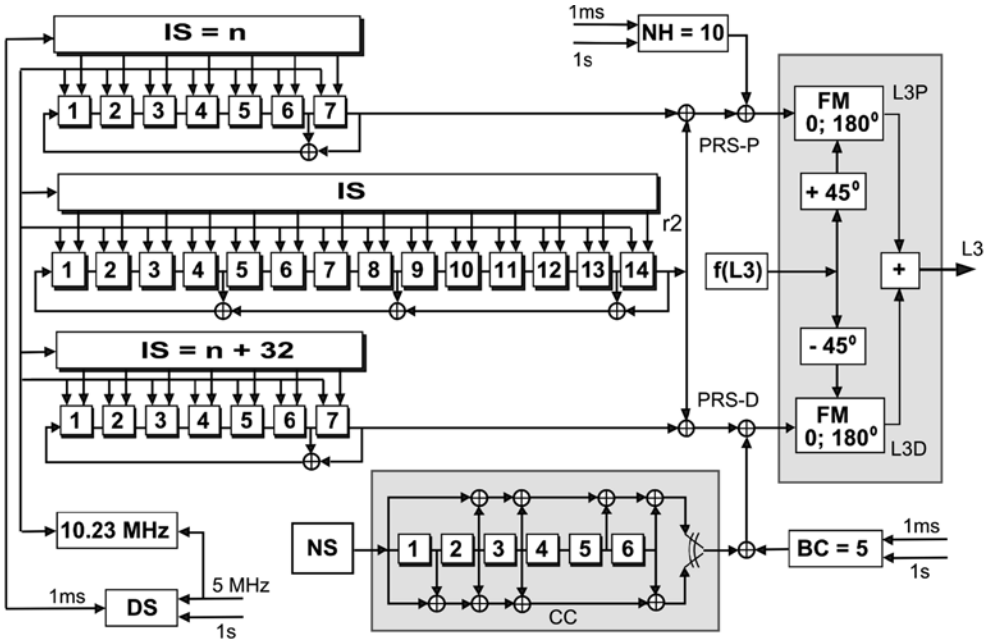


Figure 4.14 GLONASS P-code frame and message structure [38].

satellites transmit a test open service signal on L3 only, whereas future Glonass-K2 satellites will transmit signals in L1, L2, and L3, using a different design.

At the time of this writing, newly launched Glonass-K1 and Glonass-M satellites were transmitting a test civil CDMA signal in the L3 band designated as L3OC. The first satellite carrying the L3OC signal, Glonass-K1, was launched on February 26, 2011. Since 2014, newly launched Glonass-M spacecraft also carry the new L3OC signal. Although an updated GLONASS Interface Control Document was still pending publication, limited data on the signals has been published [39]:

- Frequency: The new GLONASS L3OC signal is centered at 1,202.025 MHz.
- Modulation: The L3OC signal is modulated onto the carrier using quadrature phase-shift keying (QPSK), with an in-phase data channel and a quadrature pilot channel (see Figure 4.15).
- Code properties: The L3OC signal has the following code characteristics:
  - Code type: Maximum length 9-bit shift register;



BC=Barker Code, CC=Convolution Code, NH=Neuman-Hofman Code,
NS=Navigation message symbol, IS=Initial State

Figure 4.15 Glonass-K1 L3 CDMA signal generator [39]. (Courtesy of Brian Terrill.)

  - Code rate: 10.23 Mchips/s;
  - Repeat rate: 1 ms [39].

- Navigation message: The L3OC navigation message consists of 8 navigation frames to facilitate complete information for a 24-satellite constellation broadcasting the L3OC CDMA signals. Each frame will include 5 strings and will last 15 seconds. The entire superframe will rebroadcast every 2 minutes. When the constellation reaches 30 satellites broadcasting CDMA L3OC, the navigation frames will increase to 10, and the length of the superframe to 2.5 minutes, respectively. Every navigation frame has a full set of ephemerides for the current satellite and part of the system almanac for three satellites. The full system almanac is broadcast in one superframe. A time marker is located at the beginning of a string and given as a number of a string within the current day in the satellite time scale [40].
- Future Glonass-K2 CDMA signals: At the time of this writing, only limited, fragmented information was available on the modulation signal structure and navigation message for the future civil and military CDMA signals on Glonass-K2 spacecraft. Like planned signals for GPS, Galileo, and Beidou, Glonass-K2 satellites will transmit signals using BPSK and BOC modulations. The two-channel structure (information/data and dataless pilot channel) already introduced with the text L3OC will be retained. Additionally, time division multiplexing will be added, as well as encrypted military channels.

  The ICD covering future CDMA signals is yet to be published; however, some preliminary information is available in Russian academic literature

Table 4.5 Future CDMA Signals on Glonass-K2

Signal	*Carrier Frequency (MHz)*	*Modulation*	*Code Rate (Mchips/s)*
L1OCd	1,600.995	BPSK 1	10.23
L1OCp	1,600.995	BOC (1,1)	10.23
L1SCd	1,600.995	BOC (5, 2.5)	2.5×10.23
L1SCp	1,600.995	BOC (5, 2.5)	5×10.23
L2OCd	1,248.06	BPSK 1	10.23
L2OCp	1,248.06	BOC (1,1)	10.23
L2SCd	1,248.06	BOC (5, 2.5)	2.5×10.23
L2SCp	1,248.06	BOC (5, 2.5)	5×10.23
L3OCd	1,202.025	TBD	10.23
L3OCp	1,202.025	TBD	10.23

[41]. The main known attributes of the future CDMA K2 signals are in Table 4.5.

- Future Glonass-KM CDMA signals: At the time of this writing, no design data was available on the signal structure on future Glonass-KM spacecraft. It is expected that the satellites will continue to carry the legacy FDMA signals, the CDMA signals already introduced on Glonass-K2, and introduce a new L5 signal centered at 1,176.45 MHz [42].

Acknowledgments

The information presented on GLONASS is revised and updated from previous editions of this book. The authors would like to thank Richard Clark and Jay Purvis, who coauthored the 1996 version of this chapter, and Richard Clark, who coauthored the 2007 version of this chapter. The primary source for the 1996 version was "Russia's Global Navigation Satellite System," which was produced under U.S. Air Force Contract Number F33657-90-D-0096. The authors would like to thank all the contributors and participants involved in effort. The contract was performed by ANSER (Washington, D.C.) with some assistance from the Russian Space Agency. ANSER assembled a team of Russian GLONASS experts in Russia to compile and author the report. The Russian authors included V. F. Cheremisin, V. A. Bartenev, and M. F. Reshetnov of the NPO Prikladnoy Mekhaniki (Applied Mechanics); Y. G. Gouzhva and V. V. Korniyenko of the Russian Institute of Radio Navigation and Time; N. E. Ivanov and V. A. Salishchev of the Scientific Research Institute of Space Device Engineering; Y. V. Medvedkov of the Russian Space Agency; V. N. Pochukaev of the Central Scientific Research Institute of Machine Building; M. N. Krasilshikov and V. V. Malyshev of the Moscow Aviation Institute; V. I. Durnev, V. L. Ivanov, and M. Lebedev of the Russian Space Forces; and V. P. Pavlov of the Flight Control Center. The team from ANSER included E. N. O'Rear and R. Turner from the Arlington office and S. Hopkins, R. Dalby, and D. Van Hulle from the Moscow office. In addition, the authors would like to thank the

following navigation experts who reviewed the initial draft of the ANSER report and provided many valuable comments: P. Misra of Lincoln Laboratory; L. Chesto, former chairman of the RTCA Special Committee 159; and J. Danaher and Jacques Beser, formerly of 3-S.

References

[1] ANSER, "Russia's Global Navigation Satellite System," Arlington, VA, ANSER, U.S. Air Force Contract Number F33657-90-D-0096, May 1994.

[2] Anodina, T. G., "The GLONASS System Technical Characteristics and Performance, Working paper FANS/4-WP/75," International Civil Aviation Organization, Montreal, Canada, 1988.

[3] Kazantsev, V. N., et al., "Overview and Design of the GLONASS System," *Proc. Int. Conference on Satellite Communications*, Volume II ,Moscow, Russia, October 18–21, 1994, pp. 207–216.

[4] "On the Activity on Application of the Global Navigation Satellite System GLONASS," Russian Federation Governmental Decree 237, March 7, 1995, http://www.glonass-center.ru/decree.html.

[5] Global Navigation System GLONASS Interface Control Document (Version 5.1), Moscow, 2008, http://aggf.ru/gnss/glon/ikd51ru.pdf.

[6] "The Decree of the President of the Russian Federation," Decree 38-RP, February 18, 1999, http://www.glonass-center.ru/38rp_e.html.

[7] "Declaration of the Government of the Russian Federation," Russian Federation Governmental Decree 346, March 29, 1999. http://www.glonass-center.ru/decl_e.html.

[8] "On Use of GLONASS for the Benefit of Social and Economic Development of the Russian Federation," Presidential Decree 638, May 17, 2007.

[9] "Maintenance, Development and Use of the GLONASS System 2012-2020," Federal Targeted Program, Russian Federal Space Agency, http://www.federalspace.ru/115/.

[10] Feairheller, S., "The Russian GLONASS System: A US Air Force/Russian Study," *Proc. 7th Intl. Technical Meeting of Satellite Division of U.S. Institute of Navigation*, Salt Lake City, UT, September 20–23, 1994, pp. 293–304.

[11] Lebedev, C. M., "Space Navigation System 'GLONASS'-Application Prospective," *Scientific Information Coordination Center for Military Space Forces, Proc. RTCA 1994 Symposium*, Reston, VA, November 30–December 1, 1994, pp. 199–210.

[12] Bassevich, A. B., et al., "GLONASS Onboard Time/Frequency Standards: Ten Years of Operation," *Russian Institute of Radionavigation and Time*, 1996, www.dtic.mil/dtic/tr/fulltext/u2/a501103.pdf.

[13] Kulik, S. V., "Status and Development of GLONASS," *First United Nations/United States of America Workshop on the Use of Global Navigation Satellite Systems*, Kuala Lumpur, Malaysia, August 20–24, 2001. (Removed from the Internet in 2004), http://www.jupem.gov.my/gnss_bm.htm.

[14] Kulik, S. V., "GLONASS: Status and Progress," *Second United Nations/United States of America Regional Workshop on the Use of Global Navigation Satellite Systems*, Vienna, Austria, November 26–30, 2001, http://www.oosa.unvienna.org/SAP/act2001/gnss2/presentations/index.html.

[15] Revnivykh, S., "Status and Development of GLONASS," *Third UN/USA Workshop on the Use and Applications of Global Navigation Satellite Systems, for the benefit of Latin America and the Caribbean*, Santiago, Chile, April 1–5, 2002, http://www.oosa.unvienna.org/SAP/act2002/gnss1/presentations/index.html.

[16] Revnivykh, S., "Status and Development of GLONASS," *Fourth UN/USA Workshop on the Use of Global Satellite Positioning Systems, for the benefit of Africa*, Lusaka, Zambia, July 15–19, 2002, http://www.oosa.unvienna.org/SAP/act2002/gnss2/presentations/index.html.

[17] Polischuk, G., et al. "The Global Navigation Satellite System GLONASS: Development and Usage in the 21st Century," *34th Annual Precise Time and Time Interval (PTTI) Meeting*, tycho.usno.navy.mil/ptti/ptti2002/paper13.pdf.

[18] Revnivykh, S., "Developments and Plans of the GLONASS System," *UN/USA International Meeting of Experts the Use and Applications of Global Navigation Satellite Systems*, Vienna, Austria, November 11–15, 2002.

[19] Revnivykh, S., "Developments of the GLONASS System and GLONASS Service Interface," *Joint Meeting of Action Team on Global Navigation Satellite Systems and Global Navigation Satellite Systems Experts of UN/USA Regional Workshops and International Meeting 2001-2002*, Vienna, Austria, December 8–12, 2003.

[20] International Telecommunication Union, "GLONASS System Information, December 8, 2003," *First Consultation Meeting Forum*, Geneva, December 8–9, 2003, International Telecommunications Union Web site.

[21] Vagyn, Y., et al., "Kosmicheskaya Sistema Kontrolya Soblyudeniya Soglasheniy O Zapreshcheniy Ispitaniy Yadrenovo Oruzhiya," *Aerospace Courier*, Vol. 6, No. 83, 2012, pp. 68–69.

[22] "Sanctions Delay Russia's Glonass-K2 Program," *GPS World*, December 17, 2014, http://gpsworld.com/sanctions-delay-russias-glonass-k2-program/.

[23] Revnivykh, S., "GLONASS Status and Modernization," *International GNSS Committee IGC-7*, Beijing, China, November 4–9, 2012, http://www.unoosa.org/pdf/icg/2012/icg-7/3-1.pdf.

[24] Koshelyaevsky, N. B., and S. B. Pushkin, "National Time Unit Keeping over Long Interval Using an Ensemble of H-Maser," *Proc. 22nd Annual Precise Time and Time Interval Applications and Timing Meeting*, Vienna, VA, December 14–6, 1990, pp. 97–116.

[25] Qualcomm, "Performance by Utilizing GPS and GLONASS Satellite Networks for Greater Location Accuracy," May 23, 2011, https://www.qualcomm.com/news/releases/2011/05/23/qualcomm-enhances-mobile-location-performance-utilizing-gps-and-glonass.

[26] International Telecommunication Union, "Technical Description and Characteristics of Global Space Navigation System GLONASS-M - Information Document," Document 8D/46-E, November 22, 1994.

[27] International Telecommunication Union, "Technical Description and Characteristics of Global Space Navigation System GLONASS-M - Information Document," Document 8D/46(Add.1)-E, December 6, 1994.

[28] Boykov, V. V., V. F. Galazin, and Y. V. Korablev, "Geodesy: Application of Geodetic Satellitesfor Solving the Fundamental and Applied Problems," *Geodeziya i Katografiya*, No. 11, November 1993, pp. 8–11.

[29] Boykov, V. V., et al., "Experimental of Compiling the Geocentric System of Coordinates PZ-90," *Geodeziya i Katografiya*, No. 11, November 1993, pp. 18–21.

[30] Military Topographic Department of The general Staff of Armed Forces of the Russian Federation, "Parametry Zemli 1990 (PZ-90.11) Reference Document," Moscow, Russia, 2014.

[31] Vdovin, V., and M. Vinogradova, "National Reference Systems of the Russian Federation, Used in GLONASS Including the User and Fundamental Segments," *International GNSS Committee IGC-8*, Dubai, United Arab Emirates, November 11, 2013, http://www.unoosa.org/pdf/icg/2012/icg-7/3-1.pdf.

[32] Grygoriev, M. N., and C. A. Uvarov, *Logistics*, Moscow, Russia: Yuright Publishing, 2012.

[33] Lyskov, D., "GLONASS Policy, Status and Evolution," *International GNSS Committee ICG-8*, Dubai, United Arab Emirates, November 10, 2013, http://www.unoosa.org/pdf/icg/2013/icg-8/2.pdf.

[34] Beser, J., and J. Danaher, "The 3S Navigation R-100 Family of Integrated GPS/GLONASS Receivers: Description and Performance Results," *Proc. of the U.S. Institute of Navigation National Technical Meeting*, San Francisco, CA, January 20–22, 1993, pp. 25–45.

[35] Stein, B., and W. Tsang, "PRN Codes for GPS/GLONASS: A Comparison," *Proc. of the US Institute of Navigation National Technical Meeting*, San Diego, CA, January 23–25, 1990, pp. 31–35.

[36] Lennen, G. R., "The USSR's Glonass P-Code - Determination and Initial Results," *Proc. of 2nd Intl. Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1989)*, Colorado Spring, CO, September 1989, pp. 77–83.

[37] Biradar, R. L., "Architecture and Signal Structure of GLONASS," *IJITE*, Vol. 3, No. 1, January 2015, http://www.ijmr.net.in email id- irjmss@gmail.com.

[38] Daly, P., and S. Riley, "GLONASS P-Code Data Message," *Proc. of the 1994 National Technical Meeting of The Institute of Navigation*, San Diego, CA, January 1994, pp. 195–202.

[39] Thoelert, S., et al., "First Signal in Space Analysis of Glonass-K1," *Institute of Communications and Navigation German Aerospace Center (DLR) and Stanford University*, 2011.

[40] Urlichich, Y., et al., "Innovation: GLONASS. Developing Strategies for the Future," *Russian Space Systems*, http://www.spacecorp.ru/upload/iblock/a57/glonass_eng.pdf.

[41] Lypa, I., "Development and Research of Algorithms for Future GLONASS Signals Acquisition with Modulation on the Subscriber Frequencies," *Specialy 05.12.14 – Radiolocation and Navigation, Moscow Power Engineering Institute*, 2016.

[42] Revnivykh, S., "GLONASS Status, Development and Application," *Second International Committee on Global Navigation Satellite Systems*, Bangalore, India, September 4–7, 2007, www.unoosa.org/pdf/icg/2007/icg2/presentations/o5.pdf.

CHAPTER 5

Galileo

Daniel Blonski, Igor Stojkovic, and Sylvain Loddo

GNSS has become the standard means of navigation, positioning, and timing with widespread applications in a large variety of fields. Recognizing the strategic importance of these applications, Europe developed its own GNSS strategy in the early 1990s. This strategy resulted first in a European SBAS, EGNOS (see Section 12.6.1.2), and then continued towards the implementation of a complete European SATNAV system referred to as Galileo. The focus of this chapter is to provide a detailed description of the technical aspects of Galileo.

5.1 Program Overview and Objectives

The Galileo program is Europe's initiative for a state-of-the-art SATNAV system, providing a highly accurate global positioning and timing service under civilian control. Galileo will provide Europe with independence in satellite navigation but it will also be interoperable with the other SATNAV systems.

In 1999, the European Commission (EC) and the European Space Agency (ESA) established the fundamental need for the development of a European GNSS component [1, 2]. Based on the experience with EGNOS and consultations with global stakeholders, the following key objectives were identified:

- To increase control of satellite-based safety critical navigation systems;
- To ensure positioning and timing services for European users with the objective of reducing the risk in case of a policy change affecting access to GPS;
- To support the competitiveness of European industries in the global SATNAV market and to grant access to the development of GNSS technologies.

These objectives were analyzed as part of the Galileo comparative system studies conducted by ESA in the period 1999 to 2000. The result of the studies was a recommendation to develop a global SATNAV system with a similar design as the existing GPS and GLONASS systems. The early design phases were cofunded by both ESA and EC.

The European Union (EU) is the owner of the Galileo system and its 28 member states are important stakeholders of the program. The EC, being the executive body of the EU, is the Program Manager of the European GNSS program.

ESA is the technical design authority of the Galileo system. Since the 1990s, ESA has led SATNAV activities in Europe. In the 2000s, ESA and its industrial partners were driving the consolidation of the system design based on the Galileo System Test-Bed (GSTB) activities targeting the validation of ground processing techniques. The GIOVE A (Galileo In-Orbit Validation Phase Element) and GIOVE B satellites, launched in 2005 and 2008, respectively, validated new space technologies in orbit (e.g., the clocks) [3]. Those early activities were the foundation of the system development leading to a successful In Orbit Validation (IOV) campaign based on the first complete version of the Galileo ground segments and four IOV satellites in 2013 [4]. Accomplishing a successful service validation campaign, performed throughout 2016, the European Commission declared the start of the Galileo Initial Services on December 15, 2016 [5].

Today, ESA is leading the completion of the system through its FOC, expected to be reached in 2020. ESA is responsible for the finalization of the development and deployment of the Galileo system. As part of this role, ESA is in charge of the operational validation and will hand over the infrastructure in incremental system builds to the EC and the European GNSS Agency (GSA) for service provision and exploitation. ESA is supported by industrial contractors providing system engineering technical assistance (SETA). The SETA prime contractor is Thales Alenia Space Italia (TAS-I), which is supported by Thales Communications France (TCS) Airbus Defence and Space Germany (ADS-G).

GSA is supporting the EC for the promotion, commercialization, operation, and exploitation of the European GNSS infrastructure: EGNOS and Galileo. The GSA, on behalf of the EC, is managing EU GNSS Framework Programme activities and ensures the certification and accreditation of the system components of Galileo and EGNOS.

5.2 Galileo Implementation

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 are:

1. The IOV phase provides 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, implementing all key functions of a global SATNAV system.
2. The FOC phase 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 the number of remote elements and satellites.

As part of the IOV phase, two experimental satellites, GIOVE A (launched in December 2005) and GIOVE B (launched in April 2008), contributed to the characterization of the radiation environment in MEO and the validation of ground processing techniques that evolved from the early GSTB experimentation. The broadcast of the experimental GIOVE signals secured the spectrum required for Galileo in accordance with International Telecommunication Union (ITU) World Radiocommunication Conference frequency allocations to the RNSS. Furthermore, both satellites enabled performance testing of critical payload technology (e.g., AFSs, radiation hardened digital technology) in the Galileo MEO target orbit. The satellites, together with a ground segment prototype, allowed end-to-end testing of the fundamental system concepts before the development of elements of the final system was completed [3].

The objectives of the IOV phase were accomplished with the completion of the IOV Test Campaign, during which the core functions of the final Galileo system have been successfully tested.

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 GSA for service provision and exploitation.

5.3 Galileo Services

The Galileo system is expected to meet a variety of user needs. The set of specified services form the basis of the system design and operations and have been used to consolidate the main features of the Galileo navigation system. However, the capabilities of the system will serve a much larger range of applications, well beyond the scope of the defined services. This section focuses on a description of the Galileo services identified to form the core mission. These services will be provided worldwide and independently from other SATNAV systems, by using the signals broadcast from the Galileo satellite constellation.

The reference services envisaged for the Galileo FOC phase are: the Open Service (OS), the Commercial Service (CS), the Public Regulated Service (PRS), and the support for the satellite-aided Search and Rescue (SAR) service. Beyond providing the reference services, the signals emitted by Galileo satellites are interoperable with other GNSS signals, and therefore enable a much wider range of applications relying on utilization of multiple GNSS constellations in parallel. Key aspects of interoperability are addressed in Section 5.6. The expected system performance of the FOC system is presented in Section 5.8.

5.3.1 Galileo Open Service

The Galileo OS will provide publicly accessible PVT information to worldwide users through the ranging signals on three frequencies designated as E1, E5a, and E5b. This service is suitable for mass-market applications, such as in-car navigation or personal navigation by mobile phones. The targeted dual-frequency performance of the OS, as defined in the High Level Definition document [6], is summarized in Table 5.1.

Table 5.1 Galileo OS Performance Design Targets

Open Service—Dual Frequency E1/E5a or E1/E5b	
Coverage area	Global
Position accuracy [95%]	4m/8m
UTC timing accuracy [95%]	30 ns
Availability of service over system lifetime	99.5%

This performance will be achievable by users at any point in the global service area with a very high availability provided that the users are equipped with receivers that track and process signals of all Galileo satellites with an elevation higher than 5° above the local horizon. The minimum performance levels of the Galileo Initial Open Service have been published in the Open Service Definition Document [7], following the declaration of Initial Services by the EC (see Sections 5.8.2.2 and 10.2.4.1).

5.3.2 Public Regulated Service

The public regulated service (PRS) will provide PVT capabilities to government-authorized users requiring a higher level of protection (e.g., increased robustness against interference or jamming). The PRS signals will be encrypted, and access to the service will be controlled through a government-approved secure key distribution mechanism. The PRS will only be accessible through receivers equipped with a PRS security module loaded with a valid PRS decryption key. The PRS service will be provided in the E1 and E6 bands.

5.3.3 Commercial Service

The CS will allow the development of professional applications by supporting the dissemination of value-added data on a dedicated commercial service signal. The CS features a global broadcast of such value-added data in real time in the E6 band. The currently envisaged services that might be provided by means of the CS signal are related to high accuracy and authentication [8].

5.3.4 Search and Rescue Service

SAR/Galileo, the Galileo search and rescue service, comprises the forward link alert service (FLS) providing timely and accurate detection and localization of emergency beacon alerts and the return link service (RLS) providing a means to deliver short messages to emergency beacons equipped with Galileo receivers.

The Galileo constellation is equipped with repeaters that relay alarms from 406-MHz distress beacons to globally distributed MEO local user terminals (MEOLUTs). An interface with the Galileo infrastructure is implemented to enable return link messages to the SAR users providing information such as that the rescue operation is engaged. This return link capability will be provided through the Galileo navigation signals themselves. The SAR/Galileo service is fully integrated into the MEOSAR cooperative international effort by members of the COSPAS-SARSAT organization. The minimum performance levels of the Galileo Initial SAR

Service have been published in the SAR Service Definition Document [9], following the declaration of Initial Services by the EC.

5.3.5 Safety of Life

The initial Safety of Life (SOL) service was intended for safety-critical user applications. As part of the Galileo Mission Consolidation Review, carried out by the EC with the involvement of the major program stakeholders, the provision of a European SOL service through Galileo has been revised. The SOL service is currently under reprofiling and the implementation of dedicated SOL functions has been postponed until later phases of the program. A future SOL service might rely on existing regional solutions and might envisage also a collaborative approach with other GNSS constellation providers [10, 11].

This re-profiling of the SOL service allowed to alleviate the driving SOL targets imposed on the global infrastructure of Galileo. The resulting relaxation of the demanding performance targets allowed a significant reduction of system infrastructure. As part of this reprofiling, the Galileo system is required to support the SOL application by means of the OS signals. In addition, a future version of the EGNOS will be designed to provide the related corrections and integrity information to safety critical user communities.

The focus of this Galileo chapter is on the public signals of the OS and on the SAR/Galileo services.

5.4 System Overview

The core infrastructure of the Galileo system consists of three segments: the ground segment, the space segment, and the user segment (see Figure 5.1). The objective of this section is to provide additional information on the ground and space segments and a summary of the Galileo System elements is provided in Table 5.2. Each segment has distinct functions to allow the overall system to perform its mission, providing navigation services on a global scale.

The Galileo space segment will consist of a constellation of 24 operational satellites distributed over three planes with additional in-orbit spare satellites. The total number of satellites in orbit will be 30. Each satellite will broadcast the navigation signals, the navigation data provided by the ground segment together with its own time stamps, and will relay SAR alerts.

The Galileo ground segment is composed of two parts: the ground mission segment (GMS) and the ground control segment (GCS). The ground mission segment contains all functions necessary to determine the navigation message data and to disseminate those data to the satellites. For this purpose, it comprises a network of 16 globally distributed sensor stations monitoring the satellite signals, the relevant processing facilities to determine the orbit and clock correction, message generation, service monitoring and contact planning through two Galileo control centers (GCCs), and 5 globally distributed mission uplink stations (ULSs). A global communication network interconnects the centers and the stations. In addition, the GMS provides interfaces to external service providers and service facilities.

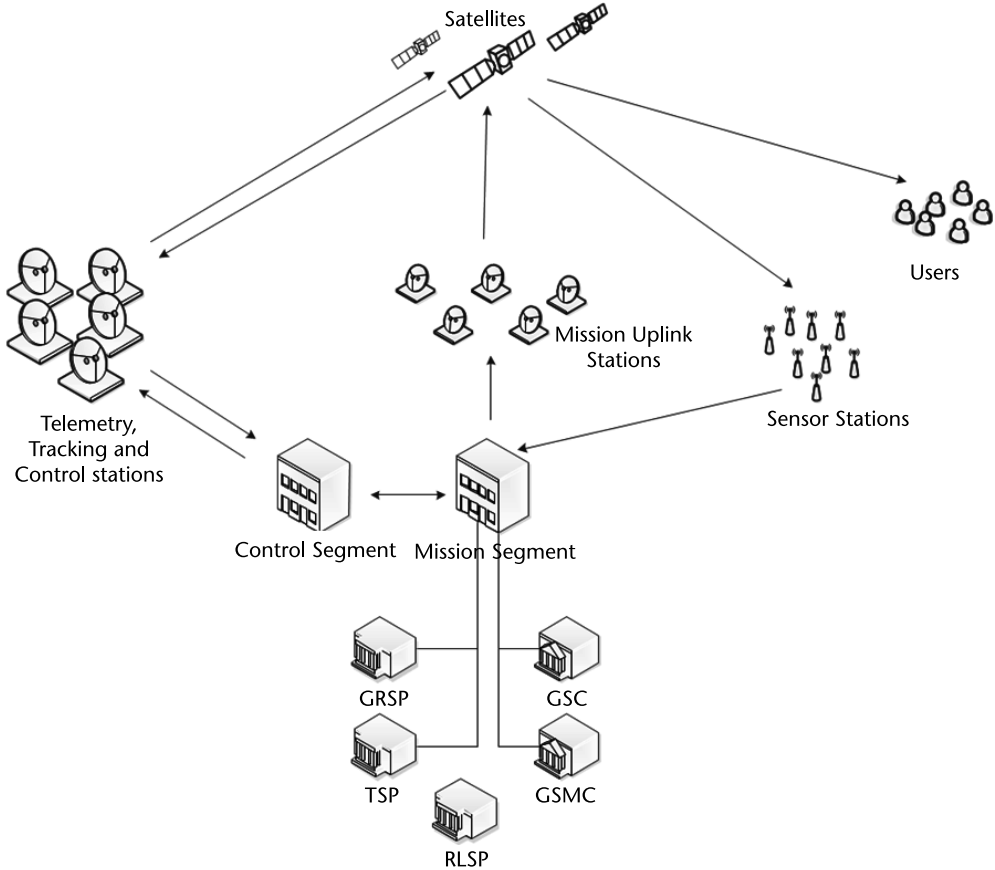


Figure 5.1 Galileo high-level system architecture and system context.

The GCS contains all functions necessary to operate each individual satellite and to maintain the overall constellation geometry. For this purpose, it contains 5 telemetry, tracking and control (TT&C) stations distributed globally and the related ground processing facilities at both control centers to monitor, control, and maintain each satellite platform and its payload to allow the dissemination of the mission data with high availability. Each segment is described in more detail in the following sections. The prime contractor for the GMS has been Thales Alenia Space (TAS-F) and the prime contractor for the GCS Airbus Defence & Space (ADS-UK).

The Galileo system will provide its services to various types of end-user applications which rely on SATNAV receivers for the utilization of the Galileo signals. A variety of such applications can be envisaged exploiting the different Galileo signals and messages. For the purpose of end-to-end verification and testing, several Galileo test user receivers have been manufactured and deployed in the field. Furthermore, dedicated test campaigns for mass market and professional receivers are carried out by ESA, GSA and the EC Joint Research Center. The user segment is not further elaborated as part of this chapter, for which the focus is on the system aspects.

In addition to the above mentioned core elements, the system is supported by a number of external service facilities [7].

Table 5.2 Galileo FOC Architecture Elements Supporting the Navigation Services

Galileo Infrastructure	*Final FOC configuration*
Space segment	
Constellation	30 satellites in MEO Walker 24/3/1 constellation
Satellites	4 IOV satellites, 24+ FOC satellites
Ground segment	
Galileo Control Centres	Two complete control centers, GCC-I in Fucino, Italy and GCC-D in Oberpfaffenhofen, Germany, each equipped with operational GMS for service provision operational GCS for satellite control
TT&C stations	Up to 6 TT&C stations
Mission uplink stations (ULS)	5 ULS sites
Galileo sensor stations (GSS)	16 sites
Communications network	Diverse-routed links between each GCC and the remote sites
In-orbit test (IOT) center	Redu IOT connected to both GCCs
External support	
Launch site (LS)	Soyuz and Ariane 5 Launch sites in Kourou connected to GCSs
External satellite control centers	LEOP Centre in CNES Toulouse, LEOP Centre in ESA/ESOC Darmstadt, connected to GCSs and external TT&C stations
Service facilities	Galileo Security Monitoring Centres (France and United Kingdom)
	Time service provider
	Geodetic reference service provider
	Galileo Reference Centre
	Galileo Service Centre
	Return link service provider

- The Galileo Time Service Provider (GTSP) delivers the relevant steering data to allow a highly accurate realization of UTC at the user level based on Galileo System time (GST).
- The Geodetic Reference Service Provider (GRSP) ensures the alignment of the Galileo Terrestrial Reference Frame (GTRF) with the ITRF by regularly computing the sensor station locations in this reference frame. The GRSP will also carry out an independent verification and calibration of the orbital products of Galileo using satellite laser ranging data from the International Laser Ranging Service (ILRS).
- The European GNSS Service Centre (GSC) provides detailed information on the status of the Galileo system to the public and acts as a point of contact for users of the OS. Furthermore, the GSC provides the interface between the core system and CS providers.
- The Galileo Security Monitoring Centre (GSMC) provides security monitoring of the system as well as management functions for the PRS users. It acts as the interface for governmental entities to the Galileo system.
- The Galileo Reference Centre (GRC) will independently monitor the quality of the provided Galileo services during the Galileo Exploitation Phase.
- The SAR/Galileo Service Centre hosts the infrastructure for coordinating and supporting the provision of the SAR FLSs and RLSs and will perform

SAR performance monitoring based on reference beacons. Three European MEOLUTs, located on the far corners of the European SAR coverage area (ECA), provide the Galileo/SAR/FLS by detecting and localizing emissions of SAR distress beacons. The SAR/Galileo RLS will rely on the dissemination capabilities of the Galileo core system to notify SAR/Galileo users with an acknowledgement of their distress transmission.

- The Galileo system has interfaces to external satellite control centers for the support of the Launch and Early Operations Phase (LEOP). The LEOP support is provided by 2 LEOP control centres located at the ESA/European Space Operation Centre and at the Toulouse Space Centre of the French National Centre for Space Studies (CNES).
- The In-Orbit Test (IOT) activities are supported by a dedicated IOT station located in the ESAs Redu Centre in Belgium. The IOT station is equipped with a high-gain antenna and a measurement system for detailed characterization of the Galileo L-band signal-in-space. Furthermore, ultrahigh frequency (UHF) transmitters are available to support SAR transponder commissioning. The IOT station also contains a C-band uplink to support satellite commissioning and operations.

5.4.1 Ground Mission Segment

The Galileo GMS is performing the mission related tasks such as computation and dissemination of the navigation data as well as generating GST. The GMS is composed of centralized elements located at the 2 GCCs and remote elements distributed globally to allow good coverage of the Galileo satellite constellation. The key processing functions of the ground segment and their interactions are shown in Figure 5.2.

The remote elements are:

- A network of 16 Galileo GSSs used for L-band ranging measurements from each Galileo satellite and for monitoring the Galileo signals in space. The measurements are used for orbit determination, time synchronization (ODTS) and the supervision of the products provided by the GMS.
- A network of 5 Galileo mission ULSs disseminates the mission-related data (navigation, SAR, CS, other navigation-related products) from the GMS to each Galileo satellite via C-band.

The remote elements are connected to the two GCCs through a high-performance communication network denoted as the Galileo Data Dissemination Network (GDDN). Both geographically redundant GCCs, one in Oberpfaffenhofen (Germany) and one in Fucino (Italy), are fully redundant and contain all the facilities for the navigation processing and monitoring and control of the system and its elements.

The GMS is connected to external facilities which provide data to the system such as the time and terrestrial reference products needed for the ODTS processing and GST generation.

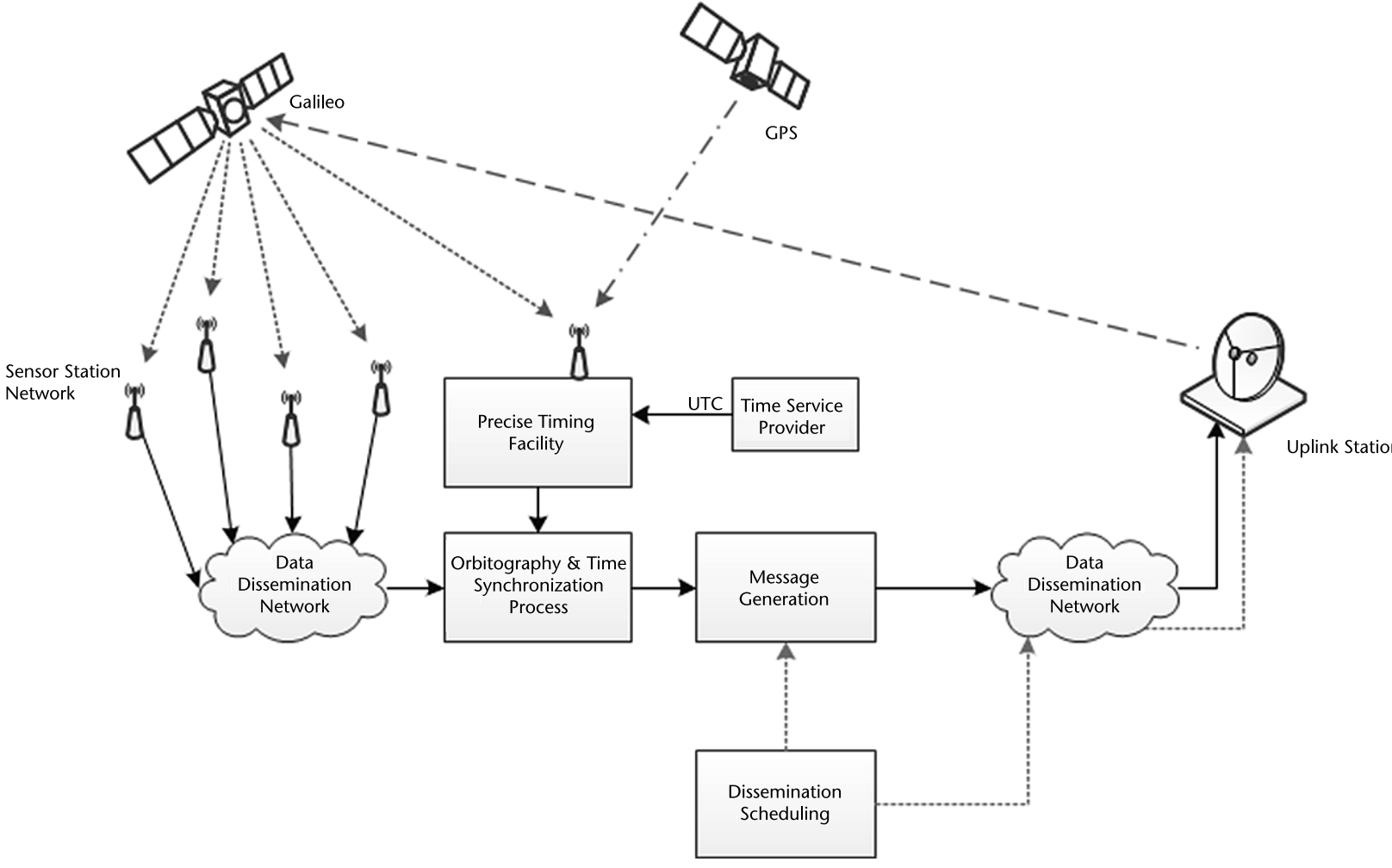


Figure 5.2 High-level functional chain for Galileo navigation processing.

In addition to the remote stations, the GMS contains several mission-essential facilities that are centralized at the GCCs. In addition to the data-processing, timekeeping, and message generation facilities that will be further described in this section, the GMS also contains auxiliary functions such as monitoring and control of the ground assets, data archiving, mission planning and ULS contact scheduling, online mission monitoring, and offline data analysis and data exchange with external entities. Furthermore, the GMS provides operations training, system security, and communication network-related functions. These functions are rather generic for systems of such complexity and importance as Galileo and are therefore not further discussed in detail here.

The core functions supporting the navigation mission are implemented in the orbit and synchronization processing facility (OSPF) that estimates the orbital location of the satellite and the offset and drift of its clock and, based on this data, predicts satellite ephemeris and clock bias to be used for the navigation message broadcast.

The Precise Time Facility (PTF) is the time source of Galileo, generating GST and synchronizing all system assets to GST.

The Message Generation Facility (MGF) builds the navigation messages to be uplinked. The functions and interactions of the key elements of the GMS are best described by elaborating the Navigation Message determination and dissemination process.

5.4.1.1 Galileo System Time Generation

The Galileo GMS generates GST using the atomic clocks of the PTF. The GST is linked to TAI and the required time synchronisation is performed with either TWSTFT or Common View technique links with the TSP. There are two PTFs deployed in the Galileo GMS, one in each GCC, one acting as the master PTF and the other acting as the slave PTF. The slave PTF is steered to the master PTF using PTF-to-PTF links. The Galileo-to-GPS time offset (GGTO) is estimated by the PTF, which relies on synchronization links with GPS system time via the USNO and a coordination interface aligned to a PTF check algorithm. (Note that the Galileo-to-GPS Time Offset is one version of the GGTO cited in Tables 3.17 and 3.18.

Each of the two PTFs contains two active hydrogen masers (AHM) in hot-redundancy and four Caesium clocks. The output of the master AHM steered to UTC modulo 1 second constitutes the physical realisation of the GST.

The Galileo system follows the recommendation 460-4 of the ITU and provides information that allows users to derive UTC from GST. The steering correction required to maintain the close alignment of the GST with UTC is provided by the GTSP. The future GTSP will ensure that GST is maintained within 50 ns (95%) to UTC modulo 1 second.

In the context of multiple GNSS constellations providing their services to users globally, there is a need to provide sufficient information to allow transparent utilization and seamless transition between the various systems. The Galileo and GPS system time scales are established independently from each other, but both time scales are steered to UTC. The difference between both timescales is the GGTO. The GGTO is determined by the Galileo PTF and broadcast as GGTO offset and drift through the Galileo navigation message. In the Galileo system, there

are redundant techniques implemented to derive the GGTO. The baseline for the IOV phase has been to derive the GGTO based on TWSTFT between the USNO and the Galileo PTF. For better accuracy and improved stability, the final Galileo Ground Segment implements a calibrated combined Galileo/GPS receiver to derive the GGTO based on signal measurements of both systems.

5.4.1.2 Navigation Data Generation

A worldwide-distributed network of GSSs continuously collects observables of the Galileo signals-in-space. The GSS receivers make pseudorange measurements referenced to the time provided by the GSS local AFS-based clock. The products of the ODTS algorithm depend on the quality and quantity of these observables. The quality of the observation data is to a large extent also driven by the receiver design and the local receiver environment (in particular for multipath and RF spectrum interference). The GSS reference receiver is a high-accuracy receiver especially manufactured for this function. It includes an atomic clock for time keeping and time stamping of the observation data.

The GSS sites were carefully selected to be at locations with benign RF environments. Before being selected as a reference site, the local RF environment had been well characterized. The coverage of the GSS network is shown in Figure 5.3. A regular monitoring of data quality parameters and network connectivity is performed for each individual receiver to ensure the high quality and availability of the observations used to compute the navigation messages.

The performance and sizing of this real-time network was initially driven by the demanding SOL requirements. After the reprofiling of the SOL service, the GSS network has been adapted to the needs of the OS and PRS, which still demand high data quality and availability. The resulting network consists of 16 GSS sites with redundant hardware and geographic diversity to provide tolerance to faults.

The GSS network provides the real-time observables and the received navigation messages to the GCCs for processing. The data transmission is performed via the GDDN, which consists of dedicated leased communication lines that include a very small aperture terminal (VSAT) infrastructure and landlines depending on the location of the remote sites.

The collected observations are routinely processed by the ODTS function, which generates a new navigation data set once every 10 minutes based on the predicted evolution of the satellite orbit and clock. This ODTS estimation process is based on a least squares algorithm that is estimating the orbit and clock of all satellites for each service individually. These estimates are based on signals of the respective service (see mapping in Table 5.5).

The near-future behavior of the clock and orbit are predicted based on these estimates. In order to generate the full navigation data set, the evolution of the orbit and clock for each satellite are predicted over the next 24 hours. The tuning of the algorithm is such as to ensure a minimum ranging error for lower ages of data.

The full navigation dataset is divided into 8 batches of 3 hours each, identified by a different Issue of Data (IOD) parameter. The predicted orbit and clock information is then parameterized for dissemination in the navigation message. The message encoding details can be found in [12].

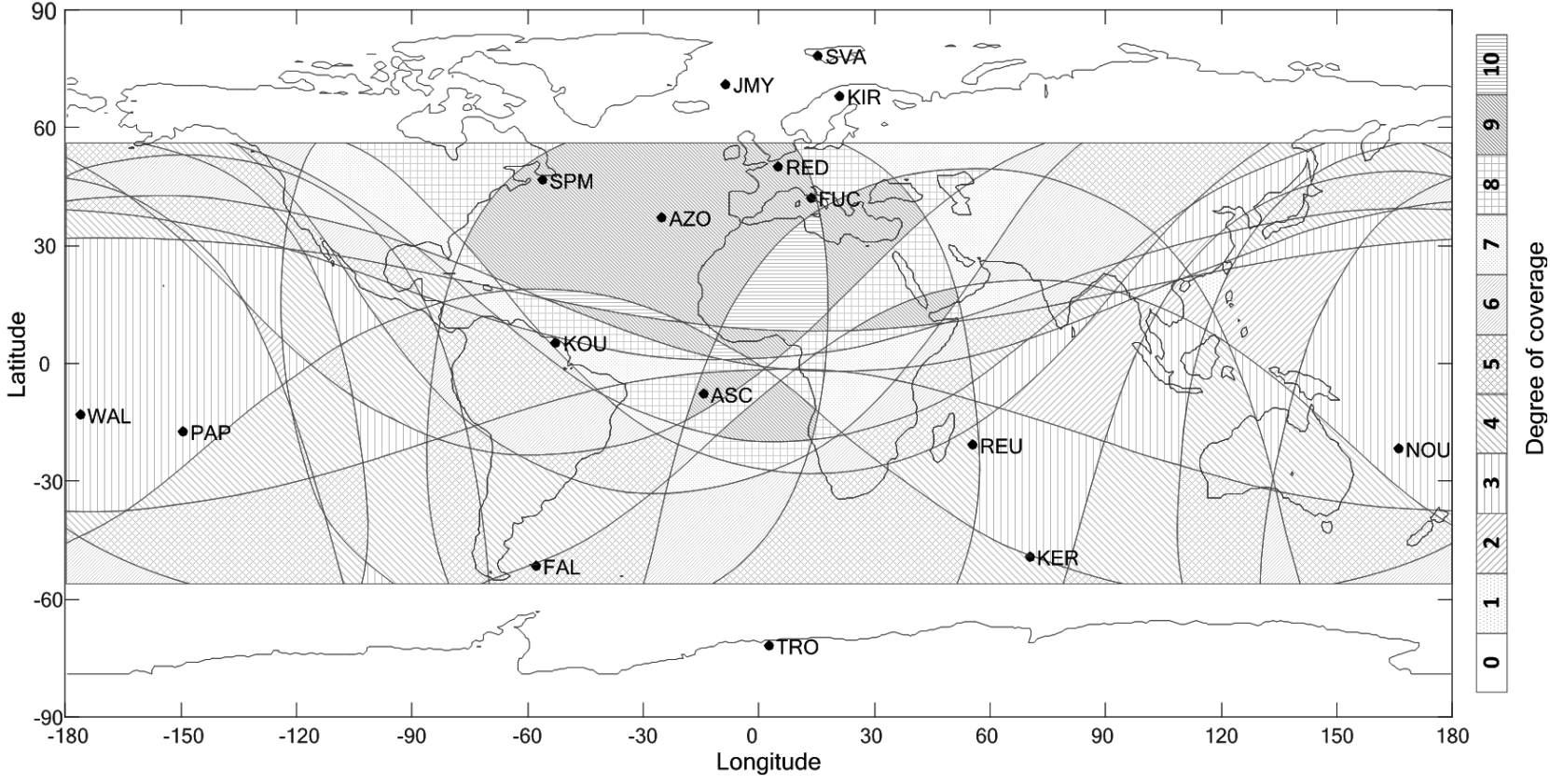


Figure 5.3 Galileo ground reference receiver network.

This split of the navigation information in batches is done to reduce errors of the message, when parameterized and encoded for dissemination. The preparation of the navigation data and the incorporation of auxiliary data (partly stemming from external sources, for example, the SAR RLM) are done by the MGF.

The first four batches of each new set are disseminated to the satellites by the ULS network, while the complete set is uplinked through the TTC. The targeted ODTS error has been a design driver for the mission uplink network. The ULS network sites, identified in Figure 5.4, have been selected such that consecutive ULS contacts to a specific satellite occur within 100 minutes for a fault-free state of the system. This duration is linked to the ODTS accuracy over time, while the design target was to ensure the user ranging error (URE) is better than 0.65m (1σ). Based on early clock experimentation results, the message refresh rate was derived to be 100 minutes for RAFS type clocks [13]. The much higher stability of the satellite-based passive hydrogen maser clock (PHM) and its better predictability by the ODTS process allows for longer prediction times in case of ground segment uplink failures (see also Section 5.4.3.2).

To achieve a regular uplink of the navigation data and ensure the dissemination of other data, such as the SAR RLS message or the CS data, a dedicated facility computes the most efficient uplink schedule to satisfy the needs of each service based on the available system resources (satellites and ULSs). The corresponding algorithm aims at maintaining the broadcast of up-to-date navigation data and health information by each Galileo satellite. The uplink scheduling function ensures the high quality of the ranging signals during nominal operations by controlling the aging of the broadcast messages.

During a scheduled ULS contact, the most recent navigation message batches are uplinked. The first batch is immediately broadcast by the satellite after complete reception. This process ensures that the most up-to-date navigation information is available to the users of the corresponding service.

Until another ULS contact is made, the stored navigation messages are broadcast sequentially. Each single valid message is continually broadcast until it reaches an age of 3 hours. At that point, the satellite will start broadcasting the next batch stored onboard. The aforementioned dissemination scheme and the imposed constraints have been important criteria for the design of the global ULS network with its 5 uplink stations.

The end-to-end process, starting from data collection at the GSS and completed by the application of the navigation data by the user, takes some time. The difference between the reference time and the time of utilization of the predictions by the user is called the Age of Data, which is determined by the refresh rate of the navigation message stored onboard that is driven by the uplink scheduling process and the availability and latency of the dissemination infrastructure. The lower the refresh rate, the higher the potential error of the orbit and satellite clock parameters in the message.

In early 2015, a significant improvement of the ranging performance of the Galileo system was achieved through a major upgrade of the GMS [14]. An upgrade of the elements for the determination and dissemination of the navigation data was rolled out to the operational infrastructure. As part of this upgrade, the GSS network and the mission uplink network have been expanded with additional locations improving the coverage of the constellation for both observables

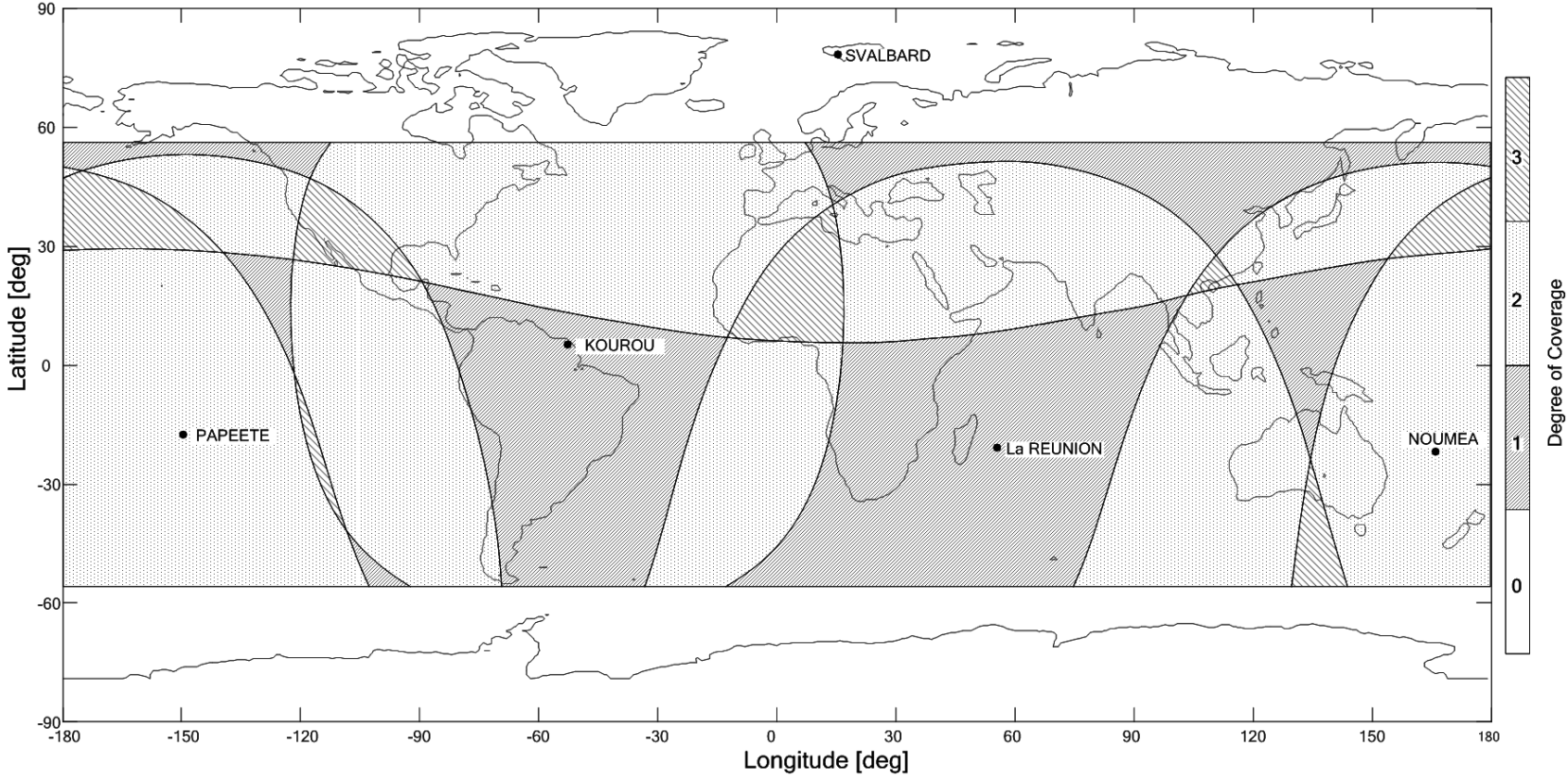


Figure 5.4 Galileo Mission Uplink network.

collection and navigation message uplink. As a result of this upgrade, the ranging performance has significantly improved and is already consistent with the targets defined for the final Galileo system.

5.4.2 Ground Control Segment

The GCS performs all functions related to the command and control of the satellite constellation. It supports operations and maintenance activities of the individual Galileo satellites. The key functions of the GCS are the following:

- Monitoring and control of the operational satellites via periodic contacts between the satellites and the TTC stations;
- Shared operations and data synchronisation between the two GCCs, acting in a master-backup configuration to ensure redundancy;
- Short-term planning of satellite operations;
- Flight dynamics;
- Support to operations preparation, training, and validation activities;
- Ground assets monitoring and control.

The platform and payload operations and maintenance activities such as upgrades of the onboard software, telemetry analysis, and planning and execution orbit-keeping maneuvers are core tasks of the GCS. The maintenance of the constellation geometry includes also recovery operations in order to address contingency situations and satellite failures with the objective of minimizing the time a satellite is not contributing to the service provision.

The GCS consists of centralized redundant elements inside each GCC and a network of 6 remote Galileo TT&C stations with 13-m antennas working in S-band for command and control of the Galileo satellites. The real-time GCS functions comprise the transmission of satellite tele-commands, that is, the reception and processing of satellite telemetry and the monitoring and control of ground assets. The nonreal-time GCS functions provide support for the real-time operations through satellite contact planning, flight dynamics, operations preparation, and secure key management.

Routine GCS operations are automated and performed in accordance with a short-term plan, the execution is supervised by operators. In contrast, critical operations are performed manually, with the support of machine executable procedures.

5.4.3 Space Segment

5.4.3.1 Constellation Geometry and Orbit Design

The reference geometry of the Galileo constellation is the result of detailed studies optimizing the number of satellites for the provision of the end user services. The system design initially resulted in a satellite constellation with 27 operational satellites in a Walker 27/3/1 constellation. Each of the three orbital planes was intended to have one inactive spare satellite to recover faster from satellite ultimate failures [15, 16]. This configuration was found to be the optimum for the provision of the

SOL and OS services. The SOL was driving the constellation geometry with its demands on satellite failure tolerance, low UERE, and the resulting minimum satellite elevation angle of 10°. As a result of the SOL reprofiling [10], the minimum user elevation angle has been reduced to 5° resulting in an optimization of the reference constellation geometry and a reduction of the number of operational satellites. The Galileo space segment after the completion of the deployment will comprise 24 operational satellites in a Walker 24/3/1 constellation. The three orbital planes are equally spaced and inclined 56° with respect to the equator. Each plane in the nominal constellation contains eight orbital reference slots separated by 45°.

Another major objective for the orbit selection was high service availability and, derived from this, the need to reduce the number of orbit-keeping maneuvers [17]. The selected orbit has a semimajor axis of 29,600.318 km (or 23,222-km altitude); this leads to a repeat cycle of the satellite-Earth geometry of 17 orbits in 10 sidereal days.

This cycle is short enough to allow repeatability of measured characteristics while being long enough to minimize gravitational resonances. After the initial orbit fine positioning, only one station-keeping maneuver is needed during the lifetime of a satellite.

To maintain the quality of the provided services over the system lifetime, two spare satellites will be deployed in each orbital plane. These spare satellites will reduce the time needed to recover from failures in the constellation. In case one operational satellite is terminally failed, a spare satellite will be manoeuvred to replace it within a few days, rather than to prepare and launch on-ground spare satellites which can easily take up to several months. Table 5.3 provides the Keplerian parameters of the Galileo reference slots.

Table 5.3 Galileo Constellation Orbital Parameters

Parameter		*Value/Derivation*
Semimajor axis	a	29,600.318 km
Inclination	i	56°
Galileo constellation reference epoch	T_0	21 March 20 1000 : 00 : 00.0 UTC
Right ascension of ascending node	Ω_{ref}	$\Omega_{ref} = \Omega_0 + 120° \cdot (n_{plane} - 1) + \dot{\Omega} \cdot (T - T_0)$
RAAN of Plane A at Galileo reference epoch	Ω_0	25°
Mean RAAN drift	$\dot{\Omega}$	$-0.02764398 \frac{°}{\text{day}}$
Plane identifier	n_{plane}	1, for Plane A 2, for Plane B 3, for Plane C
Argument of latitude	u	$u_0 + 45° \cdot (n_{slot} - 1) + 15° \cdot (n_{plane} - 1) + D_{nom} \cdot (T - T_0)$
Argument of latitude Slot A01 at Galileo reference epoch	u_0	338.333°
Mean rotation	D_{nom}	$613.72253566 \frac{°}{d}$
Slot identifier	n_{slot}	1 to 8

$\dot{\Omega}$ has been computed considering the nonspherical nature of the Earth's field of gravitation as well as the effects of the Sun and the Moon. D_{nom} is derived from the ground track repeat cycle of 17 orbit revolutions over 10 sidereal days.

In order to maintain the good geometric properties of the Walker constellation and the resulting DOP, the actual position of each satellite is allowed to deviate from the ideal slot position with a tolerance of ±2° in both the along-track and across-track directions. This along-track tolerance is important for maintaining the relative distance to neighbouring satellites. This requires precise adjustment of the satellite's velocity.

During the fine positioning of the satellite at the end of the LEOP phase, the inclination and RAAN are optimized within the tolerances of the station keeping box (pre-biased) to ensure that the satellite's across-track position over several years is nominal. The biases are selected considering the satellite drift due to gravitational forces, solar radiation pressure and other satellite external and internal disturbances. The objective is to maintain the orbital position within the slot tolerance limits without the need for fuel-intense out-of-plane orbit keeping maneuvers.

The first Galileo satellites (GSATs) 0101/0102 were launched into the assigned slots of the initially planned Walker 27/3/1. In order to not jeopardize the objectives of the IOV campaign, it was agreed to launch the second pair of IOV satellites into slots of the same constellation geometry (Walker 27/3/1). The IOV SV slot positions were taken into account in the design of the new reference geometry (Walker 24/3/1), while launching into slots of the originally planned 27/3/1 constellation allowed for an execution of the IOV test campaign according to the initial plans. Furthermore, the inclusion of the IOV slots within the reference 24/3/1 geometry avoided fuel-consuming orbit correction maneuvers and related major operational efforts. Figure 5.5 depicts the current state of deployment and indicates the reference geometry.

Satellites at the end of their operational life or after a nonrecoverable payload failure will be decommissioned and placed into a graveyard orbit that is at least 300 km above the nominal operational constellation taking into account the need for reducing collision avoidance maneuvers and the resulting service degradation.

5.4.3.2 Satellites

The Galileo constellation currently under deployment is composed of two families of satellites. The satellites of both families are 3-axis attitude controlled and have an approximate mass of 700 kg and a design lifetime of 12 years.

The first four satellites, GSAT0101 up to 0104, were procured and launched to form the space component of the Galileo IOV phase. These satellites, manufactured by EADS Astrium GmbH (now Airbus D&S) as the satellite prime contractor, have a size of 2.7m × 14.5m × 1.6m (x, y, z in satellite reference frame) and a mass of approximately 700 kg and are generating 1,420-W power. The first two IOV satellites were launched into the constellation B plane on October 21, 2011.The second pair of IOV satellites were launched into the C plane on October 12, 2012. Both launches employed a Soyuz-ST launch vehicle with a Fregat upper stage.

The satellites of the second family, GSAT02xx, are under production by OHB System AG as prime contractor. The contracted 22 satellites will form the core of

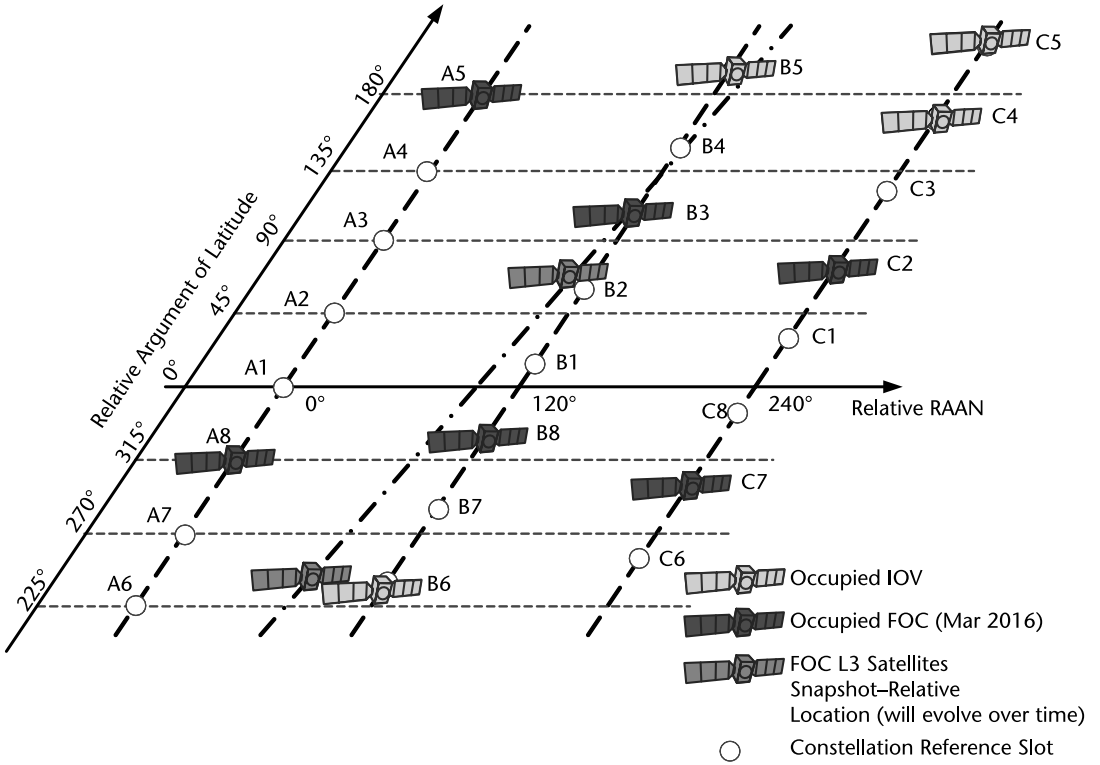


Figure 5.5 Galileo constellation geometry.

the Galileo FOC constellation. Eight FOC satellites have already been launched by Soyuz/Fregat from Kourou. The first FOC launch, in August 2014, injected GSAT0201/0202 into eccentric orbits (see more details in Section 5.4.3.3). The subsequent four Soyuz launches (March, August, and December 2015 and May 2016) successfully deployed 8 satellites into the Galileo constellation. On November 17, 2016, the first quadruple launch of Galileo satellites was performed successfully by an Ariane 5 launch vehicle [18]. The GSAT02xx type satellites have a slightly higher mass and are able to generate more power than the IOV satellites (approximately 730 kg and 1.9 kW). The dimensions of the FOC satellite with deployed solar arrays are 2.5m × 14.7m × 1.1m (x, y, and z).

The satellites of both families, although different by design, have similar components and a similar architecture that is described in more detail in the next sections [19, 20].

Galileo Satellite Platform Architecture

The Galileo satellite platform houses all subsystems required to operate the satellite: the TT&C subsystem, the attitude and orbit control, the propulsion and the data handling subsystems, as well as the thermal control and power subsystems (see Figure 5.6). Because the deployment of Galileo satellites is based on direct orbit injection, the propulsion subsystem is designed only for orbit correction manoeuvres requiring limited delta-v capability. The basis of the propulsion system is a set

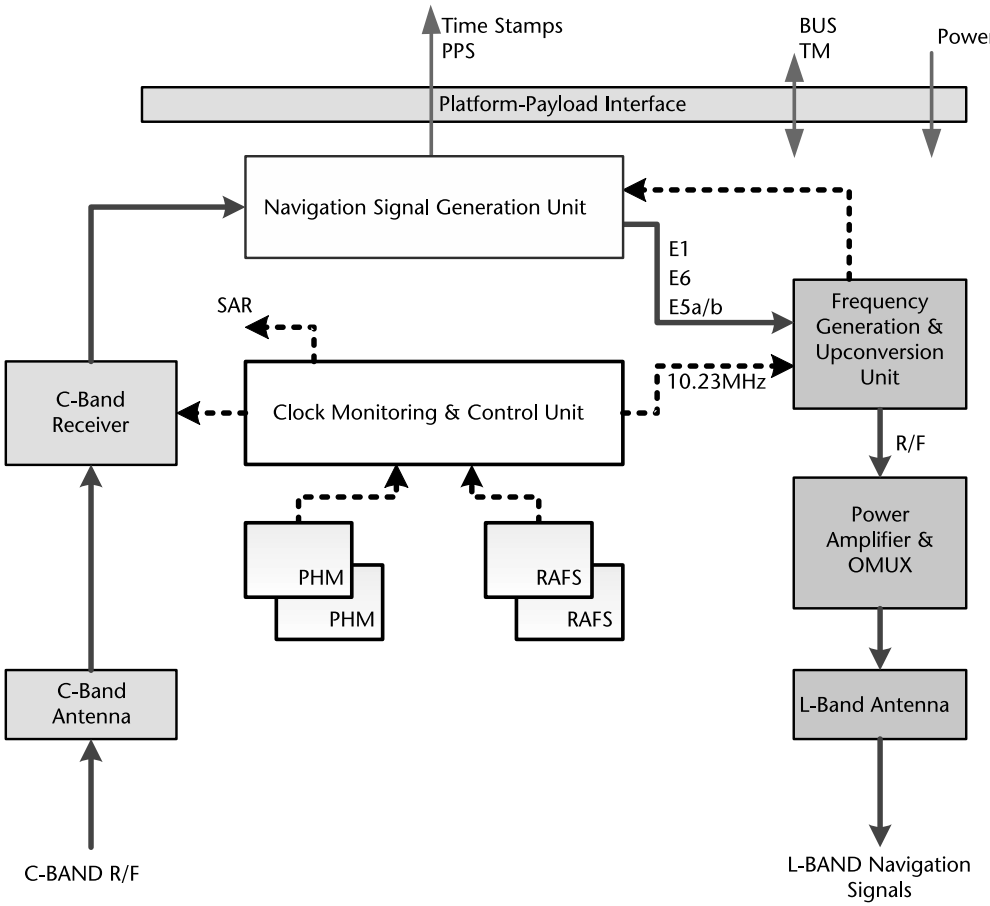


Figure 5.6 Galileo satellite platform simplified architecture diagram.

of eight monopropellant thrusters. Each thruster provides a nominal thrust of 1N using monopropellant grade hydrazine.

The attitude and orbit control subsystem (AOCS) performs the 3-axis attitude control during all phases of flight and during orbit manoeuvres. To achieve the full pointing performance during nominal operations of the satellite, the AOCS relies on Earth and Sun sensors to determine the orientation of the satellite. The Earth sensor operates with infrared light to detect the edge of the Earth-based on the temperature difference between the cold background of deep space and its warm atmosphere. The Sun sensor determines the angle to the Sun based on its emissions in visible-light wavelengths. Gyroscopes are used to monitor the angular rates of the satellite.

The active parts of the AOCS are the reaction wheels that generate the angular momentum to steer the satellite attitude in the 3 axes. When the reaction wheels reach their operational limits, magneto-torquers are used to discharge the accumulated momentum. The magneto-torquers are electromagnetic coils fixed to the structure of the satellite, that when powered, generate a magnetic field that interacts with the Earth's magnetic field to produce a mechanical momentum. In

non-nominal operational modes, during the LEOP as well as during contingency operations, safe modes (Earth acquisition, Sun acquisition), and end-of- life operations, the AOCS can also use the thrusters for orbit and attitude control. The AOCS controls the attitude of the satellite such that it rotates twice per orbit around its yaw axis. This change in attitude is performed to ensure an optimum orientation of the solar arrays towards the Sun to maximize the efficiency of solar cells and harvest the maximum amount of solar power for the satellite.

The power subsystem generates, stores, distributes, and regulates the power needed by the satellite to perform its mission. The Galileo satellite power subsystem is based on a 50-V bus architecture (for both satellite families). The main elements are the two solar arrays collecting the electrical power during Sun exposure times, a Li-Ion battery to provide the needed power during eclipses and a distribution and conditioning unit that ensures that each payload and platform element has sufficient power to operate. Redundant elements of the satellite are powered off to reduce power consumption.

The thermal control subsystem ensures that the temperatures inside the satellite stay within tight operational limits to ensure a stable environment for the sensitive navigation payload, especially the atomic clocks and the RF subsystem. Temperatures inside the satellite are controlled by thermistors radiating heat inside the satellite during the eclipse periods and by radiators on the sides of the satellite to dissipate excess heat when the satellite is exposed to sunlight.

The onboard computer, the central element of the data handling subsystem, steers all subsystems of the satellite. It collects and stores key information of the platform and payload elements that are embedded in the downlinked telemetry to assess the operational health of the satellite. The onboard computer receives its instructions through the TT&C subsystem from the GCS.

The TT&C subsystem provides redundant two-way communications in S-band in both ESA standard TT&C mode and spread spectrum mode. This subsystem supports accurate ranging and range-rate (Doppler) measurements as input to the flight dynamics facility of the GCS. The TT&C system relies on two orthogonal circularly polarized hemispherical helix antennas situated on opposite sides of the satellite for communication with the ground. Both antennas together ensure omni-directional coverage for reception and transmission of telemetry and telecommands during any operational mode and any orientation of the satellite.

As a passive component of the platform, the Galileo satellites carry a laser retro reflector that allows ranging between the ILRS stations and the satellites with an accuracy of a few centimeters.

Galileo Satellite Payload

Figure 5.7 shows the main elements of the Galileo payload: the C-band antenna, the mission receiver, the timing subsystem, the navigation signal generation unit (NSGU) and the L-band antenna.

The C-band antenna receives the mission data as a dedicated CDMA C-band uplink from the mission ULS. The uplink data stream is processed by the mission receiver and the contained navigation data are stored onboard. These data, together with satellite-generated data such as the clock information provided by the timing subsystem, are compiled into the navigation messages by the NSGU. Other

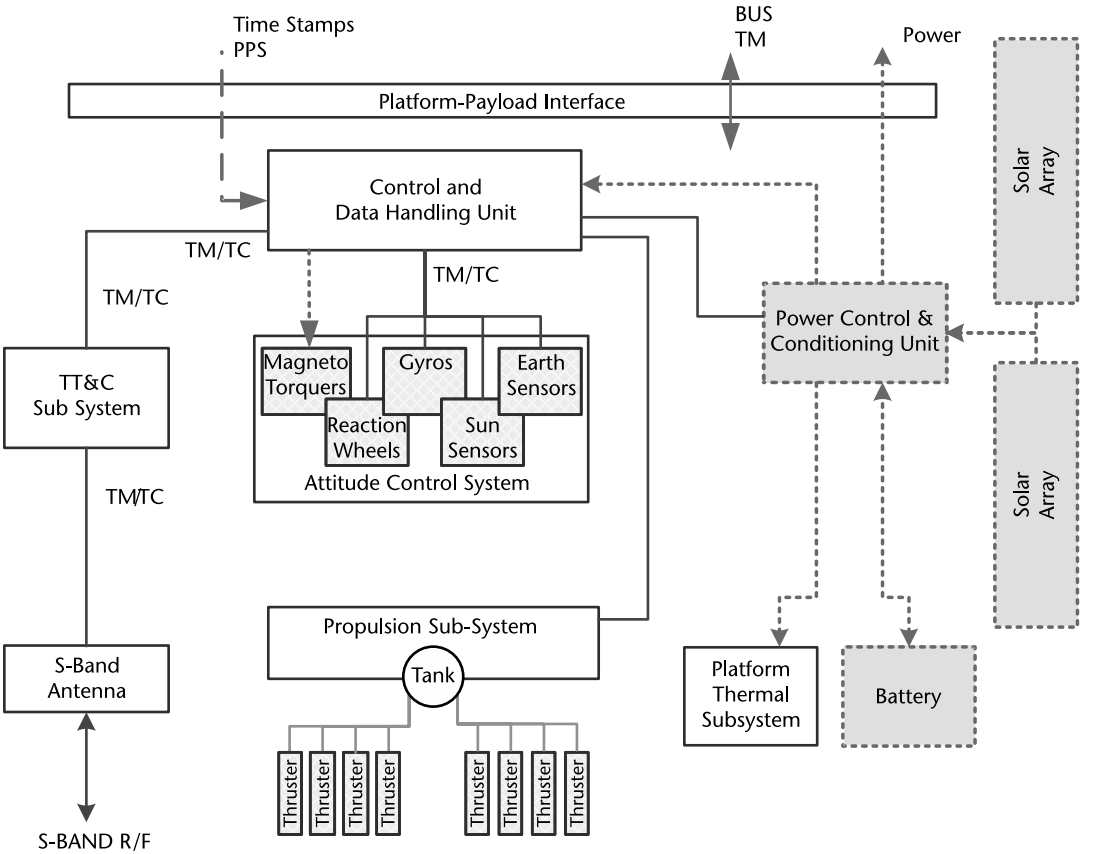


Figure 5.7 Galileo satellite payload main elements.

time-critical data such as the CS data stream are not stored but directly injected in the downlinked navigation data.

The timing subsystem, as the essential element of a navigation satellite, provides the onboard frequency reference. The stability of the onboard timing subsystem and its clocks is essential for the overall performance of a navigation system. The most important part of the timing subsystem is the four atomic clocks: two PHMs and two RAFSs.

The frequency stability of clocks is typically presented using Allan deviation (ADEV) plots that indicate the frequency instability as a function of the sampling period. Figure 5.8 provides ADEV measurement results for the operational master clocks onboard the operational Galileo satellites for the period of October 2015 to January 2016. The different clock technologies can be clearly distinguished with the PHMs, providing better long-term stability.

In order to save power and ease the thermal control of the satellite, only 2 of the 4 clocks onboard are operated at any point in time: one master, typically a PHM, and one RAFS as a backup.

The time provided by the master clock is distributed by the clock monitoring and control unit (CMCU) to elements that need a time or frequency reference. The CMCU provides the link between the timing subsystem and the NSGU. It al-

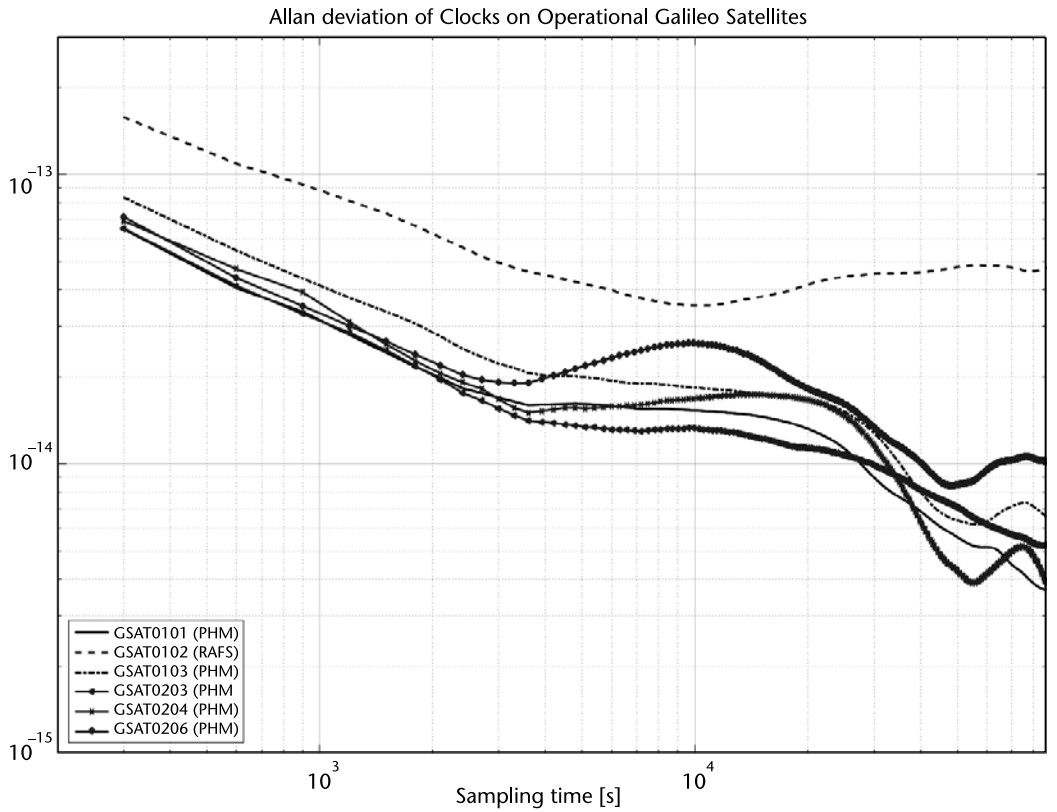


Figure 5.8 Measured clock Allan deviation for selected operational satellites.

lows one to synchronize the master and backup clock to ensure a quasi-seamless transition between the two clocks whenever needed.

The NSGU generates the coherent navigation signals, combining the time information received from the CMCU and the uplinked navigation data mission receiver. The navigation data messages are modulated onto the corresponding navigation signals (E1, E5, and E6) and then broadcast. The Galileo signal and message structure are described in Section 5.5.

In addition to the Navigation payload, there is also a SAR payload onboard the Galileo satellites (except for GSAT0101 and 0102). Details of the SAR mission are discussed in Section 5.7.

5.4.3.3 L3 Satellites

On August 22, 2014, an anomaly in a Soyuz-Fregat upper stage during the launch coasting phase caused Galileo satellites GSAT0201/0202 to be injected into an eccentric orbit preventing the nominal progress of in-orbit operations (Figure 5.9).

The cause of the anomalous injection was identified as a freeze of a propellant line. Appropriate recovery actions were put in place for the manufacturing process of the Fregat upper stage used by Galileo. Immediately after identification of the anomalous injection, analyses had been carried out by ESA in order to

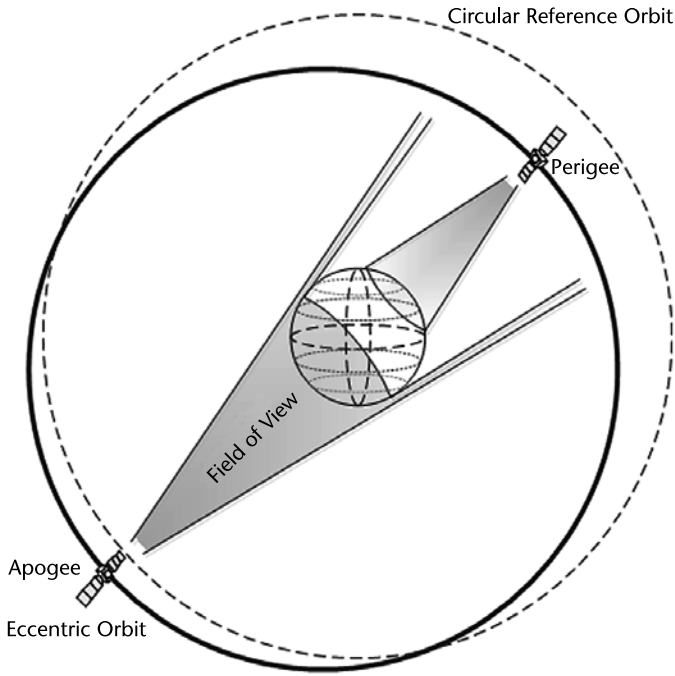


Figure 5.9 GSAT0201/0202 final orbits.

recover GSAT0201/0202. Orbit recovery maneuvers were conducted that led to a nominal mode of operations. The IOT campaign was then carried out. The IOT characterization of these first FOC satellites did confirm the anticipated in-orbit performance of the FOC satellite family [19].

The final orbits of both satellites have an orbit period of 12:56 hours and a ground track repeatability of 37 orbits in 20 sidereal days. The orbit parameters are provided in Table 5.4

After adaptation of the ground segment, these satellites can have a significant contribution to user performance especially during the Galileo deployment phase. Even after completion of the reference constellation deployment, both satellites can provide additional signals that might be useful in an environment with limited visibility. Successful position fixes have been carried out with both satellites demonstrating their usability in user receivers [19].

Table 5.4 Orbit Parameters of GSAT0201 and GSAT0202

	GSAT0201	*GSAT0202*
Semimajor axis	27,979.7079 km	27,978.0244 km
Eccentricity	0.1582	0.1581
Inclination	50.1°	50.1°
RAAN	66.455°	65.432°
Argument of perigee	44.5°	45.6°
True anomaly	331.917°	164.292°
Argument of latitude	16.4°	209.9°

5.4.4 Launchers

The Galileo constellation deployment plan foresees launches with different launch vehicles, Soyuz and Ariane 5. The Galileo satellites were designed with this requirement and are compatible with different launchers. This capability reduces the time needed for the deployment of the whole constellation.

In 2005, ESA and the Russian Federal Space Agency agreed to operate Soyuz/ST launchers from the Guiana Space Centre. Construction of the Soyuz launch pad in Kourou started in 2005 and was completed in April 2011. The first operational launch occurred on October 21, 2011, carrying the first two Galileo IOV satellites to circular MEO. This new launch facility has been used to launch all Galileo IOV and FOC satellites deployed by Soyuz. The Soyuz with its Fregat upper stage is able to directly inject two Galileo satellites into the circular MEO orbit.

After the successful completion of the IOV phase, followed by the deployment and in-orbit testing of the first FOC satellites, the constellation deployment has been accelerated by the Ariane 5 ES launch vehicle which was upgraded and qualified for Galileo. The upgraded A5 ES launches 4 Galileo satellites simultaneously. The modifications encompass a reignitable upper stage that allows for longer coasting phases between the release of the two pairs of satellites.

The Ariane 5 launches began in 2016 and are planned to continue through 2017. Plans call for a total of three Ariane 5 missions to complete the deployment of the 24 satellites into the reference constellation (Walker 24/3/1) slots. Further launches will be necessary to deploy the in-orbit spare satellites [19].

5.5 Galileo Signal Characteristics

This section provides an overview of the Galileo signal characteristics. The Galileo signal design considered a set of high-level rules and guidelines for the selection of carrier frequencies and modulation characteristics. The most important rules and guidelines were the following:

- Provide at least two, preferably three carrier frequencies to support ionospheric delay compensation and multicarrier measurements, and to provide alternative frequencies in case of local interference.
- Overlay with existing SATNAV systems where possible (i.e., use the same carrier frequencies) to allow combined use with minimum technical effort for receivers. Therefore, Galileo signals need to be compatible (i.e., have a low, controlled and coordinated interference into all other ranging signals in the same band).
- The support of combined use (i.e., interoperability with the other SATNAV systems) was desired. This implies, for example, using equivalent or similar modulation principles, to allow reception with the same receiver digital front end. Because of the same center frequency and similar bandwidth, a user receiver can acquire and track the Galileo E1 and GPS L1 signals with one single front end. Although similarities exist in the navigation data message

concepts, separate processes are needed to retrieve each data message from the SIS. Other aspects of interoperability are discussed in Section 5.6.

- Determine the bandwidth of the ranging signals by trading off support of precise and robust tracking and multipath mitigation capabilities against receiver complexity and power consumption. The former necessitates wider bandwidth and flat power spectra while the latter calls for smaller minimum required bandwidth and more easily implementable modulation sequences (e.g., two-level rectangular codes).
- Considering the rapid development of receiver technology, in particular for mass market and power efficient receivers, E1 was chosen with a slightly larger minimum bandwidth than the legacy GPS C/A signals.
- Services with special protection needs are to be separated from public services. This led to the difference in modulation parameters between the CS, OS, and PRS.
- ITU regulations for frequency ranges available for RNSS use and for the protection of other users of the foreseen Galileo bands were also to be considered.

Compatibility and interoperability of Galileo and GPS were ensured through cooperation between Europe and the United States on the definition of the Galileo E1 OS and the GPS L1C modulation parameters as well as power levels. The result was the EU-US agreement in 2004 on Galileo E1 and E5 and GPS L1C and L5 signals, respectively. Since then, similar spectrum coordination agreements have taken place with other SATNAV systems.

The Galileo satellites broadcast coherent navigation signals on three frequencies in L-band: E1, E5, and E6. Galileo provides, in the E1 frequency band centered at 1,575.42 MHz, three signal components, referred to as E1-A, E1-B, and E1-C. The E1-A signal component carries the Galileo PRS. The E1-B and E1-C components form the data and the pilot components, respectively, of the Galileo E1 OS. The signal plan is provided in Figure 5.10.

At a carrier frequency of 1,278.75 MHz, Galileo is emitting its E6 signals that include the signal components E6-A, E6-B, and E6-C. Similar to the E1 frequency band, in E6 the E6-A component refers to the Galileo PRS service. The components E6-B and E6-C are the data and the pilot components, respectively, of the Galileo CS.

The Galileo E5 signal centered at 1,191.795 MHz consists of 2 individual signals, the Galileo E5a and the Galileo E5b signal. The E5a data and pilot components are centred at 15.345 MHz below the E5 carrier frequency (1,176.45 MHz) and the E5b data and pilot component are centered at 15.345 MHz above the E5 carrier (1,207.14 MHz). Both E5a and E5b can be tracked individually as if they were modulated on separate carrier frequencies in the E5 band. The E5 signal (E5a and E5b) can also be tracked as one signal with a very large receiver bandwidth of at least 51.15 MHz allowing for better multipath rejection and Gabor bandwidth. This processing is possible because the E5 carrier is generated coherently using the AltBOC modulation scheme. In Galileo, the power split between the data and the pilot component of any signal is 50/50% [12].

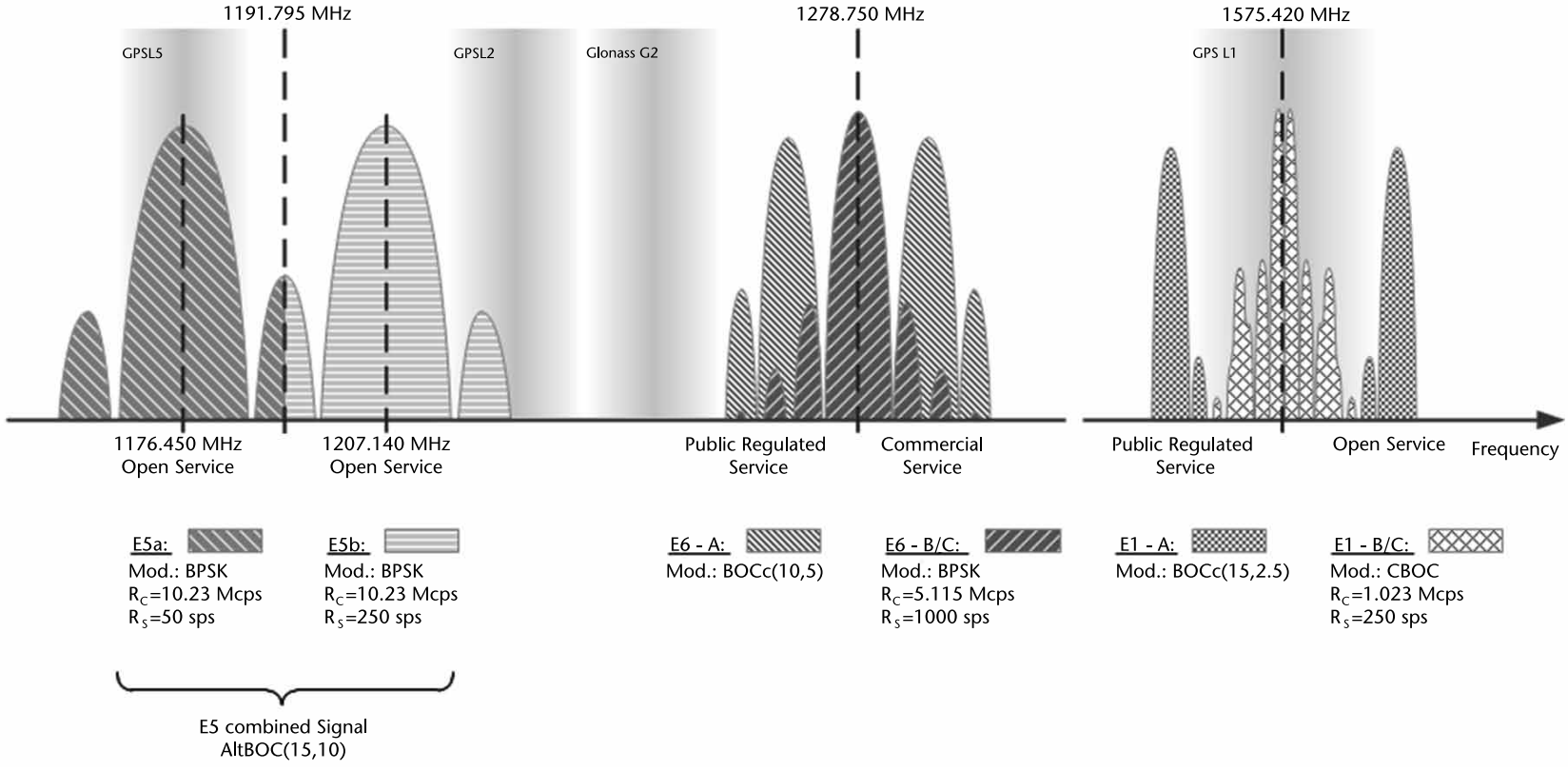


Figure 5.10 Galileo signal plan.

The conventional BPSK-R modulation on the Galileo E6 CS signal and the sine- and cosine-phased BOC modulations of the other Galileo signals are described in Section 2.4.

All Galileo signals are coherently derived from the same onboard master clock. The characteristics of the publicly accessible Galileo signals are summarized in Table 5.5, including their assignment to Galileo services.

The composite binary offset carrier (CBOC) modulation used for the Galileo E1 OS signal is based on a 1.023-Mcps spreading sequence with a two-component spreading symbol. The spreading symbol consists of the additive combination of one BOC(1,1) and one BOC(6,1) subcarrier. The BOC(6,1) subcarrier has 1/11th of the total CBOC power.

The E1 data channel (E1-B) uses the in-phase combination of the two components and the pilot channel (E1-C) uses the anti-phase combination. The resulting signal has four levels and features an alternation between the BOC(1,1) and BOC(6,1) in the time domain, due to the inversion of the phase of the BOC(6,1) component between data and pilot. The resulting signal is referred to as CBOC(6,1,1/11). The CBOC(6,1,1/11) is a particular realization of the multiplexed BOC MBOC(6,1,1/11) that was agreed as the jointly defined interoperable signal for GPS L1C and Galileo E1 OS. An alternative realization of the MBOC(6,1,1/11) is the TMBOC(6,1,1/11) selected for GPS L1C, which is based on time-multiplex-

Table 5.5 Overview of Galileo Public Signal Components and Modulations

Signal	*E1*		*E6*		*E5b*		*E5a*	
Frequency (MHz)	1,575.42		1,278.75		1,207.14		1,176.45	
Service	Open		Commercial		Open		Open	
Component	E1-B Data	E1-C Pilot	E6-B Data	E6-C Pilot	E5b-I Data	E5b-Q Pilot	E5a-I Data	E5b-q Pilot
Rx reference bandwidth (MHz)	24.552	24.552	40.92	40.92	20.46	20.46	20.46	20.46
Message type	I/NAV	—	C/NAV	—	I/NAV	—	F/NAV	—
Modulation	CBOC 1/11	CBOC 1/11	BPSK-R	BPSK-R	BPSK-R	BPSK-R	BPSK-R	BPSK-R
Primary code chip rate (multiples of 1.023 MHz)	1	1	*5*	*5*	10	10	10	10
Subcarrier frequency (multiples of 1.023 MHz)	1&6	1&6	—	—	—	—	—	—
Data rate (symbol/sec)	250	N/A	1,000	N/A	250	N/A	50	N/A
Secondary code chip rate (chips/sec)					1,000		1,000	
Primary code length (chips)	4,092	4,092	5,115	5,115	10,230	10,230	10,230	10,230
Secondary code length (chips)		25		100	20	100	4	100
Carrier phase	0°	0°	0°	0°	0°	90°	0°	90°
Min. Rx power (dBW)	–160	–160	–158	–158	–158	–158	–158	–158

ing the BOC(1,1) and the BOC(6,1) waveforms. The resulting power spectra of both CBOC(6,1,1/11) and TMBOC(6,1,1/11) are identical.

For the tracking of Galileo E1 OS, the equivalent of a four-level correlator with amplitude levels of –1.25, –0.65, +0.65, and +1.25 is necessary to take full benefit of the CBOC. Use of a conventional two-level BOC(1,1) replica for tracking is possible, but at the cost of not using all the energy of the BOC(6,1) component, representing a loss of approximately 0.4 dB before any receiver-dependent losses.

Several advanced techniques have been suggested demonstrating the possibility for implementing efficient CBOC tracking [21–23].

The Galileo E5 signal is generated using a wideband complex sideband modulation, referred to as alternative BOC (AltBOC) [24, 25]. The baseband representation of AltBOC corresponds to the sum signal of two coherently generated and individually quadrature modulated complex subcarriers, the upper E5b and the lower E5a, and of an intermodulation function to achieve constant transmit envelope [26]. The two subcarriers, before band limitation, are discrete multilevel signals with period $T_s = (15.345\ \text{MHz})^{-1}$. The ideal wideband description provided in [12] results in a signal constellation diagram representing an 8 PSK-type modulation. The main energy content of the intermodulation function is located outside the recommended AltBOC receive bandwidth of 51.2 MHz, around ±46-MHz offset. A concept to exploit the AltBOC by generating an AltBOC replica using a lookup table as part of the receiver implementation is provided in [12], together with the direct mathematical description.

Galileo does not provide a specific navigation message support for the combined use of E5a and E5b as one AltBOC. Instead, each subcarrier provides a different service-specific navigation message. E5a provides the OS-related F/NAV messages and E5b provides the I/NAV messages, in the past specifically envisaged for SOL users. The ephemerides information provided in I/NAV and F/NAV are equivalent and interchangeable; however, the clock corrections provided in the two message types are not necessarily identical, and may differ slightly. This is because each message is individually generated for the specific frequency pair. That is, F/NAV provides clock corrections for the E1/E5a frequency pair and I/NAV for the E1/E5b pair.

Each data and pilot component of the E5a and E5b sideband signals can be individually acquired and tracked as conventional BPSK-R(10) type navigation signals. For this purpose, a receive bandwidth centred on the sideband frequency is recommended. The typical bandwidth for such sideband tracking may be 20.46 MHz, the main lobe of the BPSK-R(10) modulation. Using larger receiver bandwidths will require a trade-off between the desired improvement of tracking accuracy and the increasing crosstalk from the other sideband due to the coherency of the sidebands and unavoidable related cross-correlation imperfections. A larger bandwidth in E5 may also imply an increased susceptibility to interference from other primary users of this frequency range, especially from aeronautical systems like Distance Measurement Equipment and Tactical Air Navigation systems.

A full description of the public Galileo signals as part of the OS and CS and their modulations is published in the Galileo Public Open Service Signal in Space Interface Control Document (OS SIS ICD) [12].

5.5.1 Galileo Spreading Codes and Sequences

The ranging signals transmitted by the Galileo satellites are individually spread in the frequency domain. The periodic spreading codes are unique for each signal component and also different for each satellite. The spreading code length of each signal data component (E5a-I, E5b-I and E1-B) is chosen to cover full symbols. For the data channels that would require a code length longer than 10,230 chips and for the pilot channels (E5a-Q, E5b-Q, and E1-C), the spreading codes are generated by a tiered code construction. The pilot primary code has the same length as the primary code used for the equivalent data component. The long spreading codes are constructed by sequentially XORing each consecutive primary code period with the next chip of the secondary code. The tiered code construction has been selected to limit the primary codes length to reduce the acquisition time by limiting the search space and also to provide a nonrepetitive sequence length of one symbol for data components and 100 ms for the pilot components.

The primary spreading sequences were optimized for orthogonality across the whole family to ensure sufficient isolation between the signal sources. The secondary codes are adjusted for low autocorrelation side lobes and a flat spectrum amplitude in the frequency domain.

Receivers coherently integrating over the full length of the tiered code (or over multiples of the primary code length) will observe additional correlation peaks at integer intervals of the primary code aside of the main correlation peak. The amplitude of these additional peaks is lower than the amplitude of the main correlation peak, as their autocorrelation amplitude is a function of the secondary code used to extend the primary code. This feature of the tiered code generation allows the receiver to determine code phase relative to GST with an ambiguity of 100 ms. This enables time-free position solutions for users on Earth relying on code phase measurements including the secondary code of the pilot signal provided the user has ephemerides and clock correction information available and the receiver clock misalignment is below the 100-ms ambiguity.

Another feature of the two-tiered codes is the possibility to adapt the coherent integration time in multiples of the primary code length at the receiver level taking advantage of the knowledge of the secondary code. The Galileo primary and secondary spreading codes for public use are provided in the OS SIS ICD [12].

5.5.2 Navigation Message Structure

The public signals of Galileo provide three different navigation messages:

1. F/NAV or Free Navigation message: This navigation message is providing data for the usage of the E5a signals. The message is provided on E5a-I.
2. I/NAV or Integrity Navigation message: This message has originally been envisaged to provide the safety of life-related data and alerts with a high data rate message with short page length. Since the reprofiling of the SOL service, the I/NAV provides OS data on the E1-B and E5b-I signals.
3. C/NAV or Commercial Navigation Message: This is the message that provides the data generated by the commercial service provider. It is a fast message on E6-B. (See Table 5.5.)

Both F/NAV and I/NAV messages provide equivalent and partially identical navigation-related data. The orbit information and the parameters to convert from GST to UTC and from GST to GPS time are compatible between I/NAV and F/NAV. The clock-related information is specific for each message because it is derived from measurements of E1/E5a for F/NAV and E1/E5b for I/NAV.

The utilization of broadcast information such as ephemeris, clock corrections, GST-UTC parameters, almanacs, and the related usage algorithms are in line with the GPS definitions; only minor adaptations have been made for Galileo.

The ephemeris contains information that enables the receiver to compute the satellite's orbital position at the time of transmission or more accurately the coordinates of the satellite's common apparent L-band antenna phase center in the ECEF coordinate system.

The satellite-specific realisation of GST is provided by the sequential Week Number counting from the origin of the GST and the Time of Week. The Week Number rolls over after 4,096 weeks, or approximately 78 years. The Time of Week provides the number of seconds elapsed since the start of each week, defined by the leading edge of the first chip of the first code sequence of the first page symbol. It covers one entire week (604,799 seconds) and it is reset to zero at the end of each week (00:00:00 between Saturday and Sunday).

The satellite clock correction parameter broadcast enables the user to compute the time of transmission of the satellite's signal in absolute GST. These satellite clock corrections are provided specifically for the signals or their combinations in F/NAV for E1 and E5a and in I/NAV for E1 and E5b. Because I/NAV and F/NAV support different dual frequency combinations, the clock corrections can differ between both messages, although typically they are expected to be very similar.

The ephemeris and clock corrections on F/NAV and I/NAV are computed based on corresponding dual frequency observations and can be directly applied by dual frequency receivers. However, single-frequency receivers have to apply an additional message parameter to the clock correction, the Broadcast Group Delay (BGD) correction. The BGD is used by single frequency users to correctly determine the satellite time of transmission in GST by correcting for the payload group delays. See Section 5.8.2.3 for details.

Single-frequency users also need to correct the effect of ionosphere on the signals to derive accurate pseudorange measurement. Galileo provides correction parameters for an adapted version of the NeQuick model, which can be utilized by single frequency user receivers as described in [27]. Sections 5.8.2.2 and 10.2.4.1 discuss the NeQuick model.

The UTC conversion parameters enable transfer of GST to UTC. For this purpose, the navigation messages contain information on the UTC offset, first-order term of the polynomial, and number of leap seconds. The navigation messages also contain a notification for the next leap second adjustment [12].

The GGTO parameter provides the offset and rate of change between GPS system time and GST. It is computed by the Galileo PTF and coordinated with the USNO.

Navigation data are complemented by service parameters such as the satellite ID, the satellite health status and navigation data validity flags, the checksum for the CRC, and the IOD. The IOD allows one to identify to which batch a specific parameter belongs. Two IODs are distinguished in the navigation messages.

The IODnav is provided for the ephemeris, satellite clock correction, and Signal in Space Accuracy parameter. The IODa is defined for the almanac.

Moreover, each satellite broadcasts an almanac that provides ephemeris and clock correction data for all operational satellites of the Galileo constellation with a reduced precision. The almanac information aids the receiver's satellite acquisition process.

The Navigation data stream transmitted from each satellite is formatted such that the data essential for providing the navigation mission are broadcast more often and therefore are received by the user within a well-defined maximum time. Other complementary data that are less time-critical and not or only partially relevant for navigation are broadcast over longer periods, for instance, the almanac information.

The complete set of navigation data are transmitted on the data component of the corresponding signal in the form of a sequence of frames. Each frame consists of subframes, which, in turn, are made of several pages. A page is the basic structure of the navigation message. This structure has been selected to transmit data serving different needs. Time critical data are repeated with a high rate such as SAR RLS data. Data with lower priority such as those supporting acquisition in warm start conditions are repeated with a medium rate while yet other data are provided with a low rate. The pages can be distinguished by means of a type identifier. The sequence of the messages within a subframe or frame, although typically followed, might in the future be altered to meet future requirements. User receivers should be able to identify the page contents based on the type identifier and should be able to cope with variations in the sequence of pages as well as new page types that might be introduced to support future service evolutions.

The F/NAV page has a length of 238 bits or 10 seconds excluding the synchronization pattern and the tail bits. The F/NAV frame of 600 seconds is composed of 12 subframes with each 5 F/NAV pages.

The I/NAV frame of 720 seconds is composed of 24 subframes with 15 pages each. The I/NAV transmission relies on the frequency diversity and provides the same page layout on E1-B and E5b-I. The pages are broadcast in two consecutive blocks of odd and even words, respectively. Each word starts with the I/NAV synchronization symbols followed by a block-encoded data field, and lasts one second. A full I/NAV page (odd and even words combined) takes 2 seconds for transmission and provides a useable capacity of 245 bits, excluding the synchronization pattern and tail bits. The page sequencing on E1 and E5 are different and the pages are swapped between the two signals to allow a faster reception of the complete I/NAV data set in dual-frequency receivers. The design of I/NAV also allows for single frequency usage at the cost of longer delays until the complete message is received.

The system functions related to I/NAV allow for the introduction of short one-time low latency message pages by replacing nominal transmissions in the message dissemination; such short message pages might be used in the future.

The OS SIS ICD [12] provides a detailed description of the contents of the F/NAV and I/NAV. It is important to highlight that this ICD provides the information with reservations regarding future message evolutions. Backward compatibility will be maintained, but new message features and new message types might be introduced, exploiting the existing degrees of freedom and spare capacity.

5.5.2.1 Commercial Service Data Stream

C/NAV is a near-real-time message stream with short latency. The data content of the C/NAV will depend on external CS provider needs. The CS and its applications are still under development and consolidation at the time of this writing. Therefore, no detailed description of the C/NAV content has been published yet. The C/NAV data, like all the low-latency data channels, are provided only from satellites that are in contact with the ground segment. This will allow for different C/NAV contents on different satellites.

5.5.3 Forward Error Correction Coding and Block Interleaving

The forward error correction (FEC) protecting the Galileo data components and improving the transmission robustness of the message relies on convolutional coding with a rate of ½ and a constraint length of 7. The encoder polynomials are chosen to be identical with the GPS L5 CNAV encoder. Galileo applies one additional inversion to the output of the G2 polynomial to achieve a nonconstant symbol output despite of a continuous input of zeros. The encoding comprises nonoverlapping, independent data blocks of full or half pages of the navigation message. Each FEC encoded block is interleaved using an interleaver with 8 rows and "n" columns ("n" depends on the page size in symbols, different for F/NAV and I/NAV).

The combination of convolutional and the block-wise encoding concept requires the introduction of predefined tail bits to provide FEC protection for the complete information content of each navigation page. This ensures that burst errors are de-interleaved with at least 8 symbols of separation between individual symbol errors at the decoder input, which helps the FEC decoder to correct such errors.

5.6 Interoperability

According to the International Committee on GNSS (ICG) "Interoperability refers to the ability of global and regional navigation satellite systems and augmentations and the services they provide to be used together to provide better capabilities at the user level than would be achieved by relying solely on the open signals of one system." Galileo and GPS were the first SATNAV systems pursuing and taking the necessary implementation steps to achieve interoperability between both systems.

Interoperability at the user segment allows usage of multiple SATNAV systems to achieve a combined position solution. The combined utilization of multiple systems will lead to higher accuracy or better availability, for instance, in challenging environments, than each individual system would be able to provide.

Interoperability at the system level addresses the RF signal structure as well as the alignment of time and geodetic reference systems and their realization in reference frames between the participating SATNAV systems. Depending on the alignment of the before mentioned elements, different levels of interoperability can be reached, leading in the final state to full interchangeability.

The level of interoperability is a result of an optimization process. Factors to be considered include radio frequency compatibility, complexity of the user

equipment, market prospects, vulnerability (common mode of failures), independence of the systems, and national security compatibility issues.

As presented above, on the signal level, the usage of common carrier frequencies for the OS signals (E5a and E1) eases the RF front-end design and supports interoperability. Recent receiver developments already demonstrated that small carrier offsets (e.g., between GPS L1/Galileo E1 and GLONASS G1) are not hampering interoperability at the user level. Beyond the signal characteristics, the underlying navigation message concepts are comparable between Galileo and GPS (e.g., ephemeris, almanac, clock correction, GST-UTC, BGD). The terrestrial reference frames and reference time systems are aligned, as indicated in the following sections of this chapter.

In order to allow an increased level of interoperability, Galileo is broadcasting the GGTO in its navigation messages. The GGTO allows the user to estimate his or her PVT based on an ensemble of GPS and Galileo range measurements without the need to sacrifice one observable to resolve for the mutual system time offset between GPS and Galileo. This is of particular interest for users in constrained visibility environments [28]. Section 11.2.5 covers use of multiple constellation signals to form the PVT solution.

5.6.1 Galileo Terrestrial Reference Frame

The Galileo Terrestrial Reference Frame (GTRF) is an independent realization of the ITRS. The ITRS is defined and monitored by the Central Bureau of the IERS. The GTRF is designed to be compatible with the ITRF and will therefore be a realisation of the ITRS. The GTRF station coordinates are aligned to the ITRF station coordinates within a tolerance of 3 cm (at 95% confidence level) for all stations used in both frame realizations.

WGS 84 is the coordinate reference frame for GPS. WGS 84 is also a realization of the ITRS. The WGS 84 includes the coordinates of GPS USAF monitoring stations and those of the U.S. NGA monitoring stations. The differences between WGS 84 and the GTRF are expected to be on the order of a few centimeters. This accuracy is sufficient for navigation and many other user needs.

5.6.2 Time Reference Frame

GST as generated by the PTF is a continuous time system that has no leap seconds. The reference epoch for GST is defined at 00:00:00 UTC on Sunday, August 22, 1999, midnight between August 21 and 22. The GST initial epoch coincides with the last rollover of the GPS week number [12].

The time steering of the GST to UTC is carried out with time transfer measurements between the master clocks in the Galileo PTFs and the GTSP which provides the link to UTC (k). The time transfer measurements are processed by the GTSP to predict the retroactively published TAI timescale to the current time. The GTSP determines the deviations between the Galileo master clock timescale and the predicted TAI. Based on these deviations, the necessary steering corrections for the Galileo PTF master clocks are generated. These data are provided to the GCCs on a daily basis.

GST will be kept to within 50 ns (95%) of TAI over any 1-year time interval. The offset between TAI and GST will be known with a maximum uncertainty of 28 ns (2 sigma), assuming the estimation of TAI 6 weeks in advance. Users equipped with a Galileo timing receiver will be able to predict UTC to 30 ns for 95% of any 24 hours of operation. The difference between GST and UTC at the initial epoch was 13 leap seconds. At the time of this writing, the difference between GST and UTC was 18 leap seconds.

The GGTO parameter is determined by the PTF on a daily basis and allows users to generate combined position solutions even in situations where only a total of four satellites are received from both systems together. The GGTO is determined by means of a combined GPS/Galileo receiver at the Galileo PTFs. For robustness, traditional time transfer techniques are also used to determine the offset between GPS system time and Galileo system time. For this purpose, the time transfer between the Galileo PTF and the USNO is performed using TWSTFT and the Common View technique [29].

The GGTO is coordinated between Galileo PTF and USNO before it is included in the navigation message of both systems. The accuracy of this time offset modulo 1 second is specified to be less than 5 ns with 2-sigma confidence interval over any 24-hour period.

When more satellites are received, user receivers can also resolve the Galileo/GPS time offset as part of the position and navigation processing at the cost of one additional satellite tracked (fifth satellite when determining a three-dimensional position). See Section 11.2.5 for additional details.

In the context of the ICG, specifically in the Providers Forum, discussions are ongoing to provide the offset of each SATNAV system time scale to a common reference, such as UTC. This common approach will allow user receivers to exploit measurements from multiple GNSS constellations for mixed positioning and timing services.

5.7 Galileo Search and Rescue Mission

This section gives a general overview of the SAR system deployed by Galileo. The SAR/Galileo system is designed and operated as an integral part of the Cospas-Sarsat MEOSAR system. The SAR/Galileo space and ground segments are fully integrated within the Cospas-Sarsat structure and represent the contribution of Galileo to the International Cospas-Sarsat Program. (Additional information on the Cospas-Sarsat MEOSAR system can be found in [30, 31].)

Starting in 1982 with LEO satellites, the Cospas-Sarsat program was extended to include geostationary satellites. With the advent of MEO-based GNSS, it was recognized that there is a strong and multilateral synergy between navigation and SAR functions. Eventually, three planned GNSS core constellations: Galileo, GPS, and GLONASS, announced the future hosting of SAR equipment on their satellites, leading to the concept of MEOSAR (Medium Earth Orbit SAR component). These three SATNAV systems have been in close coordination, under the umbrella of Cospas-Sarsat, to embark SAR equipment that would be fully compatible and highly interoperable both in the space and ground segments. It is important to note that the GPS SAR payload is planned for GPS III capability insertion. A new

service, the SAR/Galileo RLS, is currently introduced to MEOSAR by Galileo. It provides a means for delivering short messages to emergency beacons equipped with Galileo receivers.

5.7.1 SAR/Galileo Service Description

The Council of The EU of December 2004 confirmed the SAR service as one of the Galileo services. Consequently, the SAR/Galileo mission was defined as follows: "The Galileo system shall provide a Search and Rescue (SAR) service by performing the detection and localisation of current and foreseen Cospas-Sarsat (C/S) 406 MHz beacons, fitted or not fitted with GNSS receivers or other means of position determination, and by providing a return link capability to distress beacons."

The Galileo SAR Service consists of two main services: the FLS, which is the classical satellite-based SAR mission, and the RLS, which delivers a short message to the beacon with additional information.

For the Forward Link Alert Service, Galileo satellites (and other MEOSAR) receive signals from C/S 406-MHz emergency beacons and rebroadcast them in L-band to MEOLUTs. The MEOLUTs determine and report the beacon's alert message and position to Cospas-Sarsat operators, that is, national mission control centers (MCC).

The FLS is a global service; hence, the C/S 406-MHz beacons can be located anywhere on Earth's surface. SAR/Galileo contributes to the global MEOSAR alert service by providing, in addition to the space segment, ground segment elements for detection and localization of distress alerts in Europe.

SAR/Galileo does not include dedicated MCCs or Rescue Coordination Centres (RCCs); these essential components of the overall SAR service remain national member states' prerogatives and responsibility.

The RLS provides users with an acknowledgment message informing them that the alert has been detected and in some cases that rescue operations are under way. The RLS is a specific service introduced uniquely by Galileo to the SAR community to provide messaging capability from the SAR operational facilities directly to the beacon. It is a global service available to any beacon with an RLS-enabled Galileo receiver. The RLM is sent to the beacon through the Galileo E1B signal (1,575.42 MHz).

5.7.2 European SAR/Galileo Coverage and MEOSAR Context

The Galileo Program with its SAR ground infrastructure contributes to the Cospas-Sarsat MEOSAR system, primarily by ensuring the coverage of the European SAR Coverage Area (ECA) which is depicted in Figure 5.11 and defined as: "… the coverage of the European territories of all EU and ESA member countries (i.e. EU countries plus Norway and Switzerland) and the associated maritime and aeronautical Search and Rescue areas adjacent and belonging to these countries."

In December 2016 the European Commission declared the start of the initial Galileo Search and Rescue service. The Galileo SAR SDD defines a simplified coverage area delimited by four corners (85.00°N, 41.20°E; 29.18°N, 37.07°E; 5.00°N, 38.00°W; 75.76°N, 77.87°W) connected by arcs of great circles [9]. Figure 5.12 shows the SAR/Galileo system architecture and identifies essential components in

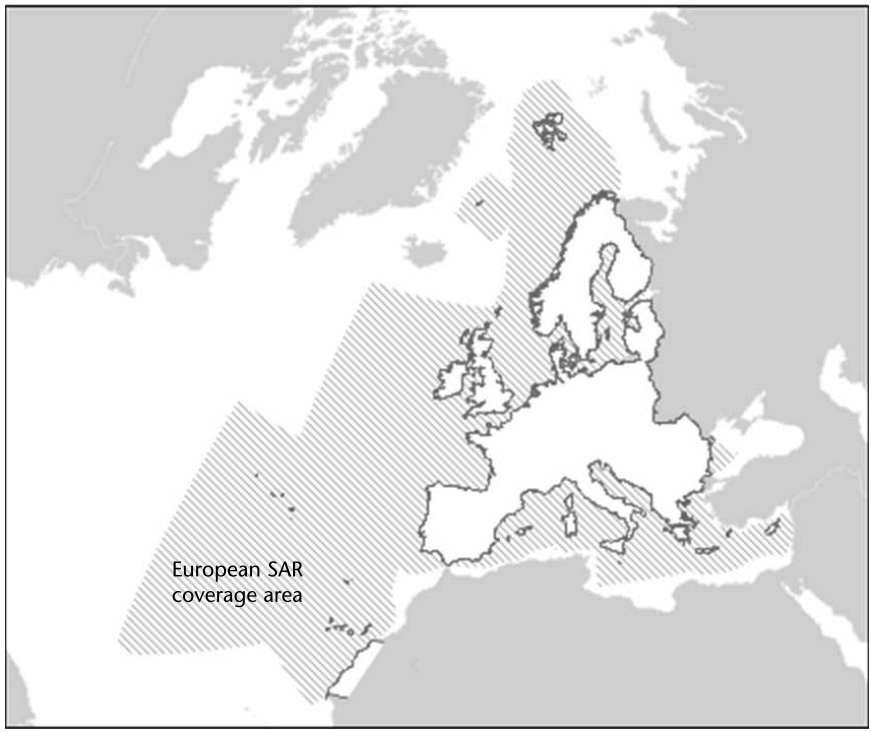


Figure 5.11 European SAR coverage area.

the integration with other Cospas-Sarsat participants. The global coverage, including coverage of European overseas territories, will be achieved through cooperation among Cospas-Sarsat members.

The EU will, in the future, further contribute to meeting the Cospas-Sarsat objective of global MEOSAR coverage by deploying additional MEOLUTs. These will be in identified coverage gaps in coordination with Cospas-Sarsat, in particular for EU member states' overseas zones of responsibility.

5.7.3 Overall SAR/Galileo System Architecture

As mentioned earlier in the section, Galileo equips its satellites with SAR repeaters that relay beacon distress signals to Earth. Relayed signals from the different satellites are received by one or several MEOLUTs. The role of the MEOLUTs is to detect the beacon alert signal, demodulate it, extract the message, and determine the location of the beacon. The alert is located by extracting the encoded positional data in the beacon message, if available, and by processing the differences in the times of arrival (ToA) of the alerts influenced by the beacon-to-satellite range and in the frequencies of arrival (FoA) influenced by Doppler shifts of the received signals. The MEOLUT then sends the estimated beacon position, the message and other relevant data to the associated MCC, which communicates with the relevant MCCs and RCCs, and via the French nodal MCC (Toulouse) with the RLS provider (RLSP) in case the alert contains an RLM request.

In order to provide the FLS, the SAR/Galileo system receives and processes distress signals from beacons by three Galileo MEOLUTs, which are coordinated by

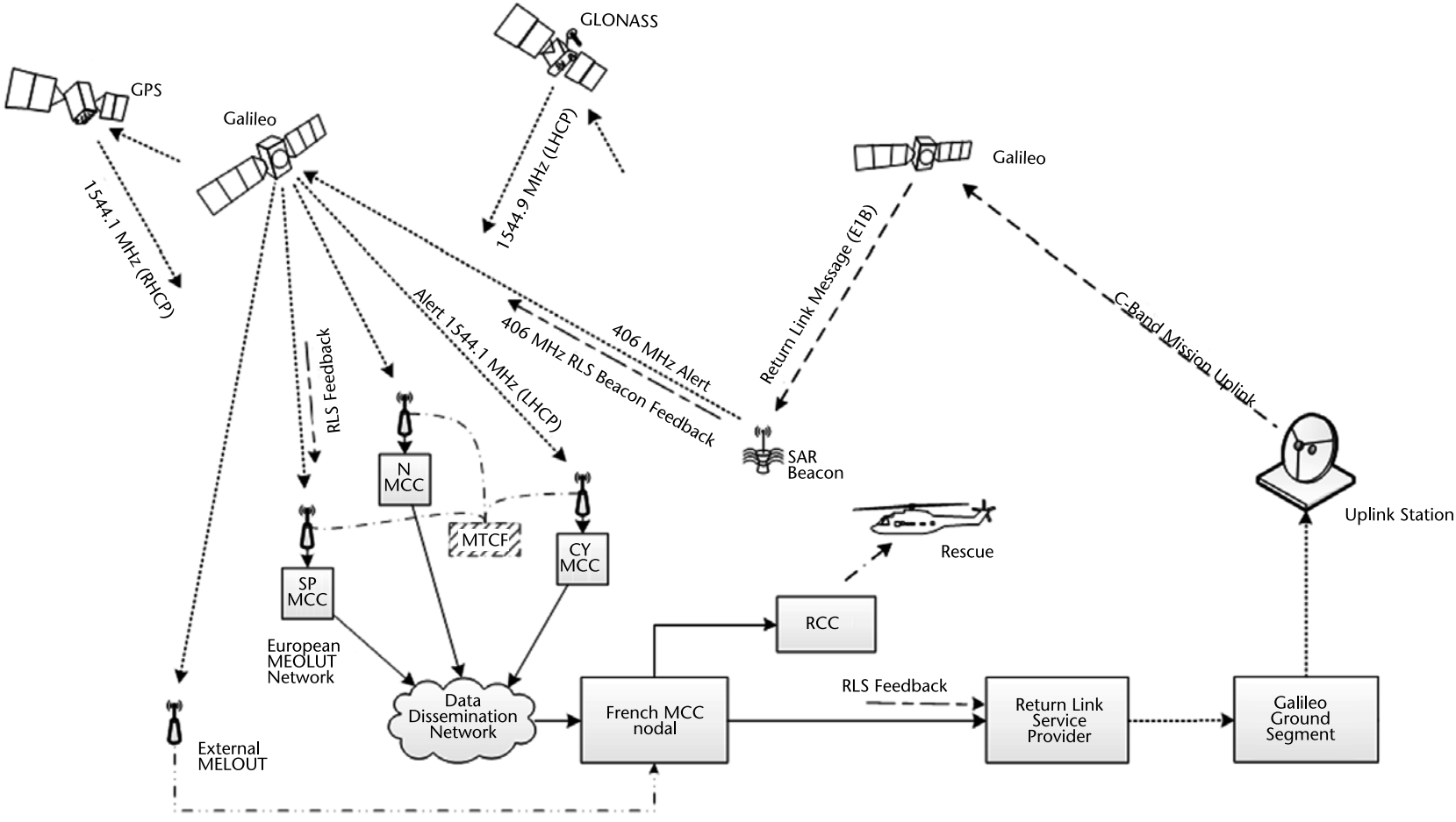


Figure 5.12 SAR/Galileo system and its MEOSAR context.

the MEOLUT Tracking Coordination Facility (MTCF). The MTCF will optimize the tracking of the three European SAR/Galileo MEOLUTs to improve the overall SAR FLS performance over the ECA region.

The space segment receives 406-MHz distress signals from SAR beacons, amplifies and translates these signals in frequency without spectral inversion, and retransmits them to the ground in the L-band (see Figure 5.13).

In order to provide the RLS, the SAR/Galileo system receives, via the RLSP, the RLM requests from the French MCC. The RLSP will identify the best satellite for the RLS broadcast based on the beacon location and additional information exchanged with the Galileo system. The Galileo ground segment incorporates the received RLMs into the navigation data of the identified Galileo satellites, which broadcast the RLM to emergency beacons within the Galileo OS E1 signal, thus closing the loop with users being notified of the detection of their alert.

The overall SAR/Galileo detection and localisation performance is continuously monitored based on the emissions of 5 MEOSAR Reference Beacons (RefBe), located within the ECA.

5.7.3.1 SAR/Galileo Space Segment

The SAR/Galileo space segment comprises Galileo satellites with SAR repeaters. The Galileo satellite payload has two principal functional elements relevant to SAR: the Navigation function and the SAR function. SAR/Galileo utilizes both of these functional elements, with the SAR function for supporting the FLS and the Navigation function for supporting the RLS.

The Galileo SAR repeaters comprise bent-pipe transparent transponders with no frequency inversion. They receive signals at the 406-MHz band and retransmit in the L6 band at 1.5441 GHz. They are designed according to the space segment interoperability requirements agreed under the Cospas-Sarsat auspices, including both the normal (90-kHz) and narrow (50-kHz) bandwidth modes, as well as the possibility to operate with fixed gain mode or automatic level control.

All space segment components of MEOSAR, that is, Galileo, GPS (planned for GPS III capability insertion), and GLONASS SAR repeaters are mutually compatible and interoperable.

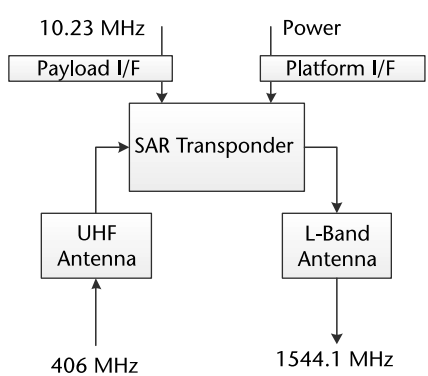


Figure 5.13 Galileo satellite SAR payload schematic.

5.7.3.2 SAR/Galileo Ground Segment

The SAR/Galileo ground segment (SGS) comprises forward and return link components. The FLS SGS consists of three operational MEOLUTs, the MTCF and the related SAR communications network (SARN), as well as five dedicated RefBe. The RLS part of the SGS includes the RLSP.

The main center for SAR/Galileo is the SAR/Galileo Service Centre in Toulouse, which is closely associated to the French MCC. The center is hosting the MTCF and the RLSP as key elements of the SGS. The three operational MEOLUTs are located close to the three corners of the ECA region in:

- Spitsbergen (Svalbard/Norway), hosting the Spitsbergen/EU MEOLUT collocated with the existing Norwegian MCC and LEOLUT;
- Maspalomas (Canary Islands/Spain), hosting the Maspalomas/EU MEOLUT collocated with the Spanish nodal MCC and other Cospas-Sarsat facilities including LEO and GEOLUTs;
- Larnaca (Cyprus), hosting the Larnaca/EU MEOLUT connected to the Cypriot MCC located at the Joint RCC in Larnaca.

The five reference beacons are located at:

- Spitsbergen/EU reference beacon collocated with the MEOLUT;
- Maspalomas/EU reference beacon collocated with the MEOLUT;
- Larnaca/EU reference beacon collocated with the MEOLUT;
- Santa Maria/EU reference beacon located on Azores Islands/Portugal;
- Toulouse/EU reference beacon located at the SAR/Galileo Service Centre.

The prime contractor for the SAR/Galileo Ground Segment has been Cap Gemini (France).

The three Galileo MEOLUTs interface to their corresponding national MCCs. The MTCF coordinates the tracking of visible satellites performed by the European MEOLUTs. The nominal SGS operational configuration of the FLS is inherently redundant and exhibits graceful degradation with failures. The system is able to also operate without the coordination of the MTCF, but the performance and reliability of the full system is significantly improved when exploiting its advanced features. These include not only the coordinated MEOLUT tracking but also the sharing of collected raw TOA/FOA data.

The RLS is typically initiated by a beacon through a Return Link Message Request. This request is part of a particular protocol for the beacon's forward link alert message. The specific RLS protocol on the 406-MHz uplink signal is routed to the RLSP. The RLM request is received at the RLSP through the Cospas-Sarsat network. The RLM delivery can also be triggered externally by the RLSP operator or by authenticated third parties interfacing with the RLSP.

The part of the RLS infrastructure that is under direct Galileo responsibility comprises the RLSP, the GMS, and Galileo satellites. The full RLS loop (with beacon feedback) consists of the following events:

- RLS beacon (operating with the RLS location protocol) issues a distress alert containing a RLM request indicating that it can accept a return link message as acknowledgement Type-1.
- At least one MEOLUT receives the alert through MEOSAR satellites and routes it through the Cospas-Sarsat data distribution network to the French MCC.
- The French MCC forwards the RLM request to the RLSP.
- The RLSP determines the appropriate time and the satellites through which the RLM shall be broadcast and passes this information including the RLM request to the GMS.
- GMS uplinks the RLM to the selected satellites.
- The originating RLS beacon receives the acknowledgement RLM via the E1B data.
- Having received the acknowledgment, dedicated bits in the forward link alert message are changed to indicate that an acknowledgement has been received.
- Following the same path: MEOSAR satellite – MEOLUT – MCC – French MCC, the confirmation that the RLM was received reaches the RLSP, which initiates appropriate actions, usually to stop further repetitions of this RLM and log this.

5.7.3.3 SAR User Beacons

The basic purpose of emergency radio beacons is to get distressed persons rescued within the golden day (the first 24 hours following a traumatic event) during which the majority of survivors can usually be saved. The following types are distinguished for the different applications and the corresponding regulations:

- Emergency position-indicating radio beacon: Signal maritime distress and comply with requirements established by the International Maritime Organisation.
- Emergency locator transmitters: Signal aircraft distress and are defined in accordance with requirements defined by the International Civil Aviation Organization.
- Personal locator beacons: Signal a person in distress who is away from normal emergency services (personal use), for example, hikers. They are also used for crew-saving applications in shipping and other specialized tasks.
- Ship security alert beacon: Provide discreet SSAS security alerts, complying with International Maritime Organization (IMO) requirements. The Cospas-Sarsat 406-MHz Ship Security Alert System (SSAS) is a system implemented

by Cospas-Sarsat and contributing to IMO efforts to strengthen maritime security and suppress acts of terrorism against shipping.

In addition to the above 4 types of emergency beacons, the Cospas-Sarsat system incorporates various system beacons, including: Test, Timing, Reference, and Orbitography beacons.

When activated, beacons transmit the alert signal in the form of short RF modulated bursts at 406 MHz. The alert signal bursts, whose duration is approximately 0.5 second, are repeated roughly every 50 seconds after beacon activation for a period of at least 24 hours. Some beacons are specified to transmit for at least 48 hours.

First-generation beacons are emergency beacons compliant with Cospas-Sarsat Specification T.001. They may, but are not required to, incorporate a GNSS receiver and include information on their location in the beacon message. First-generation beacons that include a Galileo receiver may be capable of receiving an RLM. These beacons are named RLS beacons (also known as Galileo beacons).

Second-generation beacons are presently being defined by participants to Cospas-Sarsat. These beacons will include a number of advanced features (such as modulation appropriate for accurate TOA/FOA estimation and additional data bits), and they will not be backward compatible with the first-generation beacons.

5.7.4 SAR Frequency Plan

The frequency band 406–406.1 MHz is allocated by the ITU for Mobile-Satellite (Earth-to-space) use and is used by Cospas-Sarsat for SAR satellite emergency beacons. Cospas-Sarsat has divided the 406-MHz band into channels separated by 3 kHz and is approving use of specific frequencies for batches of beacons in a planned manner. This is done to ensure an even spread of used channels for growing numbers of beacons, and therefore maximizing the capacity for the 406-MHz uplink with minimal mutual interference. Figure 5.14 indicates the channelization of the 406-MHz band used for SAR.

The Galileo program has selected one specific channel for reference beacons (channel E at 406.034 MHz), and one for testing the SAR/Galileo system, channel N centred at 406.061 MHz. This channel, within the narrowband transponder mode, is not foreseen for operational use in the foreseeable future.

With the introduction of the second-generation Cospas-Sarsat 406-MHz beacons, which use spread spectrum techniques and occupy the full uplink band, the SAR/Galileo transponders will be used in normal (wideband) mode only.

The SAR/GPS (planned for GPS III capability insertion), SAR/Galileo, and SAR/GLONASS MEOSAR constellations will operate with satellite downlinks in the 1,544–1,545-MHz band. The ITU Radio Regulations allocate the 1,544–1,545-MHz band to the mobile satellite service (MSS), space-to-Earth, for distress and safety communications (article 5.356).

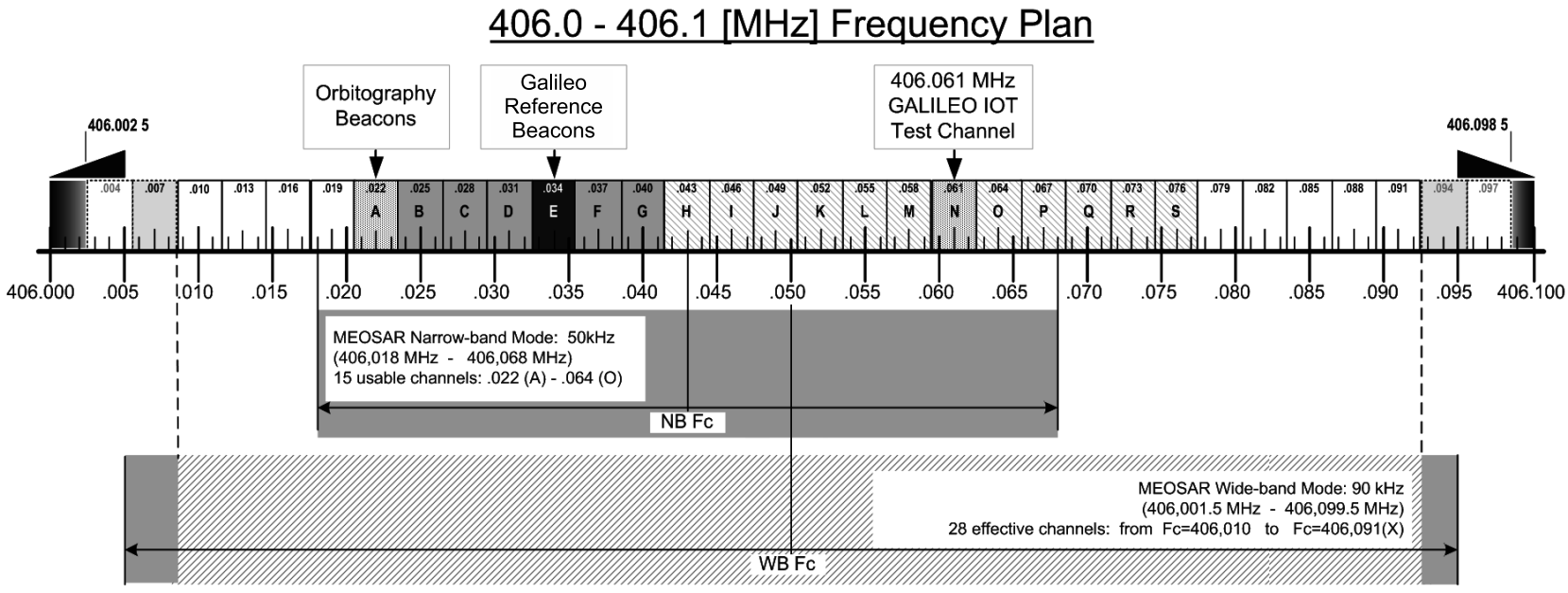


Figure 5.14 SAR UHF band.

5.8 Galileo System Performance

The fundamental theory underlying GNSS is already detailed in Chapter 2. It is not the intention of this section to repeat this information but rather to provide insight on essential performance parameters measured for the Galileo system as part of its in-orbit validation. At the end of this section, an outlook will be given by presenting the expected system performance based on extrapolation of measurements.

It has to be highlighted that the following sections present the actual measured performance as observed in January 2016, a time characterized by an incomplete system infrastructure, intense testing activities and ongoing deployment; all those factors have an influence on the results presented next.

5.8.1 Timing Performance

Users are able to derive their local realization of GST from the information provided in the navigation message. The Galileo navigation message provides users also with parameters to approximate UTC based on the receiver's realization of GST. The broadcast parameters include the number of leap seconds (i.e., integer offset between GST and UTC) and the fractional GST-UTC offset and drift. Figure 5.15 shows the achieved UTC dissemination accuracy measured during January 2016. The results present the difference between the GST-derived UTC disseminated by Galileo and the rapid UTC solution published weekly by the BIPM that is closely approximating UTC.

Galileo supports users that utilize the Galileo and GPS systems together. This is especially helpful for users with limited visibility of the satellites of both constellations. As mentioned earlier, the Galileo system provides the offset between both system time scales as part of its navigation message. The GGTO allows usage of observations from both systems without the need to compute the offset between the two timescales as an additional unknown. The achieved accuracy of the broadcast GGTO achieved during January 2016 is presented in Figure 5.16.

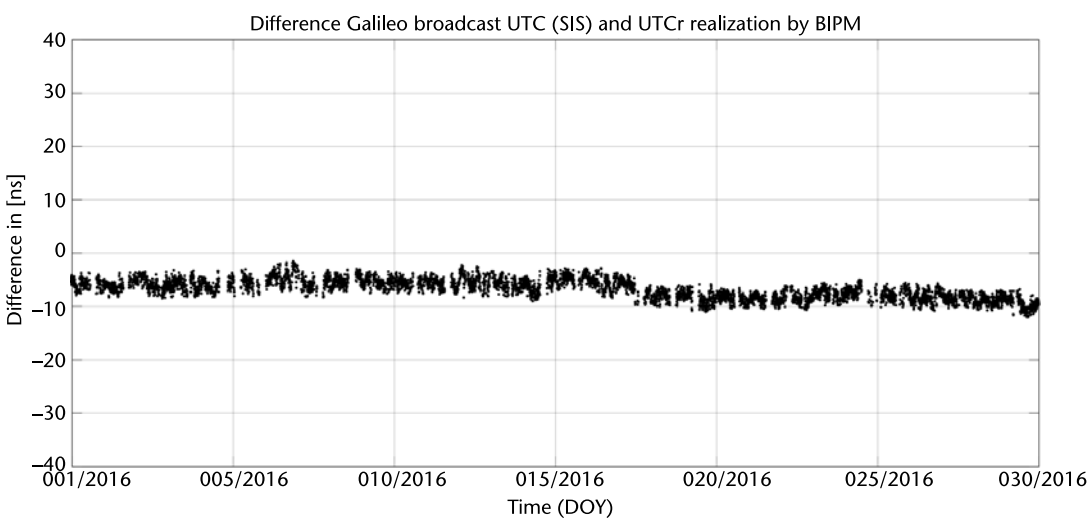


Figure 5.15 Galileo UTC dissemination performance for January 2016.

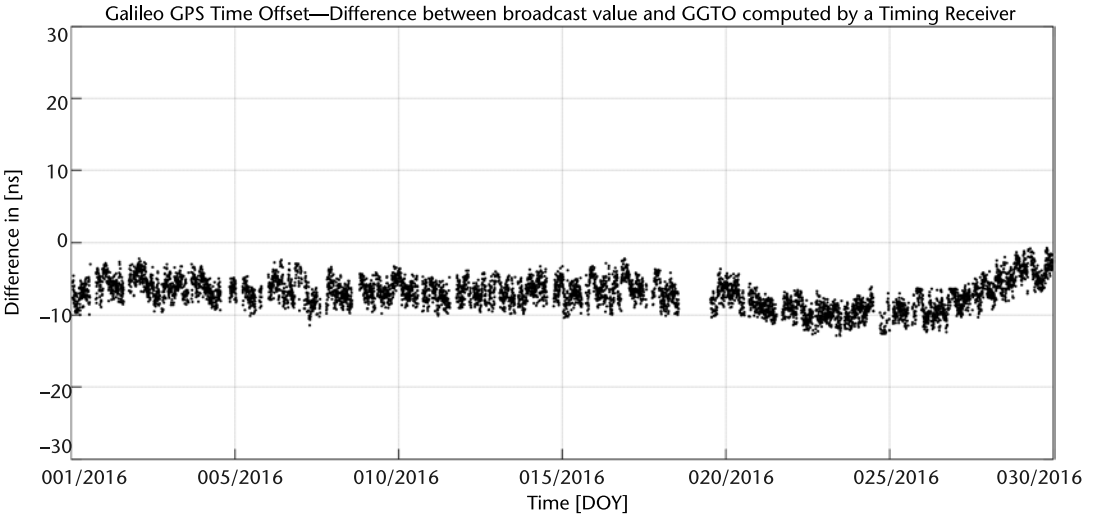


Figure 5.16 Galileo broadcast GGTO accuracy for January 2016.

5.8.2 Ranging Performance

The pseudorange measurements, which are the basis for the PVT solutions, are affected by disturbances that cause additional errors. These are sorted into three groups: space and control (e.g., ephemeris and SV clock offset prediction error, quantization errors, or signal imperfections), signal propagation environment (e.g., ionospheric and tropospheric errors and multipath) and finally the errors induced by the user receiver (e.g., measurement noise). In this section, only those contributions from the space and control segments and signal propagation environment are discussed. (Chapter 10 provides descriptions of all GNSS measurement errors.)

5.8.2.1 Orbit Determination and Time Synchronization Error

The orbit determination and time synchronization error is the error of the navigation message regarding the provided predicted orbital location and apparent satellite clock error of the satellite at a maximum operational age of the navigation message. The accuracy of the ODTS predictions provided in the navigation message is influenced even under nominal conditions by:

- The quality and availability of the observations used in the estimation process;
- The modeling of the orbit perturbations inside the orbit prediction process (including the timely variations of the satellites center of mass and antenna center of phase);
- The clock error prediction mismodeling with respect to the actual clock behavior;
- Quantization of the navigation information when generating the navigation messages;

- The refresh rate of the broadcast messages.

Figure 5.17 illustrates the geometry associated with the ODTS prediction and projection errors. This figure shows the overall orbit and clock error projected into the worst user direction.

In addition to these nominal distortions, also unexpected events can cause the degradation of the broadcast message such as, for instance, clock jumps or other failures on the ground or onboard the satellite.

The ODTS accuracy as a statistical performance characteristic of the system is defined at the worst user location at a maximum message age. For Galileo, the expected maximum age of the navigation message in the nominal system mode of operation is 100 minutes, starting from the time of the last data collected for the generation of this navigation message.

Figure 5.18 shows the cumulative distribution function of the measured ages of data as observed during January 2016 for GSAT 0101. It can be seen that about 90% of the messages have been refreshed well before the target maximum age of data of 100 minutes and that only 10% of the messages exceeded the 100 minutes. This demonstrated the feasibility to disseminate messages fast enough to ensure

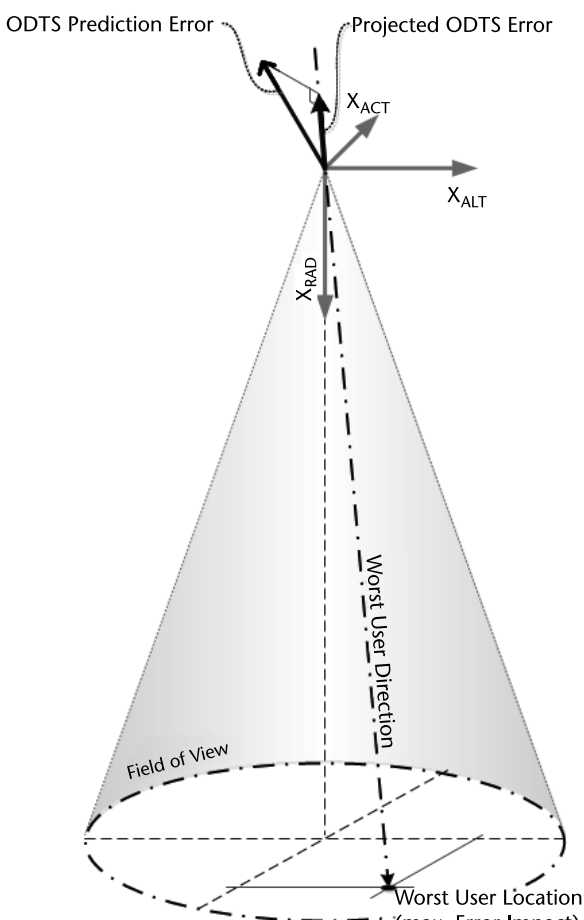


Figure 5.17 SIS ranging geometry with ODTS errors projected in the worst user direction.

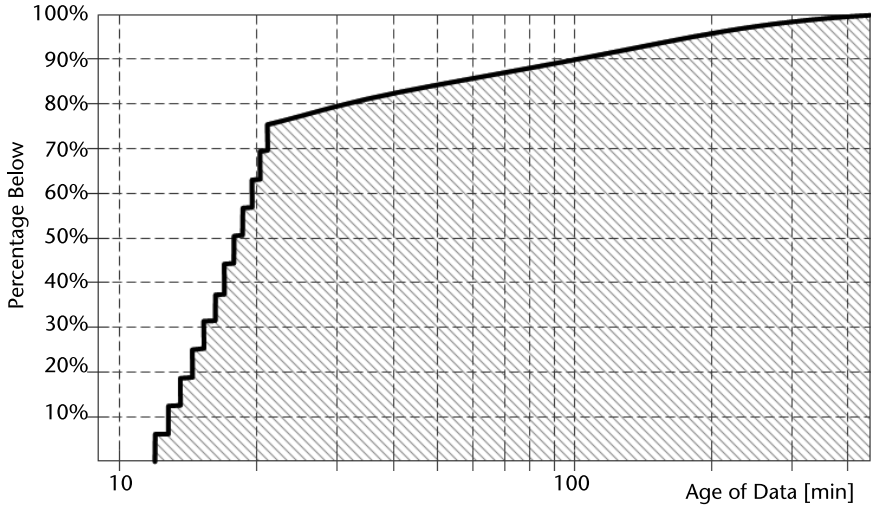


Figure 5.18 Distribution function for GSAT0101, Age of Data, January 2016.

good performance already with the Galileo system before the completion of the deployment.

Contrary to the ODTS accuracy, which is defined over a constant age of data and is used to drive the design of the system, the Signal in Space Error (SISE) is the actual measurable orbit and satellite clock prediction error when applying the correction data received by the user as part of the navigation message. The SISE is the overall orbit and clock error projected into the worst user direction (see Figure 5.17). Figure 5.19 shows the evolution of the measured SIS ranging error of GSAT0101 for the second half of January 2016. Figure 5.20 provides an overview of the ranging performance of all operational Galileo satellites for the same period (for clarity of the plot the data have been smoothed with a 14-hour moving average).

5.8.2.2 Residual Ionospheric Correction Error

The ODTS error is the dominant system contributor for dual frequency users. For the single-frequency user, the main ranging error contribution is the ionosphere, which dual-frequency users can estimate by using the different effect of the two frequencies. The single-frequency user will have to apply a model that allows him or her to reduce the impact of the ionosphere on the single frequency range measurements. These measurements can be different for each line of sight. The Galileo system provides, as part of the navigation message, the user with updated ionosphere coefficients. These coefficients allow users to determine the effective ionisation level of the ionosphere using the NeQuick G model, which is an evolution of the NeQuick model proposed by the ITU. NeQuick G is a three-dimensional empirical climatological electron density model described in [27].

Regular performance characterisation of the Galileo NeQuick G model has been done since the IOV campaign, which have shown that NeQuick G model provides better corrections than, for instance, the Klobuchar model, especially at equatorial latitudes. Figure 5.21 shows the measured correction capability of the

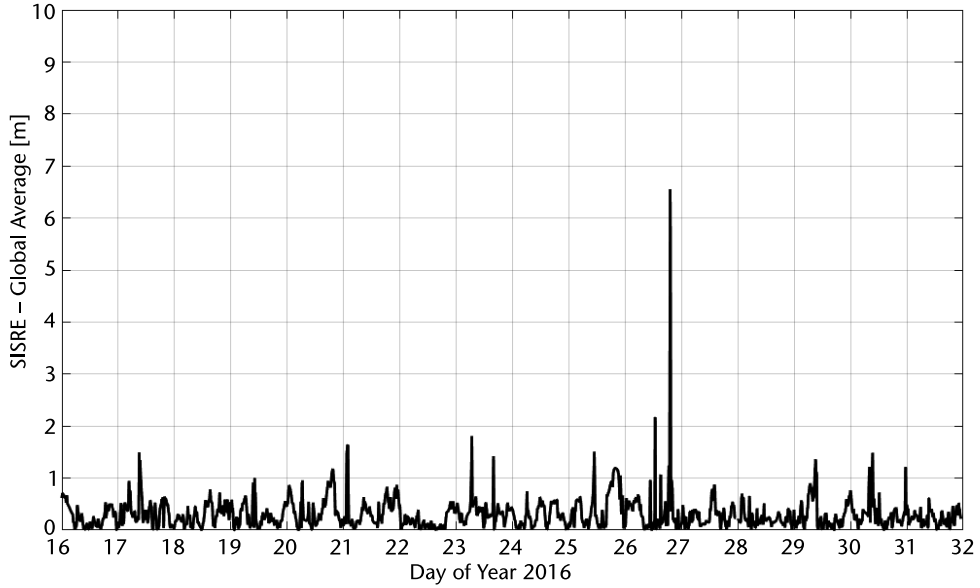


Figure 5.19 GSAT0101 SIS range error as observed January 16 to 31, 2016.

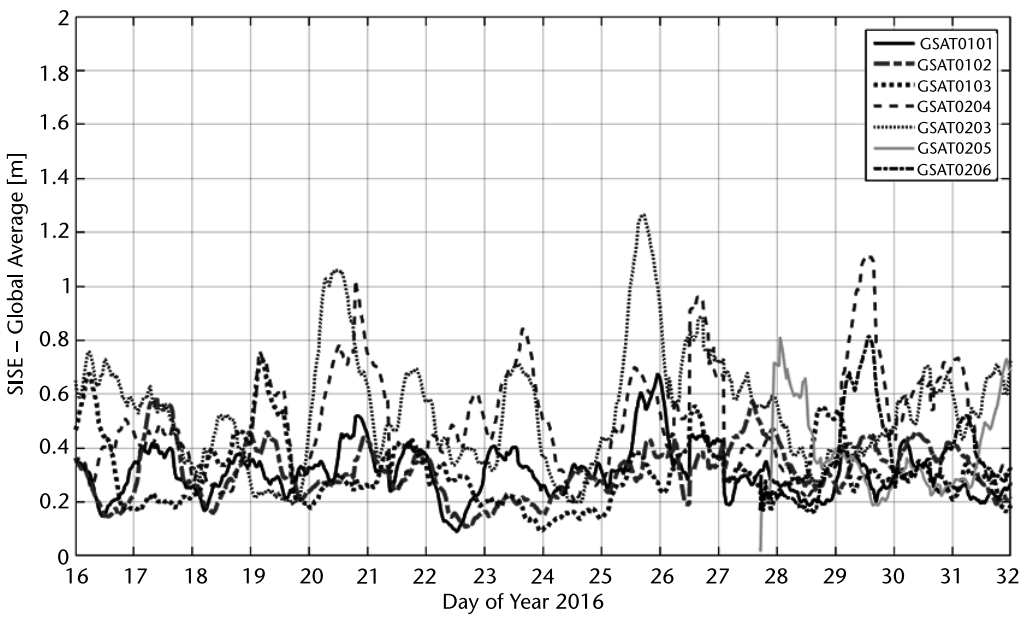


Figure 5.20 SIS range error as measured (14 hours smoothed) for all operational satellites January 16 to 31, 2016.

NeQuick G model using the broadcast IONO parameters [32, 33]. Section 10.2.4.1 provides NeQuick G model details.

5.8.2.3 Broadcast Group Delay

The BGD parameter allows the single-frequency user to correct the range measurement for signal delays in the satellite payload and RF chain. Dual-frequency users

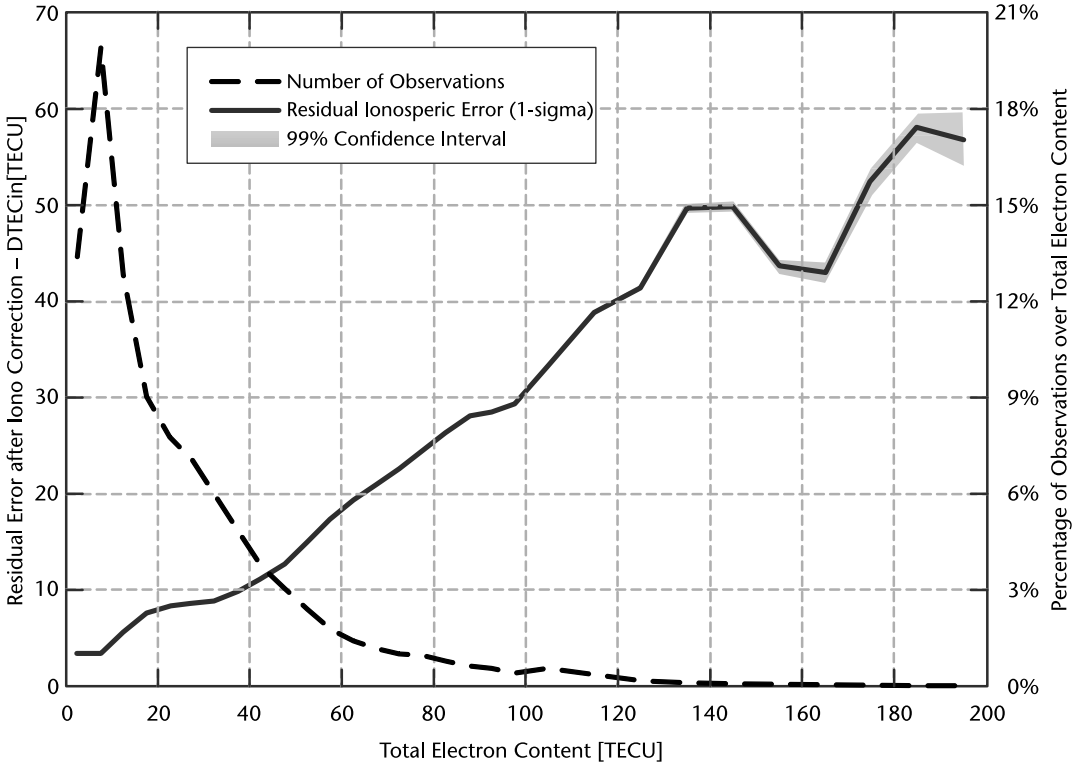


Figure 5.21 NeQuick G ionosphere model correction performance, January 2016.

do not need to correct for the BGD, because the clock corrections contained in the navigation account already for signal delays (F/NAV – E1/E5a and I/NAV for E1/E5b). The BGD between two frequencies is defined as follows:

$$\mathrm{BGD}(\mathrm{f1},\mathrm{f2}) = \frac{(\mathrm{TR}_1 - \mathrm{TR}_2)}{1 - \left(\frac{\mathrm{f}_1}{\mathrm{f}_2}\right)^2}$$

f_1 and f_2 denote the carrier frequencies of the two signals and $(\mathrm{TR}_1 - \mathrm{TR}_2)$ denote the difference of the group delays of those signals. A single-frequency user receiver can compute the correction to be applied to the dual frequency clock correction based on the following equations depending on which frequency he or she is using for the range measurements (f_1 or f_2). User receivers working on f_1 apply the following clock correction $\Delta t_{SV}(f_1) = \Delta t_{SV}(f_1,f_2) - BGD(f_1, f_2)$, and receivers using range measurements on f_2 apply $\Delta t_{SV}(f_2) = \Delta t_{SV}(f_1,f_2) - \left(\frac{f_1}{f_2}\right)^2 BGD(f_1, f_2)$.

The BGD has been characterized as part of the IOV campaign in line with the expectations on the order of 30 cm (95%). Since then, it has been continuously monitored. The contribution measured in January 2016 is also on the order of 30 cm (95%) for both the E1/E5a and the E1/E5b signal combinations.

5.8.2.4 Total UERE Budget

Other error contributors not covered by the system but impacting the range measurement are the residual tropospheric model error, the receiver dynamics and the local receiver environment in terms of interference, multipath and receiver thermal noise. Those contributors are addressed as part of the general discussions on ranging errors in Chapter 10.

Table 5.6 provides indications of the overall UERE and its contributors currently expected to be achieved by Galileo in its FOC configuration. The values are provided for single-frequency and dual-frequency users separately and given as average values over satellites and elevations and at the maximum design age of data. This summary does not differentiate between the different signals. It is a simplified version of the UERE budget used for the system design, for which the UERE contributors are considered functions of satellite elevation, user type/environment and used signal(s).

5.8.3 Positioning Performance

The positioning performance of Galileo and more generally GNSS depends on the ranging performance discussed above and is driven by the local satellite geometry. The detailed derivation of the different DOPs (HDOP, VDOP, TDOP, PDOP, and GDOP) is provided in Chapter 11 and will not be repeated here. The following table provides an overview of the different DOPs provided by the nominal Galileo constellation.

The position accuracy target for dual frequency Galileo OS users is 4m (95%) horizontal and 8m (95%) vertical. The limited satellite configuration during the ongoing deployment does not allow users to continuously derive PVT solutions. The current deployment state of the Galileo constellation allows for standalone PVT solutions for approximately 50% of the time with a Horizontal Accuracy better than 10 meters (95%), based on the operational satellites deployed by November 2016 [7]. In order to show representative PVT accuracies, a DOP-based filter has been applied, limiting the geometries to those with a HDOP less than or equal to 5. Figure 5.22 shows the measured horizontal position errors for the Galileo receiver in Noordwijk (NL) during the period March 1 to 10, 2016. As it can be seen, 95% of the position fixes are within 8.8m of the true horizontal location of the receiver antenna.

Table 5.6 Typical Design UERE Contributors Anticipated at FOC

UERE Contributor	*Single Frequency*	*Dual Frequency*
ODTS error	<65 cm	<65 cm
Satellite broadcast group delay error	<50 cm	—
Residual ionosphere error	<500 cm	<5 cm
Residual troposphere error	<50 cm	<50 cm
Thermal noise, interference, multipath, code carrier divergence	<70 cm	<100 cm
Total (RMS)	<513 cm	<130 cm

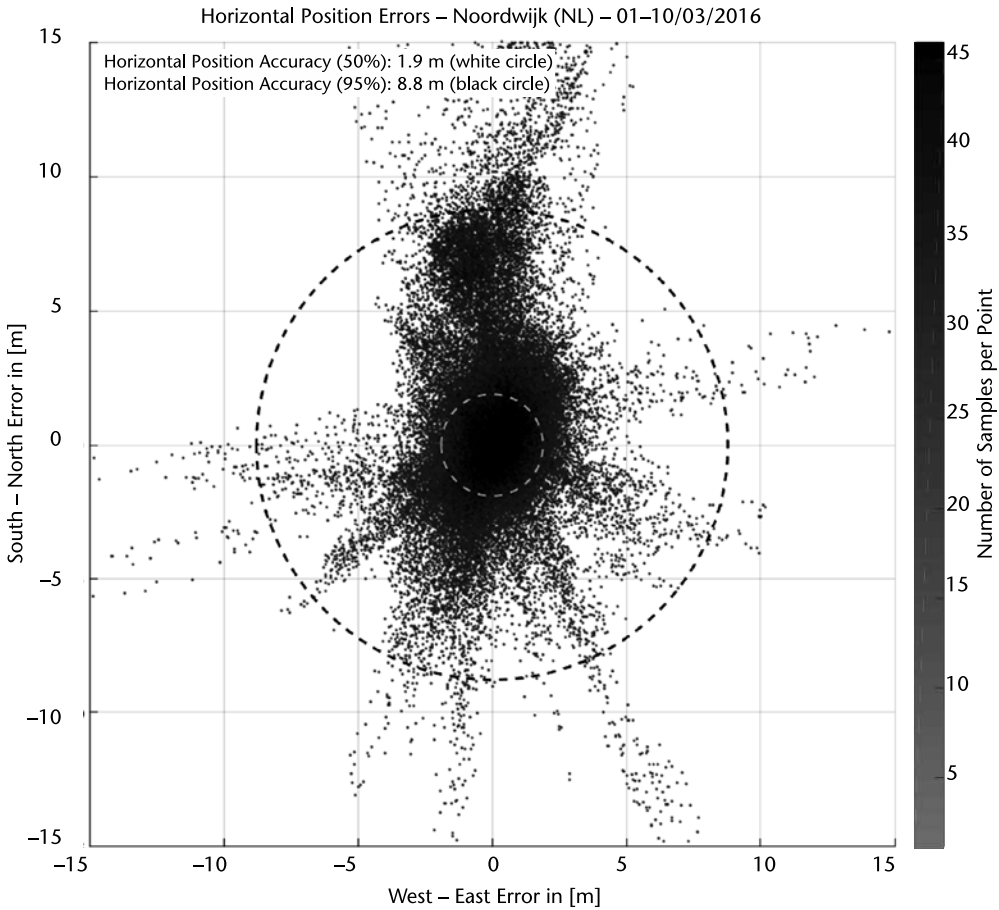


Figure 5.22 Galileo dual frequency OS horizontal position error (period from March 1 to 10, 2016).

5.8.4 Final Operation Capability Expected Performances

The extrapolation to achievable performance in FOC is based on the measurements collected during times with FOC representative conditions. The IOV performance verification has been done by means of fixed and mobile test campaigns carried out during periods with good geometric conditions and visibility of all IOV satellites. The testing did address the verification of all UERE contributors. The collected results have been filtered to remove outliers that are clearly and unambiguously caused by bad geometry conditions (e.g., PDOP above 5).

Table 5.7 DOP Values Achieved by the Galileo Walker 24/3/1 in Nominal Conditions

	Average User Location	*Worst User Location*
Horizontal DOP	1.35	1.54
Vertical DOP	2.31	2.60
Time DOP	1.48	1.58
Position DOP	2.58	2.75

The dependency of the individual performance contributors on the deployed infrastructure has been analyzed. For those parameters that were showing a clear dependence on the number of ground elements or number of satellites, extrapolation factors have been estimated from experimentation with real measurements of Galileo and GPS as well as synthetic data generated for Galileo only.

A summary of the extrapolation results is provided in Table 5.8, and the expected performance has been derived for both a typical OS dual- and single-frequency user in rural environment. The geographic distribution of the positioning performance for an OS dual-frequency user is shown in Figure 5.23.

The results of the IOV to FOC extrapolation, together with the actual test results, did confirm the feasibility of the initial design targets of positioning and timing service.

5.9 System Deployment Completion up to FOC

The Galileo system, at the time of this writing, was still under deployment. The phase between the end of the IOV and the handover of the FOC is characterized by activities linked to the full-system deployment and initial operations. During this phase, the remaining satellites will be launched and both the GCS and GMS will be completed to achieve full conformance with the mission performance and service coverage area. In parallel to the deployment activities, the operations team will ensure the maintenance of the already deployed ground and space infrastructure.

The most visible indication is the number of satellites in the constellation. At the time of this writing, three fourths of the satellites in the constellation were deployed and operated. The utilization of the Ariane 5 launch vehicle capable of injecting 4 satellites per launch was planned to continue through FOC. In addition to the constellation deployment, the ground segment is under finalization, with significant elements still to be introduced to the system configuration (such as additional mission uplink antennas). As a result of this increase of the system capabilities and the improvements in the robustness of the infrastructure, the system performance levels will gradually improve for each deployment stage.

A steady performance improvement can be observed since the first stand-alone Galileo position fix on March 12, 2013 [4, 34–36]. This trend is only interrupted by the upgrade and roll-out activities, as they occurred at the end of 2014 and early 2015 with major upgrades of the ground segment. Such long-term interruptions due to ground infrastructure deployment are expected to not reoccur since the last

Table 5.8 Expected Typical Galileo Positioning Performance

OS Dual Frequency User Vehicle—E5a/E1		*24 S/C*
Worst-case position accuracy (2σ)	Horizontal	2.9m
	Vertical	6.3m
OS single-frequency user pedestrian—E1		
Worst-case position accuracy (2σ)	Horizontal	9.9m
	Vertical	22.4m

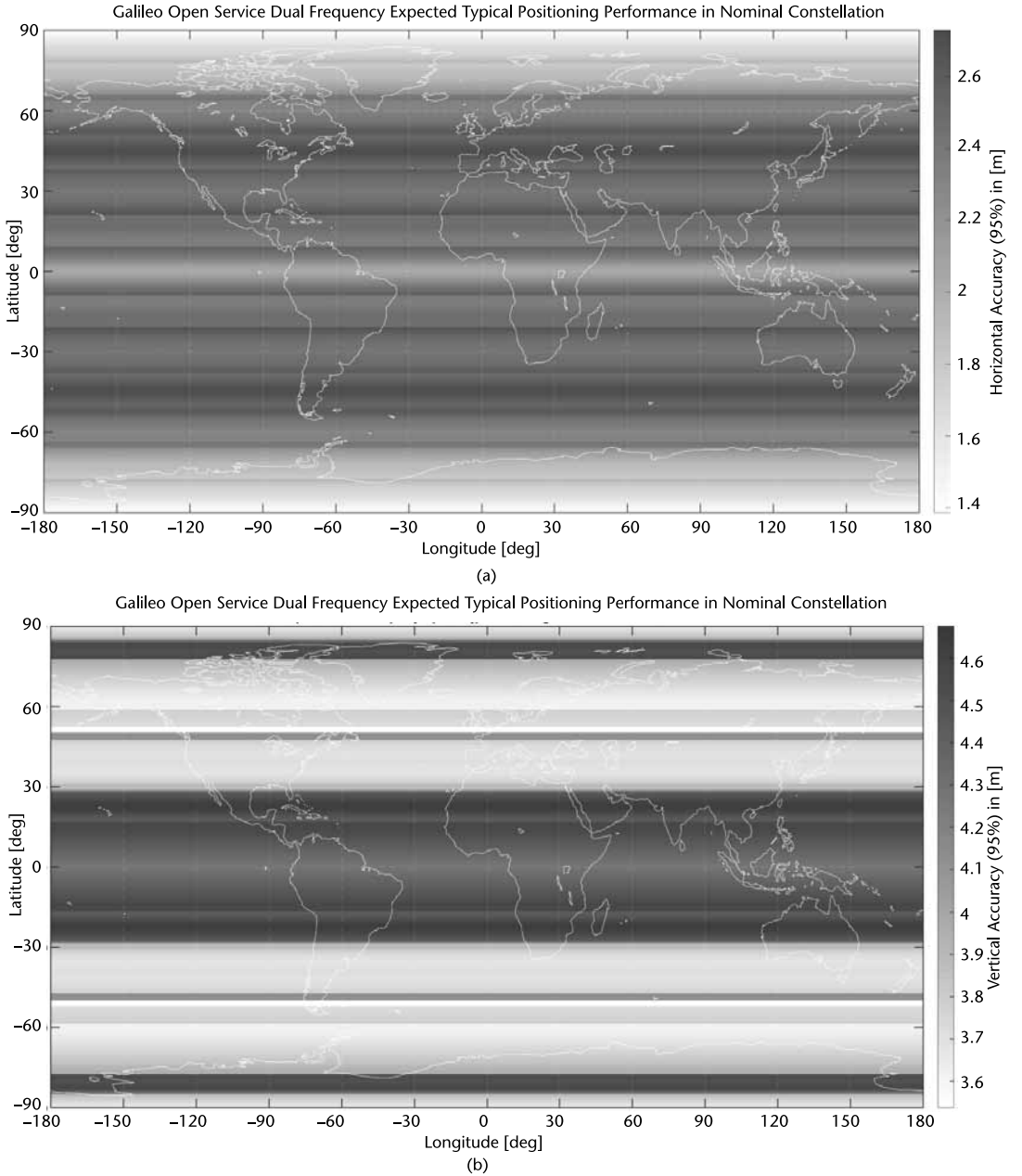


Figure 5.23 OS dual-frequency (a) horizontal and (b) vertical position accuracy (nominal constellation geometry).

ground segment upgrade. This upgrade did introduce essential elements contributing to the redundancy of the system functions. The rollout of the new version of the ground segment in 2015, for example, was carried out without a discontinuity of the provided satellite ranging signals.

The routine operations phase is intended to start with the provision of initial services while the system is still being deployed. Both the deployment and operations activities will continue in parallel up to handover of the final system

configuration. As the system configuration is expanding, the service quality will be improved with the improved system performance capabilities. After achieving the FOC configuration, the system operations are planned to last over the design system lifetime of about 20 years.

5.10 Galileo Evolution Beyond FOC

In 2007, with the awareness of the ongoing developments in the navigation field, the ESA did initialise activities on the future of the European GNSS infrastructure through an optional program supported by several of its member states. The objective of the activities is to research enabling technologies for the evolution EGNOS and Galileo systems in the presence of the increasing capabilities of existing systems as well as the growing demands of GNSS applications and users. This will ensure the competitiveness and interoperability of EGNOS and Galileo. The objectives defined for the European GNSS Evolution Program (EGEP) were [37]:

- To prepare for upgrades and evolution of EGNOS and Galileo stemming from mission evolution, improvement of performances and services, operability improvements and/or technology obsolescence;
- promote and support scientific exploitation of Galileo;
- maintain European technical know-how, competencies and infrastructures at international level;
- Sustain competitiveness and innovation capabilities.

In this context, a multitude of studies have been initiated to progress essential technologies for such system evolutions. In parallel to the technology studies, system concepts have been developed based on future GNSS mission objectives identified in cooperation with the EC.

As part of the EGEP, activities options for future GNSS system architectures are developed especially taking into account the possible deployment scenarios based on the current Galileo and EGNOS systems. The technology predevelopments are focusing along the following evolution axes: higher accuracy of the existing services, provision of a high-accuracy and certified timing service, provision of long-term ephemeris with an improved time to first fix, spoofing and jamming protection, improved SAR service, improved interoperability with other SATNAV systems, support to space users, and reduced time-to-market for future service evolutions.

The replacement of the currently deployed Galileo IOV and FOC satellites will be necessary when they reach their end of life after a lifetime of 10–12 years. New technologies, new services, and a better exploitation of the systems will maintain the competitiveness of the European navigation infrastructure.

References

[1] European Commission, "Communication from the Commission—Galileo—Involving Europe in a New Generation of Satellite Navigation Services," COM(1999)54.

[2] European Commission, "The History of Galileo," February 11, 2016, http://ec.europa.eu/growth/sectors/space/galileo/history/index_en.htm.

[3] European Space Agency; "GIOVE Experimentation Results, A Success Story," *SP-1320*, October 2011.

[4] Breeuwer, E., et al., "Galileo Works," *Inside GNSS*, March/April 2014.

[5] European Commission—Press release, "Galileo goes live!", IP-16-4366_EN, Brussels, 14 December 2016.

[6] European Commission, "Galileo Mission High Level Definition," September 2003.

[7] Galileo Initial Service—Open Service—Service Definition Document, Issue 1.0, December 2016.

[8] Fernández-Hernández, I., "The Galileo Commercial Service: Current Status and Prospects," *Proceedings of the ENC/GNSS 2014*, Rotterdam, The Netherlands; April 15–17, 2014.

[9] Galileo Initial Service—Search and Rescue—Service Definition Document, Issue 1.0, December 2016.

[10] European Commission, "Report from the Commission to the European Parliament and the Council; Mid-Term Review of the European Satellite Radio Navigation Programmes," *COM*, 2011, p. 5.

[11] Blanchard, D., "Galileo Programme Status Update," *Proceedings of the ION GNSS 2012*, Nashville, TN, September 17–21, 2012.

[12] European Commission, "European GNSS (Galileo) Open Service Signal in Space Interface Control Document (OS SIS ICD)," Issue 1.3. December 2012.

[13] Oehler, V., et al., "Galileo System Performance Status Report," *Proceedings of the ION GNSS 2009*, Savannah, GA, September 22–25, 2009.

[14] Blonski, D., et al., "Galileo as Measured Performance After 2015 Ground Segment Upgrade," *Proceedings of the ION GNSS+ 2015*, Tampa, FL, September 14–18, 2015.

[15] Zandbergen, R., et al., "Galileo Orbit Selection," *Proceedings of the ION GNSS 2004*, Long Beach, CA, September 21–24, 2004.

[16] Piriz, R., B. Martin-Peiro, and M. Romay-Merino, "The Galileo Constellation Design: A Systematic Approach," *Proceedings of the ION GNSS 2005*, Long Beach, CA, September 13–16, 2005.

[17] Navarro-Reyes, D., A. Notarantonio, and G. Taini, "Galileo Constellation: Evaluation of Station Keeping Strategies," *21st International Symposium on Space Flight Dynamics*, Toulouse, France, September 28–October 2, 2009.

[18] European Space Agency, November 17, 2016, http://www.esa.int/Our_Activities/Navigation/Galileo/Launching _Galileo/Launch_of_new_Galileo_navigation_quartet.

[19] Blonski, D., "Galileo System Status," *Proceedings of the ION GNSS+ 2015*, Tampa, FL, September 14–18, 2015.

[20] Falcone, M., "GALILEO System Status and Technology Pre-Developments," *Proceedings of the ION GNSS+ 2014*, Tampa, FL, September 8–12, 2014.

[21] Ries, L., et al., "Method of Reception and Receiver for a Radio Navigation Signal Modulated by a CBOC Spread Wave Form," Patents US8094071, EP2030039A1, January 2012.

[22] Julien, O., et al., "1-Bit Processing of Composite BOC (CBOC) Signals and Extension to Time-Multiplexed BOC (TMBOC) Signals," *Proceedings of the ION NTM 2007*, San Diego, CA, January 22–24, 2007.

[23] De Latour, A., et al., "New BPSK, BOC and MBOC Tracking Structures," *Proceedings of the ION ITM 2009*, Anaheim, CA, January 26–28, 2009.

[24] Ries, L., et al., "Tracking and Multipath Performance Assessments of BOC Signals Using a Bit Level Signal Processing Simulator," *Proceedings of the ION ITM 2003*, Portland, OR, September 9–12, 2003.

[25] Soellner, M., and P. Erhard, "Comparison of AWGN Tracking Accuracy for Alternative-BOC, Complex-LOC and Complex-BOC Modulation Options in Galileo E5 Band," *Proceedings of the ENC/GNSS 2003*, Graz, Austria, April 22–25, 2003.

[26] Lestarquit, L., G. Artaud, and J. -L. Issler, "AltBOC for Dummies or Everything You Always Wanted to Know About AltBOC," *Proceedings of the ION GNSS 2008*, Savannah, GA, September 16–19, 2008.

[27] European Commission, "European GNSS (Galileo) Open Service Ionospheric Correction Algorithm for Galileo Single Frequency Users," Issue 1.2; 2016.

[28] Hahn, J., and E. Powers, "GPS and Galileo Timing Interoperability," *Proceedings of ENC/GNSS 2004*, Rotterdam, The Netherlands, May 16–19, 2004.

[29] Galindo, F.J., et al., "European TWSTFT Calibration Campaign 2014 of UTC(k) laboratories in the Frame of Galileo FOC TGVF", Proceedings of the PTTI 2016; Monterey, CA, USA; Jan. 25-28, 2016

[30] International COSPAS-SarSat Programme; "COSPAS-SARSAT - International Satellite System for Search and Rescue"; URL http://www.cospas-sarsat.int/

[31] International Cospas-Sarsat Programme; "MEOSAR"; http://www.cospas-sarsat.int/en/2-uncategorised/177-meosar-system; 2014.

[32] Orus-Perez, R., et al., "The Galileo Single-Frequency Ionospheric Correction and Positioning Observed Near the Solar Cycle 24 Maximum," *4th International Colloquium on Scientific & Fundamental Aspects of the Galileo Programme*, Prague, Czech Republic, December 4–6, 2013.

[33] Prieto-Cerdeira, R., et al., "Ionospheric Propagation Activities During GIOVE Mission Experimentation," *4th European Conference on Antennas and Propagation (EuCAP)*, Barcelona, Spain, April 12–16, 2010.

[34] Blonski, D., "Galileo IOV and First Results," *Proceedings of the ENC/GNSS 2013*, Vienna, Austria, April 23–25, 2013.

[35] Blonski, D., "Performance Extrapolation to FOC & Outlook to Galileo Early Services," *Proceedings of the ENC/GNSS 2014*, Rotterdam, The Netherlands, April 15–17, 2014.

[36] Falcone, M., et al., "Galileo on Its Own: First Position Fix," *Inside GNSS*, March/April 2013.

[37] European Space Agency, "About the European GNSS Evolution Programme," March 3, 2015, http://www.esa.int/Our_Activities/Navigation/GNSS_Evolution/About_the_European_GNSS_Evolution_Programme, July 15–16.

CHAPTER 6

BeiDou Navigation Satellite System (BDS)

Minquan Lu and Jun Shen

6.1 Overview

6.1.1 Introduction to BDS

BeiDou Navigation Satellite System (BDS) is a global navigation satellite system independently developed and operated by China. BDS was designed to be compatible and interoperable with other GNSS constellations [1]. The name "BeiDou" comes from the Beidou constellation with seven stars, or the Big Dipper, which is near the North Star. Since ancient times, Chinese people have been using the Beidou constellation for navigation. Entering the information age, the BDS development and applications add a brand-new meaning to this ancient name.

Similar to GPS, GLONASS, and Galileo, BDS is a space-based navigation system that uses the trilateration positioning mechanism. BDS consists of a space segment, a control segment, and a user segment. The BDS space segment includes a mixed constellation of 5 GEO satellites and 30 non-GEO satellites. The BDS control segment is a distributed ground control network with a master control station, several time synchronization and information upload stations, as well as a number of monitoring stations. The BDS user segment includes various single-mode BDS terminals and multimode BDS terminals that are compatible with other GNSS systems. The main functionality of BDS is to provide 24 hours a day, all-weather, continuous, high-accuracy positioning, navigation, and timing (PNT) service for users around the world. In addition, BDS also provides a two-way short message service (SMS) and satellite-based augmentation service (SBAS) [1, 2]. BDS, together with GPS, GLONASS, and Galileo, has been identified by the United Nations (UN) International Committee on Global Navigation Satellite Systems (ICG) as one of the official GNSS providers [3].

As a national critical space-based information infrastructure, a global navigation satellite system is very important for national defense, economic development, and enhancement of people's lives. China attaches great importance to the BDS development and its applications [1]. The BDS development is aiming to construct a global space-based navigation system that is independently self-developed, open and compatible, technically advanced, stable, and reliable to promote the formation

of the satellite navigation industrial chain, to build a completed system for the support, promotion, and assurance of the national satellite navigation application industry, and to develop the extensive applications of satellite navigation in various national economic and social sectors. To achieve those goals in accordance with domestic and international requirements, including the consensus reached among members of the ICG on GNSS compatibility and interoperability, China established the following BDS development principles:

1. Openness: The BDS system construction and evolution as well as application development is open to the whole world. BDS provides high-quality services free of charge to direct users worldwide. China has been actively cooperating with other countries to promote compatibility and interoperability among GNSS components to promote the development of satellite navigation technologies and industries.
2. Independency: BDS will be independently developed and operated by China. It will be capable of independently providing services for users worldwide.
3. Compatibility: Under the framework of the ICG and the International Telecommunication Union (ITU), BDS will achieve compatibility and interoperability with other satellite navigation systems around the world to ensure that all users enjoy the benefits of satellite navigation.
4. Gradualness: China will actively and steadily promote the BDS construction and development, constantly improve the service quality, and achieve seamless interconnections among various development phases.

Under those development principles, in order to overcome difficulties of insufficient technical resources in the field of satellite navigation, limited national investment, and lack of experience in the construction and management of large-scale space-based information systems, China formulated a three-phase approach for a steady BDS development process, according to the national PNT service requirements. As a result, the BDS development follows a unique path "from regional to global; from active to passive" [1]. The three-phase development plan is as follows:

- Step 1: Start the development of the BeiDou Navigation Satellite Experimental System in 1994, to achieve the regional active service capability in 2000;
- Step 2: Start the development of the BeiDou Navigation Satellite System in 2004, and achieve the regional passive service capability in 2012;
- Step 3: Steadily push forward the development of the BeiDou Navigation Satellite System to achieve the global passive service capability by around 2020.

The three-step development path of BDS is illustrated in Figure 6.1 [2].

After continuous efforts for nearly 20 years, phase 2 of the BDS development was completed in 2012, yielding a regional navigation satellite system with 14 operational satellites in space (5 GEO+5 IGSO+4 MEO) that provided services for users in the Asia-Pacific region [1, 2]. Phase 3 of the BDS system construction process started immediately after the completion of phase 2. Four new-generation

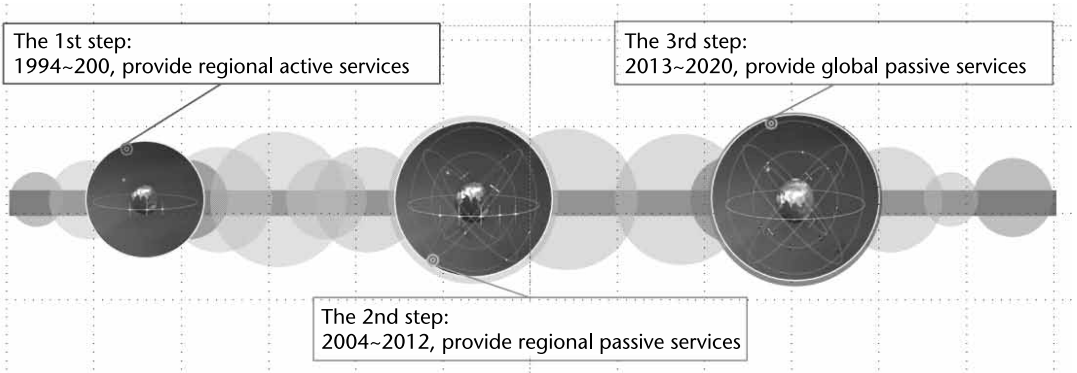


Figure 6.1 The three-phase BDS development path [2].

experimental satellites were launched in 2015 and many key technologies have been tested and verified. The deployment of the global BDS system began after the launch of the last experimental satellite on February 1, 2016. The deployment of the BDS global system with 35 satellites is planned to be completed by around 2020.

It should be noted that, according to the "Standards for Beidou Navigation Satellite System (version 1.0)", the China Satellite Navigation Office (CSNO) will gradually publish the BDS interface control documents (ICD) as well as the related performance specifications [4]. BeiDou Navigation Satellite System RNSS Signal in Space Interface Control Document (version 2.0) [5] and the Specification for Public Service Performance of Beidou Navigation Satellite System (version 1.0) [6] were released in December 2013. However, the ICD and the performance specifications corresponding to other BDS services [e.g., the radio determination satellite service (RDSS) and BDS augmentation systems] have not been published. As a result, this chapter focuses on the BDS RNSS. Information provided in this chapter related to the BDS RDSS and the BDS space-based augmentation systems is high-level and is derived from the limited information available in the public domain.

6.1.2 BDS Evolution

6.1.2.1 The Past: BeiDou Navigation Satellite Experimental System

The BeiDou project began in 1994, with a goal of establishing the BeiDou Navigation Satellite Experimental System to provide positioning, timing, and short-message services for China and its surrounding areas. The system used to be called the Beidou-1 System (BD-1). It uses the radio determination satellite service (RDSS) technique, which relies on two-way active ranging for positioning. With the completion of BD-1, China became the third country after the United States and Russia to operate a navigation satellite system. For both China and the international satellite navigation community, BD-1 represents a significant milestone [7].

According to the literature [8], China's exploration of satellite navigation systems can be traced back to the late 1960s. Inspired by the American Transit and the former Soviet Union's Tsikada, Chinese scientists carried out studies on satellite navigation systems based on Doppler measurement principles, and the work lasted

until 1980. In the late 1970s, China conducted research work on related satellite navigation and positioning systems to find a suitable solution for regional and global use. Regional satellite navigation systems were considered with 1, 2, 3, and 3 to 5 satellites. Global navigation systems with a larger number of satellites, as well as systems that provided both positioning and communication, were studied and proposed. However, none of those ideas or proposals was realized.

In 1983, Dr. Chen Fangyun, an academician of Chinese Academy of Sciences, first presented the idea of implementing rapid regional positioning and communication services by using two geostationary satellites in China, the Twin-Star Positioning System. At that time, GPS had already achieved great progress. However, the Twin-Star system combining positioning and communication services was relatively simple and economical, which made it attractive to China. In 1986, the Twin-Star Positioning System received support from the Chinese government. In June 1987, Chen et al. published a paper in which the system architecture, operating principles and mechanisms, and expected performances of the Twin-Star Positioning System were systematically introduced [9]. In 1989, demonstration and verification experiments were carried out using two DFH-2 communication satellites in-orbit, which proved the validity and feasibility of the technical platform for the Twin-Star Positioning System.

After 8 years of research and preliminary demonstration and verification, in 1994, China officially started the construction of BD-1. Two BeiDou experimental satellites (BD-1 01, BD-1 02) were successfully launched on October 31 and December 21, 2000, from the Xichang Satellite Launch Center. The two satellites were positioned at 120°E and 80°E in the geostationary orbit. BD-1 was declared to have achieved initial operational capability shortly after the successful launch of these satellites. On May 25, 2003, a third GEO satellite (BD-1 03) was launched, and was used as an in-orbit spare. On December 15, 2003, BD-1 was declared to have achieved full operational capability. China became the third country, after the United States and Russia, to own a satellite navigation system.

Around the same time that Dr. Chen Fangyun published the concept of the Twin-Star Positioning System, G. K. O'Neill from the United States proposed a similar concept named Geostar. Some patents were filed and a company with the same name was established. Geostar had a plan to establish a navigation system to cover North America and possibly the whole world. Geostar also had some communications and exchanges with the Chinese navigation community. Unfortunately, Geostar became bankrupt in 1991 [10, 11].

BD-1 utilizes two-way ranging through two GEO satellites, offering two-dimensional positioning, timing, and short-message services. It consists of three segments: a space segment with three GEO satellites (including one in-orbit spare), a control segment with one master control station and several monitoring and calibration stations, and a user segment with various types of user terminals. A schematic diagram of BD-1 is provided in Figure 6.2.

The principle of BD-1 is as follows. By using the known position coordinates of the two GEO satellites as two sphere centers, and using the distances from the satellites to the user equipment as radiuses, respectively, two spheres can be formed. The user equipment must be on the spherical intersection of the arc. Using an elevation map provided by the ground control segment, an inhomogeneous, Earth-centered sphere, with the Earth center as center and the distance from the Earth center to

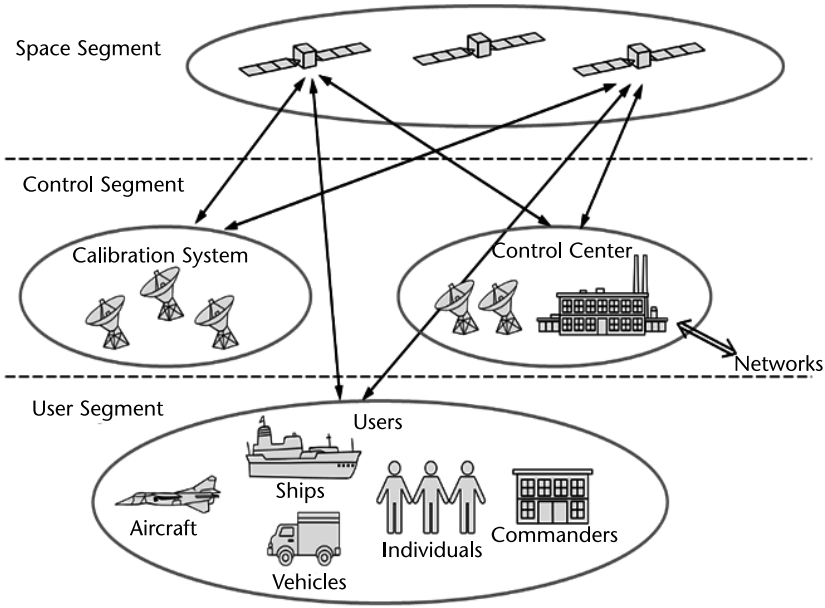


Figure 6.2 The schematic diagram of BD-1.

the user equipment as its radius, can be established. The exact position of the user equipment is the intersection of the spherical arc and the inhomogeneous sphere. The ranging information and the position calculation are performed at the ground control station through the communications with the user equipment. Meanwhile, short-message communication service can also be offered in the same time [7, 9].

Using the RDSS technique, BD-1 offers rapid positioning, short-message communication, and precise timing services. To be more specific, BD-1 provides the following services: rapidly determining the geographical position of a user; reporting the position to the user and the authorities; two-way short-message communications among the users and between a user and the ground control center; and precise one-way and two-way timing services. Since it formally began providing service in 2003, BD-1 has been used for many applications, including survey and mapping, telecommunication, water conservancy, fishery, transportation, forest fire prevention, disaster reduction and relief, and public safety. It has generated significant social and economic benefits. It played an important role in some important events and projects, such as the 2008 Wenchuan earthquake relief project and the Beijing 2008 Olympic Games [7].

The short-message service is the most unique and successful feature of BD-1. Other GNSS constellations only answer the question of "where am I?" by using a passive operational mode. Using the inherent short-message service, BD-1 can provide answers for the questions like "where are you?" Two-way precise timing is another unique feature.

Except for those advantages, BD-1 also has its weakness. Being restricted by its technical scheme, the coverage area, the user capacity, and the positioning accuracy are limited. This system does not provide velocity or height measurements. A user needs to transmit a signal to satellite when the positioning service is requested; the user velocity is limited to 1,000 km/hour. BD-1 uses active two-way ranging to

achieve two-dimensional positioning, where the position information is calculated by the ground control station and reported back to the user. This working scheme has two major shortcomings: (1) the user loses radio concealment when a position service is requested, and (2) a user terminal must contain a radio transmitter, which is a disadvantage since the user equipment has larger physical size, weight, power consumption, and cost [7].

Overall, BD-1 is a successful, practical, and economical satellite navigation experimental system. In December 2012, when the BDS regional system achieved full operational capability, BD-1 was decommissioned. However, the RDSS service that originated with BD-1 is still provided by the new BDS system.

6.1.2.2 The Present: BeiDou Navigation Satellite System (Regional)

In September 2004, phase 2 of the BeiDou development plan (construction of the BeiDou regional navigation satellite system) was initiated. The development goal of this phase was to build a regional satellite navigation system with the capability of continuous, real-time, passive, three-dimensional positioning, velocity measurement, and timing, offering PNT services to China and the Asia-pacific region. In addition to providing a Radionavigation Satellite Service (RNSS) as offered by other GNSS systems, the new system also provides RDSS service via GEO satellites. Furthermore, it provides space-based augmentation service. The system at the time was named the Beidou-2 System (BD-2). In fact, the three-phase development strategy (i.e., from experimental to regional then to global) was formally established at that time.

In April 2007, the first BD-2 MEO satellite was successfully launched, and the BeiDou navigation frequencies registered with the ITU were formally put into use. In the meantime, many technical experiments involving domestic spaceborne atomic clocks, precise orbit determination and time synchronization, and signal transmission schemes were conducted. In April 2009, the first BD-2 GEO satellite was successfully launched. In August 2010, the first BD-2 IGSO satellite was launched. With these satellites, many related technologies were validated. By April 2011, a preliminary system with 3 GEOs and 3 IGSOs was built. BD-2 began trial operation on December 27, 2011, when the system interface control document (test version) was released. In 2012, after six more satellites were launched by four rockets, the construction of the BD-2 space segment was completed. Currently, there are 14 operational BD-2 satellites in-orbit, including 5 GEO satellites, 5 IGSO satellites, and 4 MEO satellites. The BD-2 ground control segment consists of 1 master control station, 2 time synchronization and information upload stations, and 27 monitoring stations. Various BD-2 and BD-2/GNSS-compatible terminals for positioning, navigation, mapping, and surveying were also developed. With the 14 operational satellites in-orbit, BD-2 users can track at least 4 satellites anywhere, anytime in the BD-2 coverage area. BD-2 performance in its coverage area is compatible with that of the other GNSS systems.

Similar to GPS, BD-2 uses one-way passive ranging to determine a user's position. Because the user equipment works in a passive mode, the limitation on the number of users is eliminated. BD-2 broadcasts signals on 3 carrier frequencies (B1, B2, and B3). In addition to the tri-frequency RNSS service, BD-2 also integrates the RDSS service and the space-based augmentation service. At present, the in-orbit

satellites and the ground control equipment are all running in a stable condition, and the system performance meets the design specifications.

On December 27, 2012, the China Satellite Navigation Office (CSNO) announced that, in addition to the active positioning, two-way timing and short-message services, BD-2 would formally begin to provide continuous, real-time passive positioning, navigation, and timing services for China and most of the Asia-Pacific region. It was also announced that the English name of the BeiDou system would be the BeiDou Navigation Satellite System (BDS). A system interface control document entitled "BeiDou Navigation Satellite System Signal in Space Interface Control Document-Open Service Signal B1I (Version 1.0)" was also released [12].

On December 27, 2013, in a news conference held on the occasion of the 1-year anniversary of the BDS full operational capability, CSNO announced that the results of the signal monitoring and assessment over the Asia-Pacific region showed that BDS performance fully met design specifications and exceeds these specifications in some areas. Two additional system documents, "BeiDou Navigation Satellite System Signal in Space Interface Control Document-Open Service Signal (Version 2.0)" and "BeiDou Navigation Satellite System Open Service Performance Standard (Version 1.0)," were published [5, 6].

6.1.2.3 The Future: BeiDou Navigation Satellite System (Global)

Phase 3 of the BeiDou development plan is to extend the current regional system to a global system [1]. The development of the BDS global system started in 2013. It is planned that, by 2020, BDS will include a global constellation with 35 satellites to offer stable, reliable positioning, navigation, and timing services for users worldwide. It should be noted that the BDS open service signals were designed to be compatible and interoperable with the signals of GPS, GLONASS, and Galileo.

At the moment, phase 3 of the BDS development is in an engineering validation stage. The plan calls for the launch of 5 experimental satellites in various orbits to validate new technologies and new technical infrastructure for the global system. On March 30, 2015, the first new generation IGSO experimental satellite, developed and manufactured by Shanghai Micro-Satellite Engineering Center of Chinese Academy of Science, was successfully launched. This is the seventeenth member of the BDS satellite family. The satellite is positioned at an inclined geostationary orbit that is 35,786 km above the Earth with an inclination angle of 55°. This successful launch marked the beginning of BDS expansion from regional to global. On July 25, 2015, two MEO satellites were successfully launched by a shared rocket. Both satellites carried payloads with new BDS signals as well as intersatellite links. On September 30, 2015, another BDS IGSO experimental satellite was launched. This satellite utilizes the newly designed navigation satellite bus, and for the first time carried a Chinese-made hydrogen atomic clock. The fifth new-generation BDS satellite, an MEO satellite, was launched on February 1, 2016. It is the last experimental satellite of the series to test new technologies including a Chinese-made hydrogen atomic clock, intersatellite link, and new-generation satellite signals.

The requirements for the new-generation BDS satellites include improved positioning and timing accuracy, enhanced capability for self-management, a more compact system structure, and a longer life-time. Since the five experimental satellites were launched, various technical validation tasks are being conducted with

focused areas including spaceborne atomic clocks, intersatellite links, and new navigation signals. The successful launches of the experimental satellites lay a solid foundation for the global deployment of BDS.

6.1.3 BDS Characteristics

As evident from a historical review of the past 20 years of China's satellite navigation systems, BDS development has taken a path that is different from that of other global navigation satellite systems. Both GPS (see Chapter 3) and GLONASS (see Chapter 4) were built upon satellite navigation systems of previous generations. Those systems took over 20 years to complete, during which limited services were offered. Augmentation systems (see Chapter 12) for both GPS and GLONASS were constructed after the core constellations were completed, and the augmentation systems are different, operated independently from the core systems. Galileo Project, which started about the same time as BDS, also took a different approach. It first deployed an augmentation system, EGNOS (see Chapter 12), while the core constellation, Galileo (see Chapter 5) began to be deployed later. Galileo deployment has also exhibited the shortcoming of a long development cycle. Considering the national demands and technical and economic constraints, China decided to deploy BDS gradually through a three-phase plan as discussed earlier in this chapter. First, an experimental system, with lower cost and fewer technical challenges, was constructed. Second, a regional system that was more technically advanced and represented a larger financial investment was built. The regional system inherited and enhanced the functionalities of the experimental system, and also integrated augmentation system functionality into the basic navigation system. Third, with sufficiently accumulated experience, BDS will be gradually expanded to a global system. This development path reduces the investment pressure and parallelizes the construction and operational processes. However, the three-phase plan reduces the technical risks. During the phased construction stages, new technologies can be promptly introduced to the system, which ensures that the system will always be start-of-the-art. This phased approach also faces challenges, one of which is ensuring a smooth transition among the phases.

Compared with other satellite navigation systems, one of the unique BDS features is that the space segment consists of a mixed-orbit constellation of GEO, IGSO, and MEO satellites. This constellation design can continue to provide the RDSS function that began in the experimental system, while also providing an RNSS function more suited for more demanding applications. The ground track of the BDS IGSO satellites is a symmetric north-south figure-eight shape, with the middle of eight being at the equator. This orbital design means that stations inside China can track the IGSO satellites most of the time, and the utilization rate for the IGSO satellites can be over 80%. This is a very important design feature for BDS, so that it can use a minimal number of satellites to achieve regional coverage. The GEO satellites also satisfy the needs to serve a specific area (i.e., China and the surrounding region). However, the MEO satellites are more suitable to form a global constellation. The mixed space constellation with the three different orbits effectively meets various needs for a global navigation system, while providing higher-quality service in a quick manner around China.

Integrating various different services on a single platform is another unique feature of BDS. Besides the passive positioning and navigation service, BDS inherited the RDSS service from BD-1 and continues to offer active positioning, navigation, and short-message or position-reporting services. It also provides space-based augmentation service. BDS integrates RNSS, RDSS, and SBAS services, which leads to an integrated system design, an enhanced system architecture, and resource savings. The BDS satellites transmit navigation signals at three frequencies. Users can utilize the tri-frequency signals for high-precision positioning applications, such as high-precision mapping and surveying of large working areas (normally over 100 km), with significantly reduced system convergence time, enhanced positioning accuracy, and improved working efficiency.

6.2 BDS Space Segment

6.2.1 BDS Constellation

6.2.1.1 The Constellation of the BDS Regional System

The BDS development process is a phased evolution that gradually extends coverage, adds services, and enhances performance. According to the three-phase development plan, BDS has evolved from a simple constellation of 3 GEO satellites to a constellation of 14 operational satellites providing regional service. From April 2006 to December 2012, 16 satellites were launched. There are now 14 operational satellites consisting of 5 GEO + 5 IGSO + 4 MEO satellites to serve the Asia-Pacific region, as illustrated in Figure 6.3 [13].

To be more specific, the current BDS constellation includes satellites in 6 orbital planes: the GEO orbital plane, 3 IGSO orbital planes, and 2 MEO orbital planes. All of the orbits are nominally circular. The GEO satellites are operating

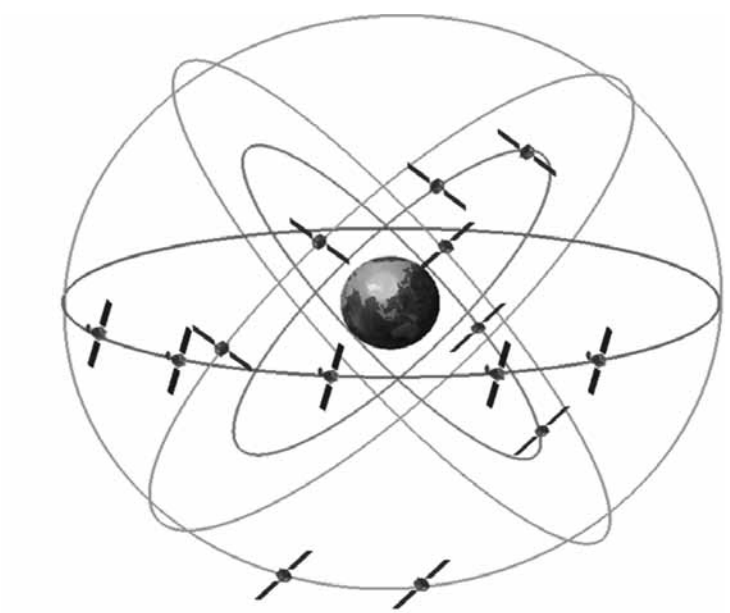


Figure 6.3 The current BDS constellation (regional, 5 GEO + 5 IGSO + 4 MEO).

in the equatorial orbit with an altitude of 35,786 km and longitudes of 58.75°E, 80°E, 110.5°E, 140°E, and 160°E, respectively. The IGSO satellites are operating at an altitude of 35,786 km and an inclination of 55° with respect to the equatorial plane. The phase difference of right ascensions of ascending nodes of the IGSO orbital planes is 120°. The satellite ground tracks for three of the IGSO satellites are coincident with an equatorial-crossing longitude of 118°E. The satellite tracks for the other two IGSO satellites are coincident with an equatorial-crossing longitude of 95°E. The MEO satellites are operating in orbit with an altitude of 21,528 km and an inclination of 55° with respect to the equatorial plane. The satellite ground tracks repeat after 13 rotations within 7 days. The MEO constellation design follows a Walker 24/3/1 construction, with a right ascension of ascending node of the satellites in the first orbital plane of 0°. The current 4 MEO satellites are in the seventh and eighth slots of the first orbital plane, and in the third and fourth slots of the second orbital plane, respectively [1, 14].

Orbital information for the current BDS constellation is provided in Table 6.1 [15]. The IGSO, MEO and GEO satellites are labeled as I, M, and G in the table, respectively.

Following the detailed description of the BDS constellation structure, we further analyze the main characteristics of the BDS constellation by examining satellite ground tracks, skymaps, and satellite coverage.

At the moment, BDS is mostly comprised of GEO and IGSO satellites over the Asia-Pacific region. The ground track repeat period of the BDS IGSO satellites is about 1 day, while the ground-track repeat period of the BDS MEO satellites is about 7 days [15]. Therefore, the ground-track repeat period for the whole constellation is 7 days. Figure 6.4 is the satellite ground track of the BDS satellites over one 7-day period from January 25, 2015, to January 31, 2015 (BDT).

Table 6.1 Orbital Information for the Current BDS Constellation*

No.	*Satellite*	*Semimajor Axis (km)*	*Eccentricity*	*Orbit Inclination (deg)*	*Argument of Perigee (deg)*	*Longitude of Ascending Node (deg)*	*True Anomaly (deg)*
1	I01	42,166.2	0.0029	54.5	174.9	209.3	220.3
2	I02	42,159.3	0.0021	54.7	187.8	329.6	87.0
3	I03	42,158.9	0.0023	56.1	187.7	89.6	326.1
4	I04	42,167.2	0.0021	54.8	167.1	211.4	201.3
5	I05	42,157.1	0.0020	54.9	183.3	329.0	65.5
6	M01	27,904.9	0.0026	55.4	182.4	108.1	118.2
7	M02	27,907.5	0.0028	55.3	180.0	107.6	167.5
8	M03	27,905.9	0.0023	54.9	170.0	227.8	325.7
9	M04	27,907.6	0.0015	55.0	190.0	227.4	351.3
10	G01	140.0°E (orbit altitude = 35,786.0 km)					
11	G02	80.0°E (orbit altitude = 35,786.0 km)					
12	G03	110.5°E (orbit altitude = 35,786.0 km)					
13	G04	160.0°E (orbit altitude = 35,786.0 km)					
14	G05	58.75°E (orbit altitude = 35,786.0 km)					

*January 25, 2013, 00:00:00 GPST.

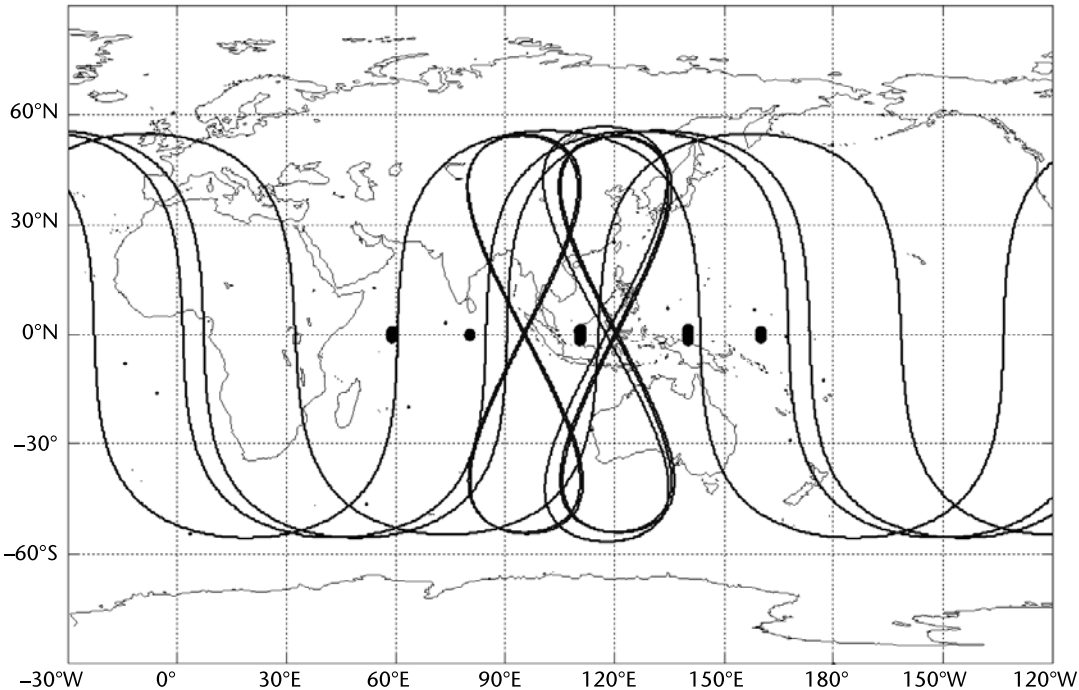


Figure 6.4 Satellite ground tracks for the current BDS constellation.

For the current constellation, a sky plot for Beijing (116.33°E, 40.00°N) is shown in Figure 6.5.

Figure 6.6 shows the number of visible BDS satellites based upon an average over 1 week. Over this period, 7 to 9 BDS satellites are visible in China and the surrounding areas.

From the above discussion, the difference in constellation designs between BDS and GPS can be easily observed. Since the current BDS constellation consists mostly of GEO and IGSO satellites, the GPS satellites are distributed more evenly across the Earth while the BDS satellites are less evenly distributed. The unevenly distributed BDS constellation, however, provides better coverage for China and the surrounding area [15].

6.2.1.2 The Constellation of the BDS Global System

According to the BDS development plan, BDS will be a global navigation satellite system upon completion. The official BDS documents indicate that the BDS global constellation will consist of 5 GEO satellites and 30 non-GEO satellites. The GEO satellites will operate in equatorial orbits with an altitude of 35,786 km and longitudes of 58.75°E, 80°E, 110.5°E, 140°E, and 160°E, respectively. The non-GEO satellites include 27 MEO satellites (with 3 in-orbit spares) and 3 IGSO satellites. With a standard Walker 24/3/1 constellation formation, the MEO satellites are evenly positioned within 3 orbital planes which are separated by 120°. The orbit is at an altitude of 21,500 km, with an inclination angle of 55°. IGSO satellites operate at altitudes of 36,000 km and are placed in 3 different inclined orbit planes with inclination angles of 55°. The subsatellite ground tracks for the IGSO satellites are

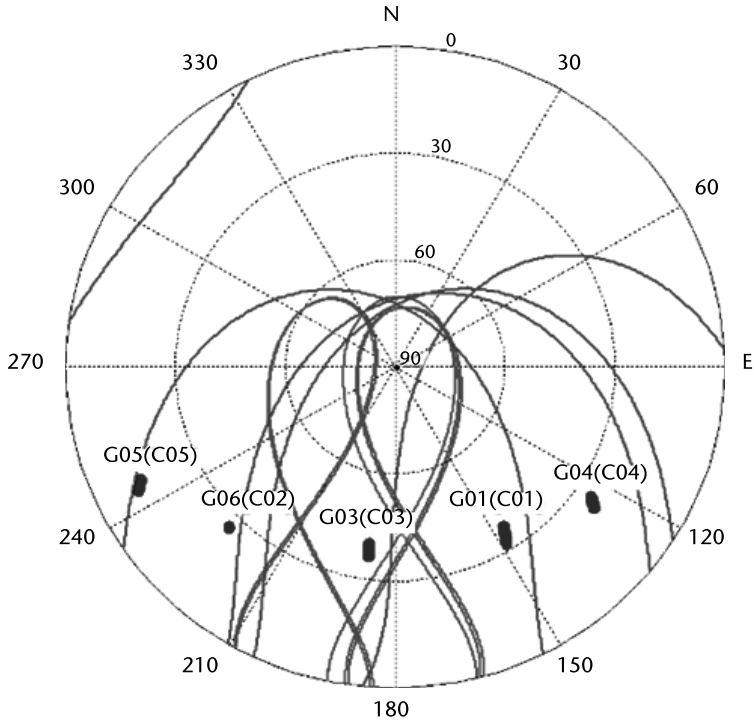


Figure 6.5 BDS sky plot for Beijing (116.33°E, 40.00°N) [data collection period: 2015/01/30 05:00~2015/01/31 09:00 (BDT)].

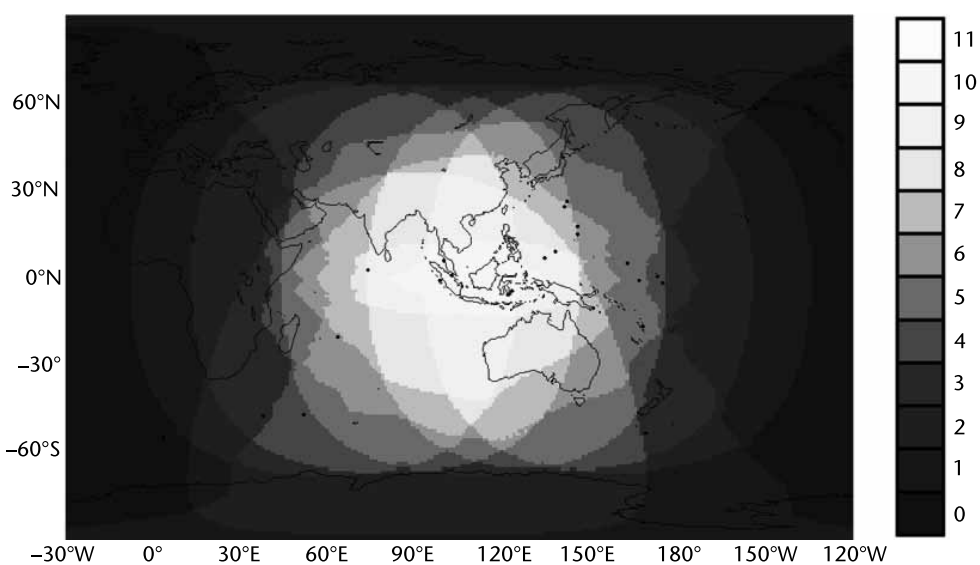


Figure 6.6 Number of visible BDS satellites over ground-track repetition period [data collection period: 2015/01/25 0:00~2015/01/31 24:00 (BDT) mask angle of 15°].

coincident while an equatorial crossing longitude of 118°E, and the satellites are phased evenly within the plane separated by 120° [1, 14].

The BDS global constellation is shown in Figure 6.7.

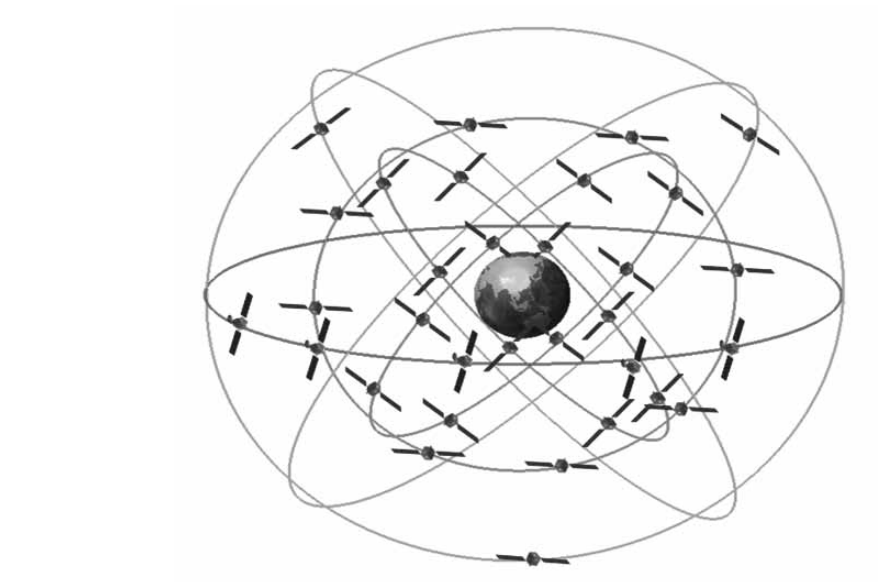


Figure 6.7 The future BDS constellation (global, 5 GEO + 3 IGSO + 27 MEO).

The BDS space segment features a mixed constellation of GEO, IGSO, and MEO satellites. Using a Walker 24/3/1 constellation design, the 27 MEO satellites are evenly distributed around the Earth to provide global coverage. The 5 GEO satellites and 3 IGSO satellites are mainly providing service in the geographic area including China and the Asia-Pacific region. Within the GEO/IGSO service area, users can track more satellites and thus receive better service, and receive additional services, such as SMS and SBAS.

Figure 6.8 presents a simulated coverage distribution for the BDS global system. It shows the geographical distribution of the BDS satellite coverage over the

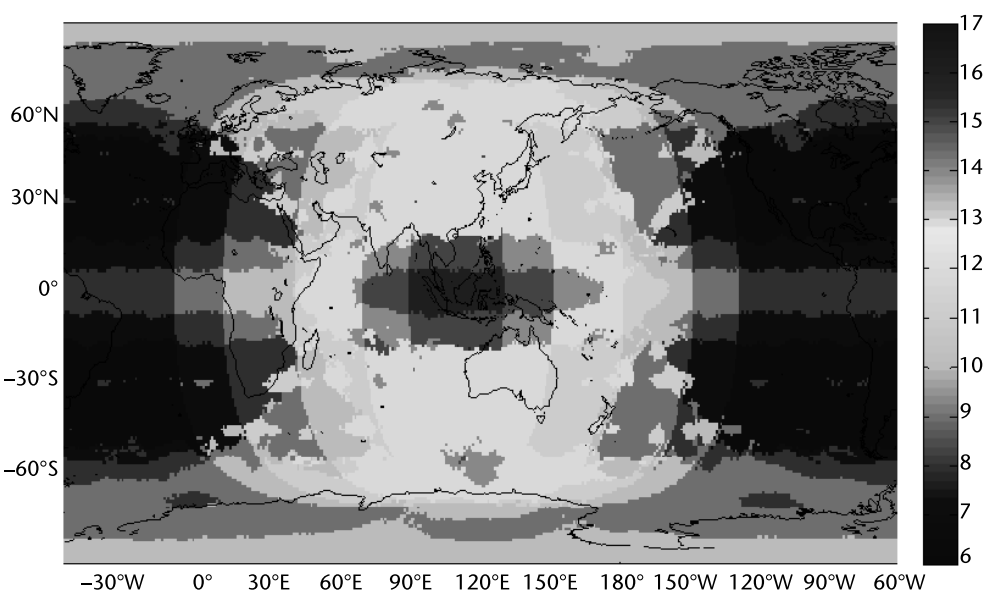


Figure 6.8 The coverage of the BDS (global, 5 GEO + 3 IGSO + 24 MEO, mask angle of 10°).

ground-track repeat period of 7 days (excluding 3 spare MEO satellites). With the GEO and IGSO satellites in the region, there will be 10 to 14 BDS satellites (95%) being tracked concurrently in the Asia-Pacific region. There will be 8 to 10 BDS satellites (95%) that can be tracked in the most of the other areas in the world.

6.2.2 BDS Satellites

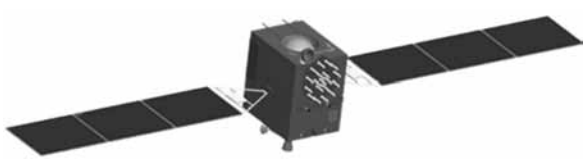
So far, all of 4 BD-1 satellites and the 16 satellites of the BDS regional system were designed and manufactured by the China Academy of Space Technology (CAST) [13]. All of the satellites were based on CAST's mature DFH-3 platform. The 4 BD-1 GEO satellites were equipped with the DFH-3 bus [16], while the following 16 BDS satellites were utilizing the enhanced DFH-3a bus [17]. Figure 6.9 illustrates the BDS GEO and IGSO/MEO satellites [7].

The platform of the DFH-3 series includes subsystems of structure, power, thermal, control, tracking and telemetry (the IGSO/MEO satellites also have a built-in data management subsystem), control, and thrust. The payload includes subsystems of navigation, antenna, while the GEO payload has the components needed for the provision of RDSS services, time and position data transmission, data uploading and precise ranging, RNSS services, while the MEO payload has components for uploads and precise ranging and RNSS services.

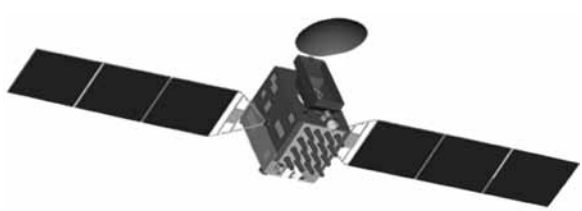
Figure 6.10 presents an expanded view of the BD-1 GEO satellites based on the DFH-3 platform. Expanded views of the BDS satellites based on DFH-3a platform have not been released [7].

Table 6.2 provides some specifications for the DFH-3 and DFH-3a platforms [16, 17].

The five BDS satellites launched after April 2015 have adopted new navigation satellite dedicated platforms. Specific information has not yet been publicly released.



BDS IGSO/MEO Satellite



BDS GEO Satellite

Figure 6.9 Illustrations of BDS satellites [7].

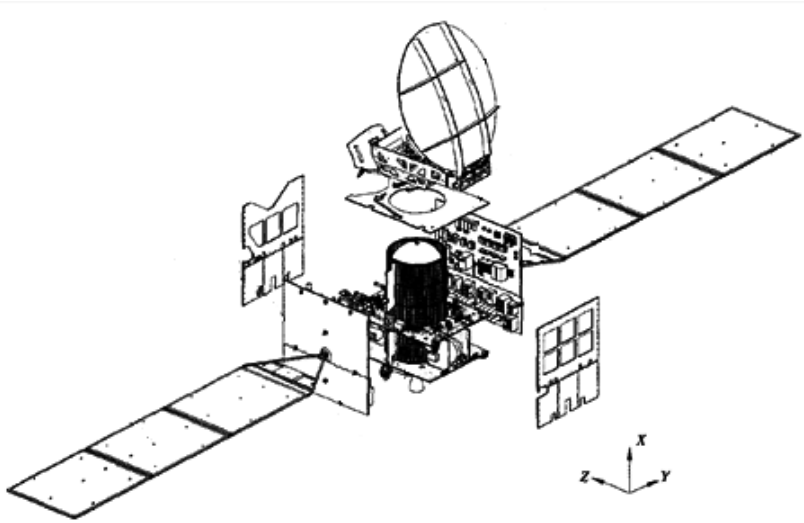


Figure 6.10 Expanded view of the DFH-3 platform based BD-1 GEO satellite [7].

6.3 BDS Control Segment

6.3.1 Configuration of the BDS Control Segment

The current BDS control segment is built upon and extended from the BD-1 control segment. The original BD-1 control segment includes a master control station together with various monitoring and reference stations across China. It is used to generate and transmit system information, to monitor and control satellite payloads, and to conduct precise satellite orbit determination [7]. The control segment of the BDS regional system consists of master control stations, upload stations, and monitoring stations. It is responsible for the operation and control of the whole system, including precise orbit determination and orbital parameter prediction, satellite clock error measurement and prediction, ionosphere monitoring and forecasting, and integrity monitoring and processing.

Currently, the BDS control segment consists of 1 master control station, 7 Class A monitoring stations, 22 Class B monitoring stations, 2 time synchronization and upload stations. The whole control segment resides within China. As the most important facility of the control segment, the master control station is located in Beijing. Sites for the upload stations were chosen based on the constellation design, so that the stations can optimally track the satellites. At the moment, the 2 time synchronization and upload stations are located in Kashi in the West and Sanya in the South, respectively. The 7 Class A monitoring stations are well distributed across China and used for satellite orbit determination and ionosphere delay calculation. The 22 Class B monitoring stations are evenly located across the country and are responsible for monitoring the system integrity [18, 19]. The distribution map of the BDS control segment is shown in Figure 6.11. Figure 6.12 presents the locations of the Class A monitoring stations.

Table 6.2 Specifications for the DFH-3 and DFH-3a Platforms (Satellite Bus)

		DFH-3	*DFH-3a*
	Description	DFH-3 satellite platform is mainly designed for communication satellites. It adopts a hexahedral structure with three compartments of propulsion, service, and communication, together with communication antennas, solar panel array. It consists of 7 subsystems: structure, control, power, tracking and telemetry, propulsion, thermal control, and communication. It uses a three-axis stabilized attitude control mechanism.	DFH-3a is an enhanced version of DFH-3.
Technical specification	*Size*	2,200 mm × 1,720 mm × 2,000 mm	2,400 mm × 1,720 mm × 2,200mm
	Mass	2,320 kg	2,740 kg
	Payload	230 kg	360 kg
	Orbit type	GEO and others	
	Antenna pointing accuracy	Pitch and roll ≤0.15° (3σ), yaw ≤0.5° (3σ)	Pitch and roll ≤0.15° (3σ), yaw ≤0.5° (3σ)
	Station-keeping accuracy	±0.1° (3σ)	±0.1° (3σ)
	Output power of solar array	1,700W	4,000W
	Effective payload power consumption		2,500W
	Designed lifetime	8 years	12 years
	Applications	Communication satellites, navigation satellites, and deep- space exploration	Communication satellites, navigation satellites, and deep-space exploration
	Missions	DFH-3 communication satellite 1997, BD-1 navigation satellites 2000, 2003, Chang'e-1 lunar mission 2007.	BDS satellites 2007 to 2011

6.3.2 Operation of the BDS Control Segment

The main tasks of the BDS control segment include [18]:

1. The master station collects all observation data from monitoring stations, processes the data to generate satellite navigation messages, monitors satellite payloads, completes mission planning and scheduling, and realizes operation control management.
2. Under the management of the master control station, the time synchronization and upload stations inject the navigation message to the satellites, communicate with the master control station, and conduct time synchronization.
3. The monitoring stations are mainly used for the continuously tracking and monitoring of the satellites, receiving navigation signals and sending them to the master control station for the navigation message generation.

Figure 6.11 The distribution of the BDS control segment [18].

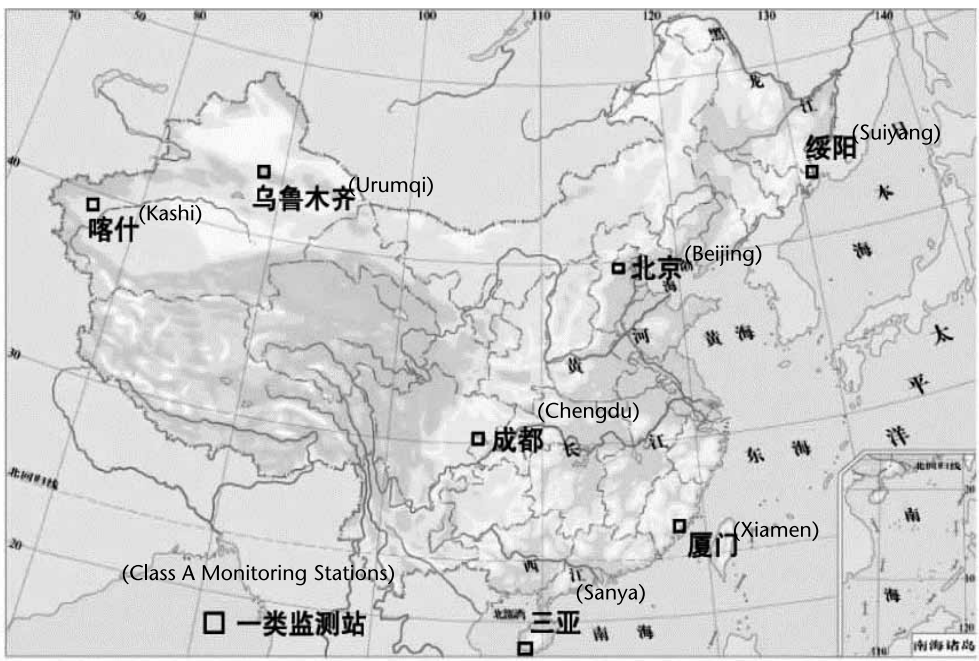


Figure 6.12 The distribution of Class A monitoring stations.

At present, the BDS control segment is significantly constrained by the fact that all of the stations are within China. As a result, it is a challenge to perform orbit determination with the precision required for the high-performance operation and control of BDS. As proposed in [18], at the current stage of development, the accuracy of the BDS coordinate system is at the centimeter level. The accuracy of BDS time is 10^{-14}, and the stability can reach 10^{-14}/week. The update cycle for the satellite orbit is 2 hours, the accuracy of orbit determination is better than 2m, yielding the corresponding user range error (URE) of 0.2m; the URE using 2-hour prediction of GEO satellite orbit is about 1m. The accuracy of the orbit prediction for the MEO and IGSO satellites is better than that for the GEOs. The URE using 6-hour prediction is about 1m, and the accuracy of 2 hours satellite clock prediction is

about 1.4 ns, where the Ionospheric corrections reduce this error component by 75%. The time for short message communication is less than 1 second, which provides for a data exchange of up to 120 Chinese characters [18].

6.4 Geodesy and Time Systems

6.4.1 BDS Coordinate System

BD-1 used the Beijing Geodetic Coordinate System 1954, with the National Vertical Datum 1985 for the height. As the system evolved, this survey and mapping reference system with horizontal two dimensions plus height could no longer satisfy the needs for a modern navigation satellite system. At the moment, BDS uses the China Geodetic Coordinate System 2000 (CGCS2000) as its coordinate system [5].

As a recent improvement of the Chinese geodetic coordinate reference system, CGCS2000 presents a concrete implementation of a global geocentric coordinate system in China. It has been formally deployed in China since July 1, 2008. The definition of CGCS2000 is consistent with that of International Terrestrial Reference System (ITRS), that is, the coordinate origin is at the center of mass of the Earth, including its oceans and atmosphere; the initial orientation value is given by the Bureau International de l'Heure (BIH) at 1984.0, where the time evolution of orientation ensures that the relative crust does not generate residual global rotation; the unit of length is the meter of the local Earth frame when considering general relativity [20].

CGCS2000 is defined as follows [21]:

1. The coordinate origin is at the mass center of the whole Earth, including oceans and atmosphere;
2. The unit of the length is the meter (SI). This scale is consistent with the time coordinate, Geocentric Coordinate Time (TCG), of the local frame of the Earth;
3. Orientation at 1984.0 is consistent with that of the BIH.
4. The evolution of orientation over time is guaranteed by a no-net-rotation condition applied to horizontal tectonic motion over the whole Earth.

The definition above can be further described as a rectangular coordinate system, with its origin and axis being defined as follows:

1. Origin: the mass center of the Earth;
2. Z-axis: points to the IERS reference pole;
3. X-axis: the IERS reference meridian plane and the equatorial plane through the origin and the orthogonal intersection with Z axis;
4. Y-axis: complete right-hand Earth-centered Earth-fixed coordinate system.

CGCS2000 defines a reference ellipsoid that rotates and is also an equipotential surface. The geometric center of the CGCS2000 reference ellipsoid is the same as the origin of the CGCS2000 coordinate system, where the axis of rotation is the same as the Z-axis of the coordinate system. The CGCS2000 reference ellipsoid also describes a normal gravity field. An equipotential, rotating ellipsoid can be

defined by four independent constants. The definition constants for the CGCS2000 reference ellipsoid are as follows:

- Semimajor axis $a = 6{,}378{,}137.0$m;
- Geocentric gravitational constant (mass of the Earth atmosphere included): $GM = 3.986004418 \times 10^{14}\text{m}^3/\text{s}^2$;
- Flattening: $f = 1/298.257222101$;
- Rate of Earth rotation: $\Omega_e = 7.2921150 \times 10^{-5}$ rad/s.

Note that the semimajor axis, flattening, and rate of Earth rotation correspond to the GRS-80 ellipsoid. However, the geocentric gravitational constant, GM, corresponds to the WGS-84 ellipsoid.

6.4.2 BDS Time System

The time reference system for BDS is the BeiDou Time (BDT) [5]. BDT adopts the international system second (SI) as the basic unit of time. BDT increments continuously without leap seconds. Its initial epoch is defined as Hour 00 Minute 00 Second 00 on January 1, 2006, of coordinated universal time (UTC). BDT counts weeks and the seconds of the week. The difference between BDT and the international atomic time (TAI) is 33 seconds.

Both BDT and TAI are atomic time scales. However, the numbers of atomic clocks used to create those time scales are different, with BDT being a local atomic time scale (i.e., it is created using clocks that are located in China and are within the BeiDou system). At the result, the lengths of the atomic second between the two systems are not exactly the same, which leads to a difference of whole seconds as well as a small daily variance C. The relationship among those systems is [15]:

$$\text{TAI} - \text{BDT} = 33s + C$$
$$\text{UTC} - \text{BDT} = -ns + C$$

where n is the number of leap seconds between UTC and BDT.

Using the reference stations at the National Time Service Center (NTSC) of the Chinese Academy of Science, BDT is compared with UTC(NTSC). Therefore, BDT is traceable to UTC(NTSC).

The difference between BDT and UTC is maintained within 100 ns (modulo 1 s). The information about the leap seconds between BDT and UTC is broadcast in the BDS navigation message.

6.5 The BDS Services

6.5.1 BDS Service Types

BDS is a multifunction global navigation satellite system that integrates many services. Upon its completion, BDS will provide global users with positioning, velocity, and timing service. 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. Those services can be classified as the following three types [1, 13]:

1. RNSS service: The RNSS service comprise of the basic navigation service that all GNSS constellations offer, namely, positioning, velocity, and timing. As with other GNSS constellations, using signals of multiple frequencies, BDS provides users with two kinds of services. The open service is available to global users free of charge. The authorized service is available only to authorized users.
2. RDSS service: The RDSS service is unique to BDS among the GNSS constellations. This service includes rapid positioning, short-messaging, and precision timing via GEO satellites for users in China and surrounding areas. This was the only service type provided by BD-1, and this functionality is now incorporated into BDS. With more in-orbit GEO satellites, the RDSS service performance has been further improved. Since the BDS RNSS service offers better passive positioning and timing performance, the short-message service 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, BDS is actually a satellite communication system with SMS services. A user identification number is required for a user to use the RDSS service; hence the RDSS service belongs to the authorized service category.
3. Wide-area differential service: 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, 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 service. 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.

Adapted to the phased BDS development plan, BDS services also gradually evolve according to the three-step plan. From the BD-1 RDSS service, the BDS services have expanded to include RDSS, RNSS, and SBAS services. The RDSS and RNSS have been formally provided by the current BDS regional system, while the BDS SBAS service has not yet formally become operational. The BDS SBAS service is expected to become fully operational when the BDS global system is fully deployed.

6.5.2 BDS RDSS Service

Using transponders from at least two GEO satellites, two-way radio links can be established between the control segment and the user segment, realizing two-way satellite ranging and data transmission, offering two-way two-dimensional positioning, communication, as well as active and passive timing. Those are called the RDSS service [7, 22].

BDS RDSS service includes:

1. Rapid positioning: After a positioning service request is sent, the system can respond within 2 seconds with the two-dimensional position coordinates of the user, providing the user and the authorities with positioning and navigation data. The positioning accuracy is 20m in the reference station coverage area, and 100m without the reference station network.
2. Short-message: Two-way short messages, up to 120 Chinese characters per message, among users and between users and ground control stations can be communicated. This type of communications can also be connected with mobile communication systems and the Internet through gateways.
3. Precise timing: The ground control center broadcasts timing information to provide users with time delay corrections.

The BDS RDSS performance has not been formally published, but the following BD-1 RDSS performance is typical [14]:

1. Positioning accuracy: horizontal – 20m (100m in the areas without reference stations);
2. Timing accuracy: One-way – 100 ns, two-way – 20 ns;
3. Short message: 1,680 bits/message (approximate 120 Chinese characters/message);
4. System capacity: 540,000 calls/hour (150 calls/second);
5. Service coverage: China and surrounding areas (70°E-145°E, 5°N-55°N);
6. Dynamic range: user velocity < 1,000 km/hour

The RDSS performance can also be illustrated from the perspective of user terminals [22]. BDS RDSS terminals can be classified into two types: one is the basic terminal for personal, vehicular and ship users, to provide users with positioning, short-message, and timing services; the other is the control terminal for user control centers, to provide control and management for the group with over 100 individual users. Different users can be assigned different service privileges with respect to service frequency and communication type. As shown in Tables 6.3 and 6.4, the users are divided into 3 types, each of which is given different service privileges [22].

6.5.3 BDS RNSS Service

Using the navigation signals at 3 different frequencies, namely B1 (1,561.098 MHz), B2 (1,207.140 MHz), and B3 (1,268.520 MHz), the BDS regional system provides positioning, velocity, and timing services for users in its coverage area, where the in-phase components of the B1 and B2 signals, B1I and B2I, are used for the open

Table 6.3 Service Frequency for Different Types of BDS RDSS Users

User type	*Service Frequency*	*Note*
Class 1	300–600 seconds	The default value is 600 seconds
Class 2	10–60 seconds	The default value is 60 seconds
Class 3	1–5 seconds	The default value is 5 seconds

Table 6.4 Communication Capabilities for Different RDSS Communication Types

Communication Type	*Length of the message length*
Type 1	110 bits (7 Chinese characters or 27 BCD codes)
Type 2	408 bits (29 Chinese characters or 102 BCD codes)
Type 3	628 bits (44 Chinese characters or 157 BCD codes)
Type 4	848 bits (60 Chinese characters or 210 BCD codes)

service, and the quadraphase components of the B1 and B2, B1Q and B2Q, together with the B3 signals, are used for the authorized service [6].

The service performance of the BDS regional system is as follows [6, 13]:

1. Coverage area: China and surrounding areas;
2. Positioning accuracy: better than 10m horizontally and 10m vertically;
3. Velocity accuracy: better than 0.2 m/s;
4. Timing accuracy: better than 50 ns.

The Specification for Public Service Performance of BeiDou Navigation Satellite System (v. 1.0) [6] describes the system performance for only the B1I signal. Within this document, detailed descriptions of the B1I coverage area, service accuracy, and service availability are provided:

1. Service area: The service area of the BDS open service (OS) is defined as the OS Signal-in-Space (SIS) coverage of the BDS satellites where both the BDS OS horizontal and vertical position accuracy is better than 10m (95%). At present, the BDS regional service capability is available, which can provide continuous OS to the area shown in Figures 6.13 and 6.14, including most of the region from 55°S to 55°N, 70°E to 150°E. It is often noted that there is a BDS-focused service area. However, this area has not been officially defined so far and is normally being considered as the coverage of the BDS RDSS services, that is, China and surrounding areas (70°E-145°E, 5°N-55°N), where the future SBAS services will be provided.
2. Service accuracy: The service accuracy of the BDS open service for positioning, velocity, and timing is described in Table 6.5.
3. PDOP availability: The BDS Open Service PDOP availability standards within its service volume are shown in Table 6.6.
4. The positioning service availability standard: The BDS OS positioning service availability standards within its coverage area are shown in Table 6.7.
5. Service accuracy standard: The service accuracy standard is presented in Table 6.8.

The BDS regional system achieved full operational capability (FOC) in December 2012. As verified from test results, the BDS regional system provides good geometric coverage in China and the Asia-Pacific area. Using a cutoff angle of 5°, in the area of 60°S-60°N and 65°E-150°E, the number of visible BDS satellites is greater than 7, and the PDOP value is normally less than 5, which satisfies the

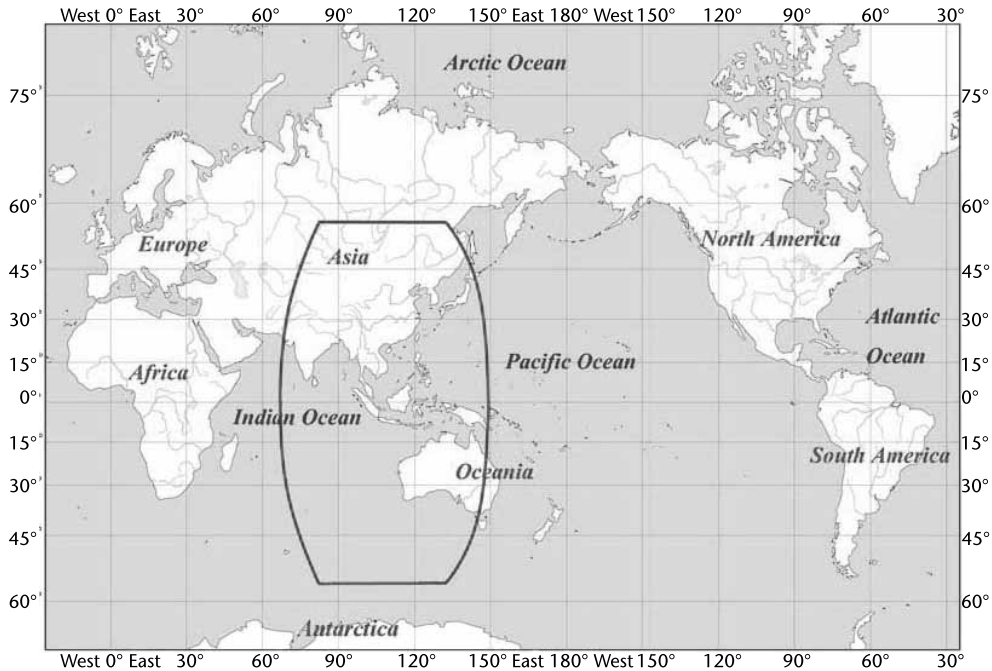


Figure 6.13 Service area of BDS regional area [6].

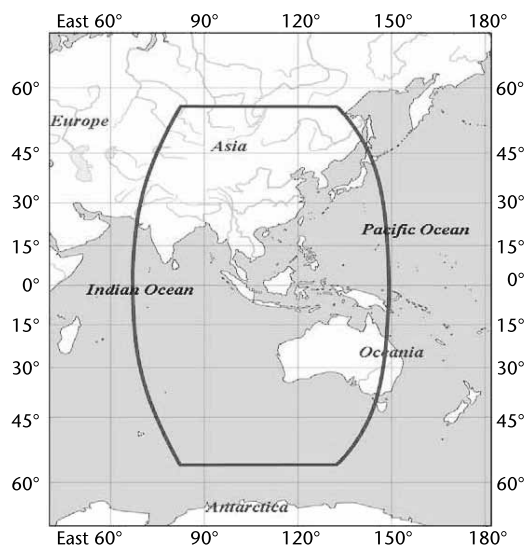


Figure 6.14 Service area of BDS regional area (partial enlarged detail) [6].

requirements for various users. The test results show that the pseudorange and carrier measurement accuracy is about 33 cm and 2 mm, repectively; the pseudorange single point positioning accuracy is better than 6m horizontally and 10m vertically; the carrier phase differential positioning accuracy in the case of an ultrashort baseline is better than 1 cm and in the case of a short baseline is 3 cm [23].

Table 6.5 BDS OS Position/Velocity/Time Accuracy Standards [6]

Service Accuracy		*Standard (95% probability)*	*Constraints*
Positioning	Horizontal	≤10m	Calculate the statistical PVT error for any point in the service volume over any 24-hour interval.
	Vertical	≤10m	
Velocity		≤0.2m/s	
Timing (Multi-SISs)		≤50ns	

Table 6.6 BDS OS PDOP Availability Standards [6]

Service Availability	*Standard*	*Constraints*
PDOP Availability	≥0.98	PDOP≤6; Calculate at any point within the service volume over any 24-hour interval.

Table 6.7 BDS OS Position Service Availability Standards [6]

Service Availability	*Standard*	*Constraints*
Positioning Availability	≥0.95	Horizontal positioning accuracy ≤ 10m (95% probability); vertical position accuracy ≤ 10m (95% probability); calculate at any point within the coverage area over any 24-hour interval.

Table 6.8 BDS OS Position/Velocity/Time Accuracy Standards [6]

Service Accuracy		*Standards (95% probability)*	*Constraints*
Position Accuracy	Horizontal	≤10m	Calculate the statistical position/velocity/time error for any point in the service volume over any 24-hour interval.
	Vertical	≤10m	
Velocity accuracy		≤0.2 m/s	
Time accuracy (multi-SISs)		≤50 ns	

6.5.4 BDS SBAS Service

BDS will comply with international civil aviation standards, through the design, validation and construction of the BeiDou Satellite-Based Augmentation System (BDSBAS), providing Category I (CAT-I) precision approach service to civil aviation users in China and surrounding areas [2]. The Satellite Based Augmentation System Interoperability Working Group (SBAS IWG) has identified the availability of BeiDou System service as part of the future global SBAS network.

In the BD-1 phase of the BeiDou development, using the BD-1 GEO satellites, some SBAS experiments were conducted. Since the BDS regional system achieved FOC, experiments and validation tests using B1I signals to implement BDSBAS service have been ongoing. Meanwhile, in addition to the B1I signal, it is being studied to offer future BDSBAS service using signals on two additional carrier frequencies,

B1C and B2a, which will be compatible and interoperable with the SBAS L1 and L5 signal family. As an integrated part of BDS, BDSBAS will offer single-frequency and dual-frequency services through 3 GEO satellites positioned at 80°E, 110°E, and 140°E. The first BDS satellite with the BDSBAS payload was planned to be launched in 2018. BDSBAS deployment is anticipated to be completed by around 2020 [2, 24].

6.6 BDS Signals

6.6.1 RDSS Signals

The RDSS service of the BDS regional system originated from the BeiDou Experimental System, or BD-1, hence its RDSS signals are fully inherited from those of BD-1. The following discussion of RDSS signals is based on publications on the BeiDou Experimental System [25–27].

As contrasted with the operation of the RNSS, where several independent navigation signals are broadcasted, the RDSS uses a scheme of "inquiry (GEO satellites)-response (user terminals)-broadcast (GEO satellites)" to realize two-way ranging and information transmission. A user terminal needs not only to receive signals from the GEO satellites (inquiry and broadcasting) but also to transmit signals to the GEO satellites. Therefore, the RDSS signals include both the downlink signals from the satellites to the user terminals and uplink signals from user terminals to the satellites. The BD-1 RDSS signal frequencies registered with the International Telecommunication Union (ITU) are in L-band 1,610 to 1,625 MHz for the uplink signals and S-band 2,483.5 to 2,500 MHz for the downlink signals [7, 25].

The BDS RDSS downlink signal is also called the outbound signal. Its carrier frequency is 2,491.75 MHz, with a signal bandwidth of 8.16 MHz (±4.08 MHz), and a minimum received signal power at the ground of –157.0 dBW [25–27]. The outbound signals use direct sequence spread spectrum (DSSS), and a modulation scheme of dual-channel offset quadrature phase-shift-keying (OQPSK) (the power level of both channels can be adjusted as needed), as well as an information format of continuous frames, where the information transmission can be divided over the time domain as super frames and fixed frames. The outbound signal includes two channels, the I-channel and the Q-channel, where the I-channel is used to transmit positioning, communication, calibration, broadcasting, and other common, public information; the Q-channel is used to transmit positioning and communication information.

The RDSS uplink signal is also called the inbound signal. Its carrier frequency is 1,615.68 MHz, with a bandwidth of 8.16 MHz (±4.08 MHz) [27]. The inbound signal uses DSSS, BPSK modulation, and a burst frame structure. Each burst frame consists of a synchronous head, service segment, and data segment, where each segment uses a different spreading code, with a spreading code rate of 4.08 MHz. The inbound signal rate is 8 kbps. Because the length of the data segment is variable, the length of the burst frames is also variable.

Since the ICD of BDS RDSS has not been publicly released, the detailed format and related parameters are not available.

6.6.2 RNSS Signals of the BDS Regional System

The BDS regional system broadcasts 5 navigation signals on 3 L-band frequencies: B1, B2, and B3. It provides both public and authorized services. The carrier frequencies of B1, B2, and B3 are 1,561.098 MHz, 1,207.140 MHz, and 1,268.520 MHz, respectively. All signals use QPSK modulation. There are two B1 channels, namely, an in-phase component B1I and a quadraphase component B1Q, where B1I provides open service and B1Q provides authorized service. The B1 signal is a QPSK signal constituted by two independent orthogonal BPSK channels. Similarly, the B2 signal also has two orthogonal BPSK channels, B2I and B2Q, to offer open and authorized services. The B3 signal is only for authorized service [5].

The B1, B2, and B3 signal characteristics are summarized in Table 6.9.

It should be noted that, in the BDS regional system, there are some differences between the B1I and B2I signals broadcasted by the GEO satellites and IGSO/MEO satellites, where the navigation message rates are 500 bps and 50 bps, respectively. The discussion in the remainder of this section will focus on the open service provided by the B1I and B2I signals [5].

6.6.2.1 Signal Structure

The B1 and B2 signals consist of ranging codes and navigation (NAV) messages for the I and Q channels, which are orthogonally modulated on carriers. The nominal carrier frequencies of B1 and B2 are 1,561.098 MHz and 1,207.140 MHz. Both signals use QPSK modulation and RHCP polarization. When the elevation angle is greater than 5°, and the RHCP user antenna gain is 0 dBi, the minimum guaranteed received power for the I-channel navigation signal is –163 dBW [5].

Table 6.9 Characteristics of B1, B2, and B3 Signals

Signal Type		*B1I*	*B1Q*	*B2I*	*B2Q*	*B3*
Service type		Open	Authorized	Open	Authorized	Authorized
Carrier frequency		1,561.098 MHz		1,207.140 MHz		1,268.520 MHz
Bandwidth (1 dB)		4.092 MHz	4.092 MHz	4.092 MHz	20.460 MHz	20.460 MHz
Multi-access scheme		CDMA	CDMA	CDMA	CDMA	CDMA
Modulation		BPSK	BPSK	BPSK	BPSK	QPSK
Pseudo code	*Length*	2,046	N/A	2,046	N/A	N/A
	Code rate	2.046 Mcps	N/A	2.046 Mcps	N/A	N/A
	Code Class	Truncated Gold	N/A	Truncated Gold	N/A	N/A
Message code rate	*GEO*	50 bps	N/A	50 bps	N/A	N/A
	IGSO/MEO	500 bps	N/A	500 bps	N/A	N/A
Error-correction code		BCH(15, 11, 1)	N/A	BCH(15, 11, 1)	N/A	N/A
Secondary coding	*Code type*	NH	N/A	NH	N/A	N/A
	Code rate	1 kbps		1 kbps		
	Length	20 bits		20 bits		
Polarization		RHCP	N/A	RHCP	N/A	N/A
Minimum received power		–163.0 dBW	N/A	–163.0 dBW	N/A	N/A
Elevation		5°	N/A	5°	N/A	N/A

The time domain expressions for B1 and B2 signals are as follows:

$$S^j_{B1}(t) = A_{B1I}\ c^j_{B1I}(t)\ d^j_{B1I}(t)\cos(2\pi f_1 t + \varphi^j_{B1I}) + A_{B1Q}\ c^j_{B1Q}(t)\ d^j_{B1Q}(t)\sin(2\pi f_1 t + \varphi^j_{B1Q})$$

$$S^j_{B2}(t) = A_{B2I}\ c^j_{B2I}(t)\ d^j_{B2I}(t)\cos(2\pi f_2 t + \varphi^j_{B2I}) + A_{B2Q}\ c^j_{B2Q}(t)\ d^j_{B2Q}(t)\sin(2\pi f_2 t + \varphi^j_{B2Q})$$

The meanings of the symbols in the above expressions are:

- The superscript j is the number of the satellite;
- A_{B1I}, A_{B1Q}, A_{B2I}, A_{B2Q}: the signal amplitude of $B1I$, $B1Q$, $B2I$, $B2Q$;
- c_{B1I}, c_{B1Q}, c_{B2I}, c_{B2Q}: the ranging code of $B1I$, $B1Q$, $B2I$, $B2Q$;
- d_{B1I}, d_{B1Q}, d_{B2I}, d_{B2Q}: the data contained in $B1I$, $B1Q$, $B2I$, $B2Q$;
- f_1, f_2: the carrier frequency of B1and B2;
- φ_{B1I}, φ_{B1Q}, φ_{B2I}, φ_{B2Q}: the initial phase of the B1I, B1Q, B2I, B2Q signal carrier.

Signal generation block diagrams for the GEO and MEO/IGSO satellites are shown in Figure 6.15 and Figure 6.16, respectively.

It should be noted that there is no integer relationship between the carrier frequency of BDS B1 and the nominal frequency of 10.23 MHz, but B1 is an integer multiple of 1.023 MHz, that is, 1,561.098 MHz = 1526 × 1.023 MHz.

6.6.2.2 Ranging Codes

The two open service signals of B1I and B2I on a satellite use the same ranging code, c_{B1I} and c_{B2I}, with a code rate of 2.046 Mcps and a code length of 2,046 [5].

Each c_{B1I} code and c_{B2I} code is made up of two linear sequences, G_1 and G_2, modulo 2 summed to generate a balanced Gold code that is then truncated by 1

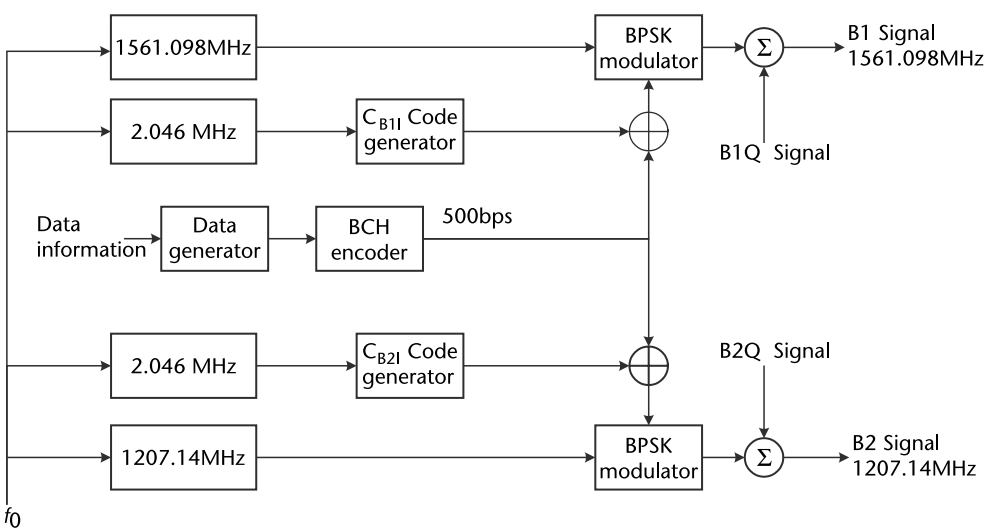


Figure 6.15 GEO satellite signal generation block diagram.

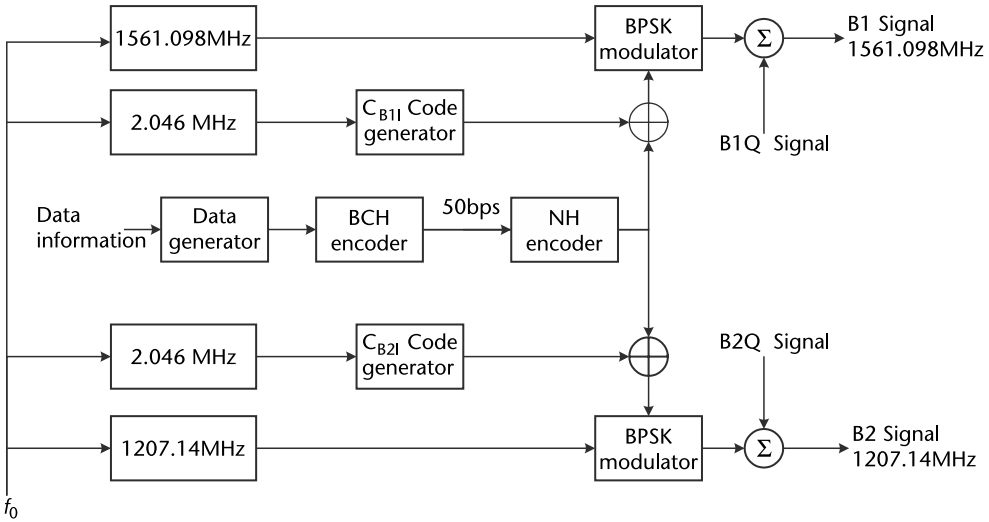


Figure 6.16 MEO/IGSO satellite signal generation block diagram.

code chip. The sequences G_1 and G_2 are generated by two 11-stage linear feedback shift registers, which can be expressed as follows:

$$G_1(x) = 1 + x + x^7 + x^8 + x^9 + x^{10} + x^{11}$$
$$G_2(x) = 1 + x + x^2 + x^3 + x^4 + x^5 + x^8 + x^9 + x^{11}$$

where the initial phases of G_1 and G_2 are:

The initial phase of G_1 is: 01010101010
The initial phase of G_2 is: 01010101010
The code generator for c_{B1I} and c_{B2I} is illustrated in Figure 6.17.

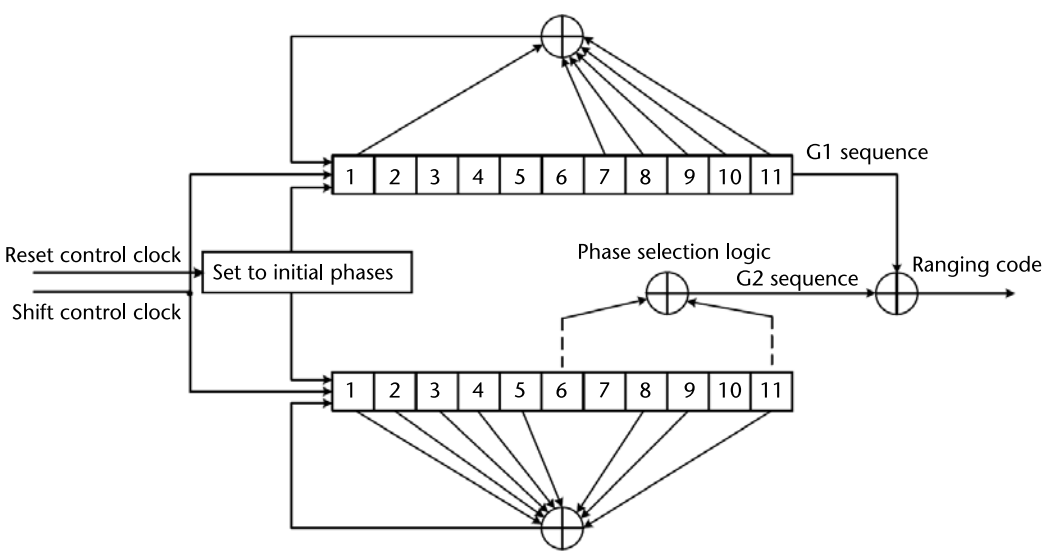


Figure 6.17 Code generator for c_{B1I} and c_{B2I} [5].

Using the modulo 2 sum of the different taps of the generated G_2 shift register, different outputs from the G_2 register can be realized, which can be used to generate different satellite ranging codes by modulo 2 summation with the G_1 sequence. The phase assignments of the G_2 sequence are shown in Table 6.10.

Correlation is the most important property of a ranging code. The auto-correlation of the BDS PRN 1 sequence and the cross-correlation of the PRN 1 and PRN 6 sequences are shown in Figure 6.18 and Figure 6.19, respectively.

Table 6.10 Phase Assignment of G2 [5]

Index	*Satellite Type*	*Ranging Code Index*	*G2 Sequence Phase Assignment*
1	GEO	1	1⊕3
2	GEO	2	1⊕4
3	GEO	3	1⊕5
4	GEO	4	1⊕6
5	GEO	5	1⊕8
6	MEO/IGSO	6	1⊕9
7	MEO/IGSO	7	1⊕10
8	MEO/IGSO	8	1⊕11
9	MEO/IGSO	9	2⊕7
10	MEO/IGSO	10	3⊕4
11	MEO/IGSO	11	3⊕5
12	MEO/IGSO	12	3⊕6
13	MEO/IGSO	13	3⊕8
14	MEO/IGSO	14	3⊕9
15	MEO/IGSO	15	3⊕10
16	MEO/IGSO	16	3⊕11
17	MEO/IGSO	17	4⊕5
18	MEO/IGSO	18	4⊕6
19	MEO/IGSO	19	4⊕8
20	MEO/IGSO	20	4⊕9
21	MEO/IGSO	21	4⊕10
22	MEO/IGSO	22	4⊕11
23	MEO/IGSO	23	5⊕6
24	MEO/IGSO	24	5⊕8
25	MEO/IGSO	25	5⊕9
26	MEO/IGSO	26	5⊕10
27	MEO/IGSO	27	5⊕11
28	MEO/IGSO	28	6⊕8
29	MEO/IGSO	29	6⊕9
30	MEO/IGSO	30	6⊕10
31	MEO/IGSO	31	6⊕11
32	MEO/IGSO	32	8⊕9
33	MEO/IGSO	33	8⊕10
34	MEO/IGSO	34	8⊕11
35	MEO/IGSO	35	9⊕10
36	MEO/IGSO	36	9⊕11
37	MEO/IGSO	37	10⊕11

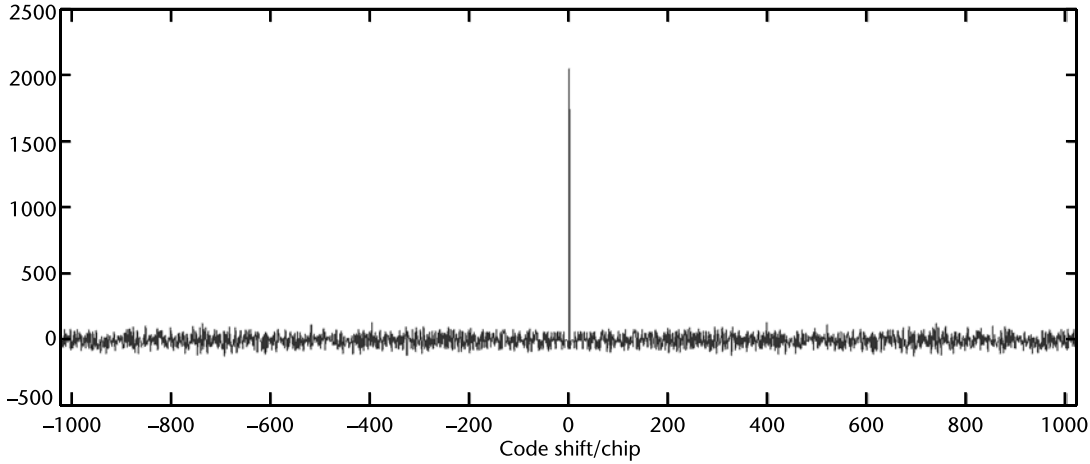


Figure 6.18 Autocorrelation of PRN 1 sequence.

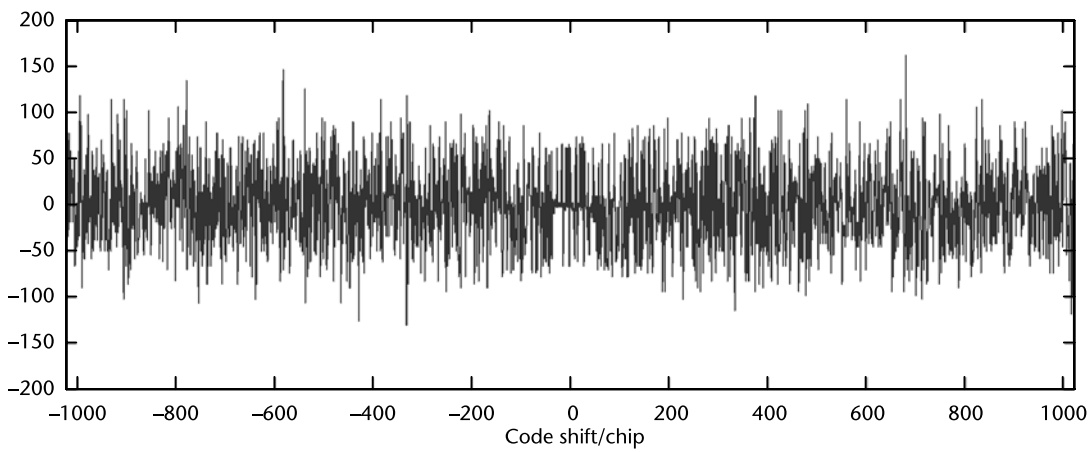


Figure 6.19 Cross-correlation of PRN 1 and PRN 6.

Figure 6.18 shows the auto-correlation for the B1I ranging code for satellite 1, where the relative delay (in chips) is specified on the X-axis. When the chip offset is 0, the auto-correlation is at the maximum value of 2,046, which is equal to the length of B1I code. When the relative delay is greater than 1 chip, the strength of the auto-correlation decreases rapidly. As illustrated by Figures 6.18 and 6.19, the BDS ranging codes have excellent auto- and cross-correlation properties.

6.6.2.3 Navigation Messages

The BDS navigation messages can be classified into two types, depending on whether they are transmitted by MEO/IGSO or GEO satellites. The NAV messages broadcast by MEO/IGSO satellites are referred to as D1. The D1 messages have a data rate of 50 bps and use a secondary code with 1-kbps code rate. The D1

message contents include the basic navigation information from the current satellite, almanac information for the whole constellation, and information regarding time synchronization with other systems. The D1 messages are broadcast simultaneously on the B1I and B2I signals from MEO and IGSO satellites. GEO satellites broadcast NAV messages of type D2, which have a data rate of 500 bps. The D2 contents include the D1 contents as well as BDS differential information, integrity information, and grid ionosphere correction information. Both NAV message types includes satellite ephemeris data in the form of Keplerian orbit elements with perturbation parameters that can be used to determine the instantaneous coordinates of the broadcasting BDS satellite in the CGCS 2000 coordinate system.

The Characteristics of the Navigation Messages

The characteristics of the BDS navigation messages can be described as follows [5]:

1. Types of NAV messages: there are two types of BDS NAV messages, D1 and D2.
2. The contents of NAV messages: The NAV messages of type D1 broadcast by the MEO and IGSO satellites only contain basic navigation information, while the NAV messages of type D2 broadcast by the GEO satellites additionally contain augmentation service information. The basic navigation information includes a frame synchronization code, or preamble (Pre), subframe counter (FraID), second-of-week counter (SOW), basic navigation message of the current satellite (ephemeris), page number (Pnum), constellation almanac information, and time synchronization information with respect to other GNSS constellations. The augmentation service information includes BDS satellite integrity information and the ionosphere grid information.

Forward error correction encoding is used for the NAV messages. A BCH (15, 11, 1) code is used with interleaving. The length of the BCH code is 15 bits, where there are 11 bits for information and 1 bit for error correction. The code generation expression is $g(x) = x^4 + x + 1$.

The NAV message bits are grouped into blocks of 11 bits. A serial/parallel conversion is made and the BCH(15,11,1) error correction encoding is performed in parallel. Parallel/serial conversion is then carried out for every two parallel blocks of BCH codes by taking bits in alternating order from the two blocks to form an interleaved code that is 30 bits in length.

Navigation Message of Type D1

The NAV message in format D1 can be characterized as follows [5]:

1. Secondary code modulation on D1. For the D1 NAV message, a Neumann-Hofman (NH) secondary code is modulated onto the ranging code. The period of the NH code is the duration of a NAV message bit. The bit duration of the NH code is the same as the period of the ranging code. As shown in Figure 6.20, the duration of 1 NAV message bit is 20 ms and the ranging code period is 1 ms. The NH code is (0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0), which is 20 bits in length clocked at a rate of 1 kbps with a

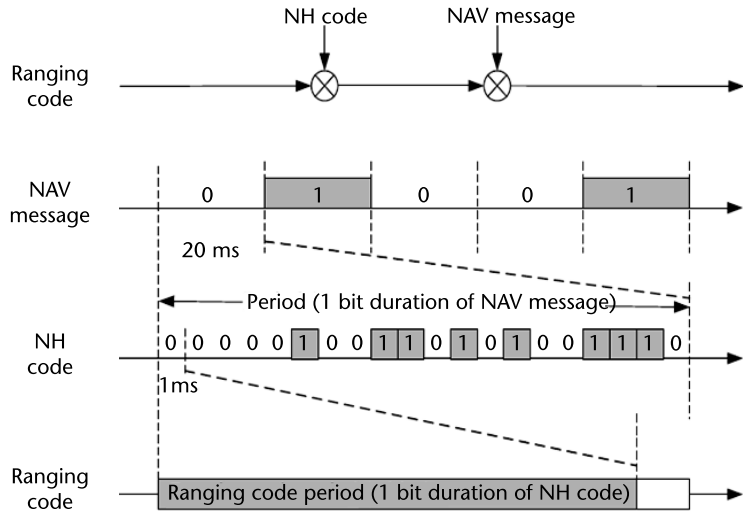


Figure 6.20 Secondary code and its timing [5].

bit duration of 1 ms. It is modulated on the ranging code synchronous with the NAV message bits.

2. D1 NAV message frame structure: The D1 NAV message data bits are organized into superframes, frames and subframes. Every superframe has 36,000 bits and lasts 12 minutes. Every superframe is composed of 24 frames (24 pages) that are each 1,500 bits in length and last 30 seconds. Every frame is composed of 5 subframes. Every subframe has 300 bits and lasts 6 seconds. Every subframe is composed of 10 words. Every word has 30 bits and lasts 0.6 second. The 30 bits in each word consist of NAV message data and parity bits. In the first word of every subframe, the first 15 bits are not encoded and the following 11 bits are encoded using a BCH (15, 11, 1) code for error correction. So within the first word, there are 26 information bits and one group of 4 parity bits. For the other 9 words in the subframe both BCH (15, 11,1) encoding for error control and interleaving are involved. Each of these 9 words of 30 bits contains two blocks of BCH parity (8 bits in total) and there are altogether 22 information bits. The D1 NAV message structure is shown in Figure 6.21.
3. D1 NAV message detailed structure: The D1 NAV message conveys basic NAV information, which includes fundamental NAV information pertaining to the broadcasting satellite (seconds of week, week number, user range accuracy index, autonomous satellite health flag, ionospheric delay model parameters, satellite ephemeris parameters and their age, satellite clock correction parameters and their age and equipment group delay differential), almanac and BDT offsets from other systems (UTC and other navigation satellite systems). It takes 12 minutes to transmit the whole NAV message. The D1 frame structure and information contents are summarized in Figure 6.22. The fundamental NAV information of the broadcasting satellite is in subframes 1, 2, and 3. The information contents in subframes 4 and 5 are subcommutated 24 times each via 24 pages. Pages 1 to 24 of subframe 4 and pages 1 to 10 of subframe 5 are used to broadcast almanac and time

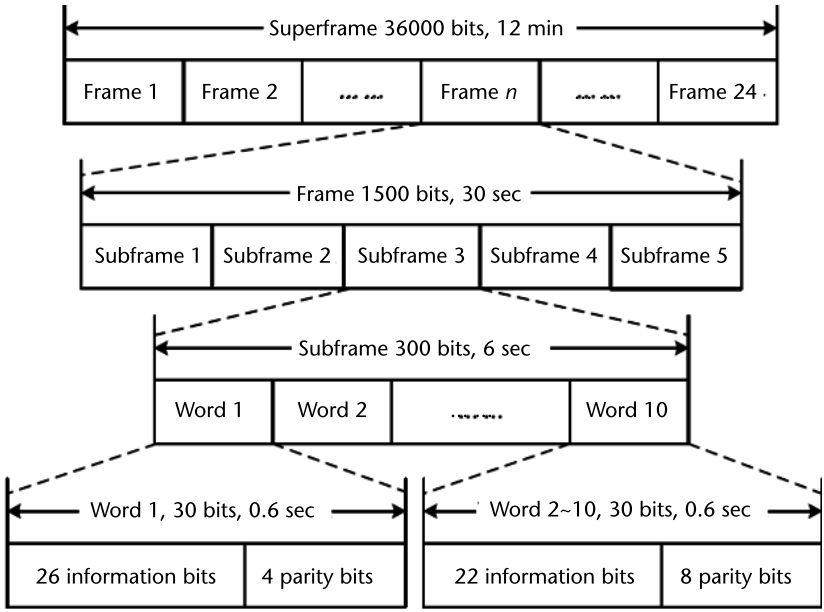


Figure 6.21 Frame structure of NAV message in format D1 [5].

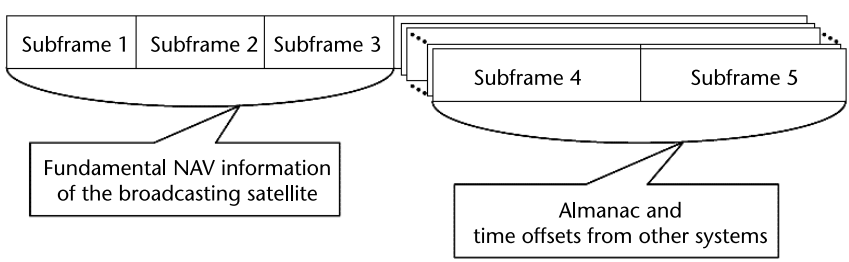


Figure 6.22 Frame structure and information contents of NAV message in format D1 [5].

offsets from other systems. Pages 11 to 24 of subframe 5 are reserved. The detailed information for each subframe is presented in the BDS ICD [5].

Navigation Message of Type D2

The NAV messages in format D2 can be characterized as follows [5]:

- D2 NAV message frame structure: The D2 NAV message is structured into superframes, frames, and subframes. Every superframe is 180,000 bits long, lasting 6 minutes. Every superframe is composed of 120 frames, each of 1,500 bits in length and lasting 3 seconds. Every frame is composed of 5 subframes, each with 300 bits and lasting 0.6 second. Every subframe is composed of 10 words, each with 30 bits and lasting 0.06 second. Every word includes NAV message data and parity bits using the same BCH encoding as described above for the D1 NAV message. Figure 6.23 provides an overview of the structure.

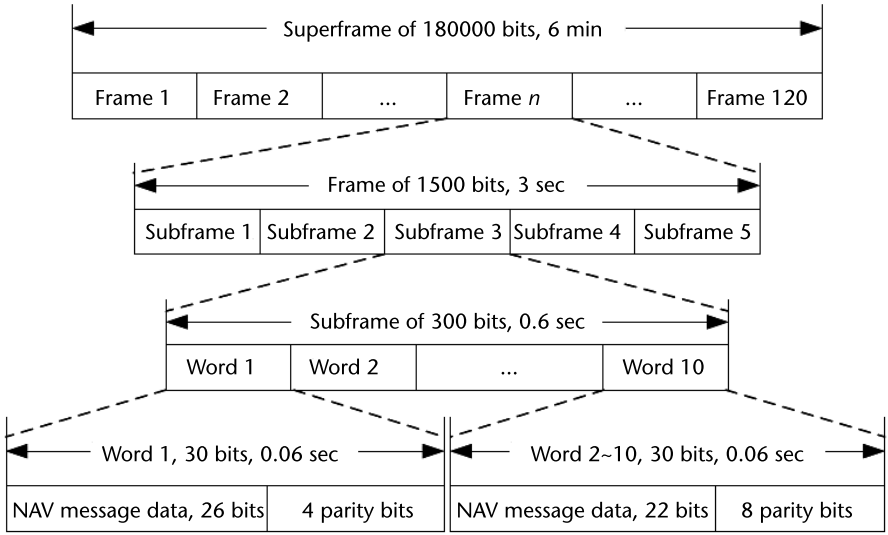


Figure 6.23 Structure of NAV message in format D2 [5].

- D2 NAV message detailed structure: The information content of the D2 NAV message includes: the basic NAV information of the broadcasting satellite, almanac, time offsets from other systems, integrity and differential correction information, and ionospheric grid information as shown in Figure 6.24. Subframe 1 is subcommutated 10 times via 10 pages. Subframes 2, 3, and 4 are subcommutated 6 times each via 6 pages. Subframe 5 is subcommutated 120 times via 120 pages. Details of the D2 NAV message are presented in [5].

6.6.3 RNSS Signals of the BDS Global System

6.6.3.1 Proposed RNSS Signals for the BDS Global System

The development process of the BDS global system started immediately after the BDS regional system was fully operational in December 2012. However, the related research and design work of the BDS global system, including design of the signals, can be traced back to around 2005. Since then, various proposed signal designs for the BDS global system have been published. Early research mainly focused on performance analysis and improvement of known BOC signal designs, while more

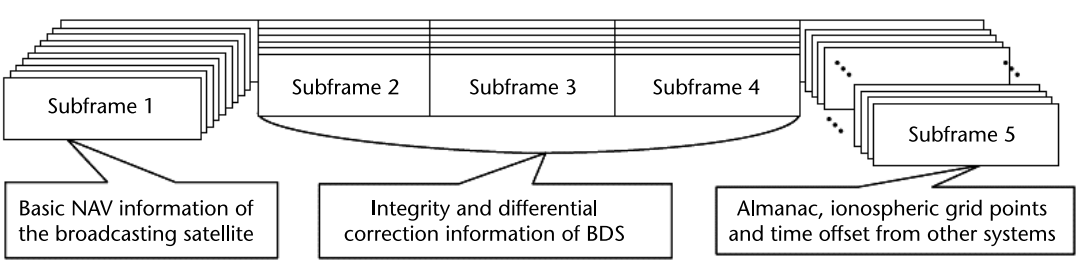


Figure 6.24 Frame structure and information contents of NAV message in format D2 [5].

novel signal designs have been proposed recently. Signal design activities for the BDS global system have focused on two areas: (1) ensuring the signal design provides enhanced performance as needed to meet the requirements of various classes of user, and (2) ensuring compatibility and interoperability with other GNSS constellations under the ICG framework.

On various occasions such as frequency coordination, bilateral and multilateral discussions, BDS representatives introduced BDS signal structures and related characteristics. For example, in July 2009, at an ICG meeting, the China National Administration of GNSS Applications (CNAGA) published the latest status of the signal design for the BDS global system (which was then referred to as COMPASS) [28]. Many publications since then have included a consistent description of the BDS signals [29].

The published signals of the BDS global system can be characterized as follows. BDS will use 3 center frequencies: B1, B2, and B3. The center frequencies for the new B1, B2, and B3 signals are 1,575.42 MHz, 1,191.795 MHz, and 1,268.52 MHz, respectively. The B1 signals provide open and authorized services; the B2 signals only provide open services; and the B3 signals only provide authorized services. It should be noted that the open service frequencies of B1 and B2 are identical to those used for GPS L1, L2, and Galileo E1, E5. The new BDS signals adopt more advanced BOC modulation schemes, and reduce the data rate of the GEO open service signals to 100 bps.

The main parameters of the B1, B2, and B3 signals are presented in Table 6.11 [29].

Important details regarding the major BDS signals include:

1. B1C: With a center frequency of 1,575.42 MHz, B1C provides both open and authorized services, where the open service signal utilizes an MBOC(6,1,1/11) scheme and contains a data channel B1cx and a pilot channel B1cy. The authorized signal B1cz uses a BOC(14,2) modulation scheme.
2. B2a/B2b: With a center frequency of 1,191.795, B2 signals only provide open services. Using the AltBOC(15,10) modulation scheme, there are two

Table 6.11 Characteristics of the B1, B2, and B3 Signals for the BDS Global System

Signal Name	*B1c*			*B2a/B2b*				*B3*		
Signal components	B1cx	B1cy	B1cz	B2ax	B2ay	B2bx	B2by	B3x	B3y	B3z
Service type	Open		Authorized	Open				Authorized		
Carrier frequency (MHz)	1,575.42			1,191.795				1,268.52		
Modulation	TMBOC(6,1,4/33)		BOC(14,2)	AltBOC(15,10)				BOC(15,2.5)		BPSK(10)
Pseudo code rate (Mcps)	1.023	1.023	2.046	10.23	10.23	10.23	10.23	2.5575	2.5575	10.23
Message rate (bps)	100	—	100	50	—	100	—	100	—	500
Secondary Coding (bits)	N/A	200	N/A	10	200	10	200	N/A	N/A	N/A

signals, B2a and B2b, that carry different NAV messages. The B2a signal has a center frequency of 1,176.45 MHz, and is comprised of a data channel B2ax and a pilot signal B2bx. The B2b signal has a center frequency of 1,207.14 MHz, and is comprised of a data channel B2ax and a pilot signal B2by.

3. B3: With a center frequency of 1,268.52 MHz, B3 signals are only for authorized services. The first B3 signal uses the BOC(15,2,5) modulation scheme and is comprised of a data channel B3x and a pilot channel B3y. The second authorized signal, B3z, uses BPSK(10) modulation.

No messages are modulated on any of the pilot channels. The open service pilot channels only convey fixed symbol data (a 200-bit secondary code). A secondary code is also used for the B2a data channel with a fixed length of 10 bits. The carriers of each open signal's data channel and pilot channel are generated in phase quadrature. The data channels of different frequencies use different secondary codes, while the data channels of the same frequency of all satellites use the same secondary code.

Figure 6.25 presents the spectrum of the B1, B2, and B3 signals for the BDS global system [29].

6.6.3.2 Recent Advances in RNSS Signal Design for the BDS Global System

In recent years, Chinese researchers have been studying more advanced signal designs for the BDS global system [30, 31]. This research has been motivated by: (1) the desire to improve the performance of the preliminary signal designs, (2) the need to address problems encountered during the BDS regional system operation, and (3) the desire to avoid potential patent risks involving the TMBOC and AltBOC modulation schemes. Published results include a new modulation scheme known as quadrature multiplexing BOC (QMBOC) for the B1 signal [32, 33], new multisignal multiplexing technologies known as time division AltBOC (TD-AltBOC) and asymmetric constant envelope BOC (ACE-BOC) for the B2 signals [34, 35], and dual-QPSK for the B3 signals [36]. All of these new designs have been implemented on the newly launched next generation BDS satellites and are currently being tested and evaluated.

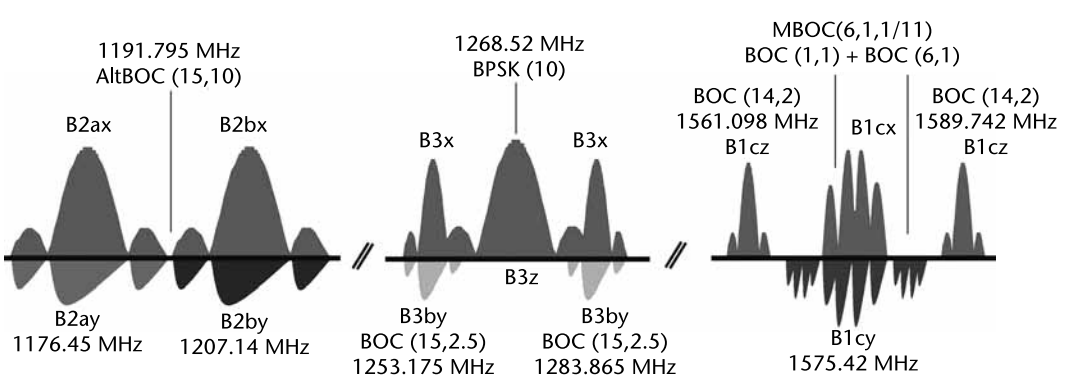


Figure 6.25 Spectrum of B1, B2, and B3 for the BDS global system [29].

QMBOC: A New Modulation for B1

The Quadrature Multiplexing BOC (QMBOC) technique modulates BOC(1,1) and BOC(6,1) signal components onto two orthogonal phases. This method avoids some problems that can arise when these two signal components are mixed together on one signal phase as in other MBOC variants. The power spectrum of QMBOC is identical to that of TMBOC. QMBOC can be processed as a BOC(1,1) signal by less-complex receivers that do not wish to take advantage of the performance enhancement possible from additionally processing the BOC(6,1) component. QMBOC provides excellent compatibility and interoperability with the open service signals of GPS and Galileo in the same frequency segment [32, 33].

From a receiver performance perspective, TMBOC and QMBOC have the same Gabor bandwidth, hence they provide identical receiver tracking performance. For simpler receivers that do not process the BOC(6,1) component, the sensitivity of QMBOC in signal acquisition and tracking is better than that of TMBOC. Also, as compared with TMBOC, QMBOC has advantages in terms of signal transmission flexibility; with QMBOC, since the BOC(6,1) component is orthogonal in phase with the BOC(1,1) component, the relative amplitudes of the two components can be freely adjusted without affecting the structure of fielded receivers. With respect to receivers, the orthogonal phases of BOC(6,1) and BOC(1,1) in QMBOC will be beneficial for low- to mid-range receivers since they can easily process only the BOC(1,1) component of the signal, sacrificing performance for lower cost while offering better interoperability with GPS L1C and Galileo E1 OS. High-end receivers would be expected to process the BOC(6,1) component to enhance tracking performance, particularly in multipath [32, 33].

TD-AltBOC and ACE-BOC: Multiplexing Techniques for B2

In order to provide interoperability with GPS L5 and Galileo E5, the BDS global system will broadcast broadband signals on two center frequencies, B2a (1,176.45 MHz) and B2b (1,207.14 MHz). The B2 signal components should enable better ranging capability and be robust against in-band interference. To minimize power required by the satellite transmitter and to reduce multiplexing losses, the B2a and B2b signals should be multiplexed into a constant envelope signal within the satellite transmitter (i.e., B2 centered at 1,191.795 MHz). Generating B2 in this manner allows flexibility on the receiver side. Receivers can individually track B2a or B2b as separate QPSK(10) signals, or treat the entire B2 as an ultrabroadband signal for high-performance applications.

TD-AltBOC and ACE-BOC are two possible methods to generate B2. TD-AltBOC time division multiplexes the data and pilot components of each B2 component (B2a and B2b) to generate two binary signals and then uses AltBOC to merge them together for transmission. Asymmetric Constant Envelope BOC (ACE-BOC) utilizes orthogonal data and pilot components to form an integral signal with constant envelope, which can support any power ratio among the components hence offering great flexibility in implementation. TD-AltBOC provides greater multiplexing efficiency (close to 100% for a receiver processing B2 as a single broadband signal) and is simpler to implement than ACE-BOC [34, 37]. However, the correlation properties of B2 generated using TD-AltBOC are degraded with respect to B2 generated using ACE-BOC. The nonideal cross-correlation characteristics of

TD-AltBOC may cause an inherent bias in a receiver's pseudorange measurement [38]. Further, to avoid a 50% signal-to-noise ratio (SNR) loss after the correlator within a TD-AltBOC receiver, each sideband of the TD-AltBOC signal should be processed individually as a TD-QPSK signal. In comparison, both the GPS L5 signal and the Galileo E5a/ E5b signal can be received as a QPSK signal. Therefore, for the multisystem receiver processing signals in the B2/L5/E5 band, it is not possible to receive the TD-AltBOC signal with the original correlator architecture supporting the GPS L5 and Galileo E5 signals, indicating less compatibility of TD-AltBOC with the other signals on the same band.

ACE-BOC avoids the problem caused by time division multiplexing within TD-AltBOC, by means of transmitting the data component and pilot component on each sideband on orthogonal phases. This allows ACE-BOC to allocate more power to the pilot components, which is desirable to improve the accuracy of pseudorange and carrier measurement as well as to improve the robustness of acquisition and tracking in low SNR conditions. The orthogonality between data and pilot components in ACE-BOC also provides better backward compatibility with earlier receiver designs. The orthogonality also provides for flexibility in the future in that it facilitates the future adjustment of the relative power of the B2 signal components, without significantly impacting fielded receivers. Lastly, ACE-BOC provides better interoperability with the GPS and Galileo signals in the same band [35]. A low-complexity implementation of ACE-BOC has been developed that is as simple to implement as AltBOC [39].

Dual-QPSK for New Signal

BDS B3, located at 1,268.52 MHz, is intended for authorized use only in the global stage of BDS. In the B3 band, a new modernized signal B3A will be broadcast in addition to the earlier B3 signal. Dual-QPSK [36] is a multiplexing technique that has been proposed to generate B3A and the earlier B3 signal within the satellite transmitter. Dual-QPSK solves the problem of combining a BOC(15,2,5) and a QPSK(10) component with equal power in the signal generator, where the BOC(15,2.5) signal has a data and a pilot channel with modulation phases orthogonal to each other. In addition, generalized Dual-QPSK supports flexible adjustment of the power among the signal components.

References

[1] China Satellite Navigation Office, "Development Report of BeiDou Navigation Satellite System (v. 2.2)," December 2013.

[2] Ran, C.Q., "Status Update on the BeiDou Navigation Satellite System (BDS)," Tenth Meeting of the International Committee on Global Navigation Satellite Systems (ICG), Boulder, Colorado, United States, November, 2015.

[3] International Committee on Global Navigation Satellite Systems (ICG): Members, http://www.unoosa.org/oosa/en/ourwork/icg/members.html.

[4] National Standardization Technical Committee of BeiDou Navigation Satellite System, China, "Standards for BeiDou Navigation Satellite System (v. 1.0)," November 2015.

[5] China Satellite Navigation Office, "BeiDou Navigation Satellite System Signal in Space Interface Control Document (v. 2.0)," December 2013.

[6] China Satellite Navigation Office, "Specification for Public Service Performance of Beidou Navigation Satellite System (v. 1.0)," December 2013.

[7] Fan, B.Y., Li, Z.H., and Liu, T.X., "Application and Development Proposition of BeiDou Satellite Navigation System in the Rescue of Wenchuan Earthquake," *Spacecraft Engineering*, Vol. 17, No. 4, 2008, pp. 6-13.

[8] Yu, H.X., and Cui, J.Y., "Progress on Navigation Satellite Payload in China," *Space Electronic Technologies,* No. 1, 2002, pp. 19-24.

[9] Chen, F.Y., et al., "The Development of Satellite Position Determination and Communication System," *Chinese Space Science and Technology*, No. 3, 1987, pp. 1-8.

[10] O'Neill, G.K., "The Geostar Position Determination and Digital Message System", *National Tele system Conference,* 1983, pp. 312-314.

[11] O'Neill, G.K., "The Geostar Satellite Navigation and Communications System", *the 40th ION Annual Meeting*, 1984, pp. 50-54.

[12] China Satellite Navigation Office, "BeiDou Navigation Satellite System Signal in Space Interface Control Document-Open Service Signal B1I (Version 1.0)," December, 2012.

[13] Xie, J., "Technology Development and Prospect of BeiDou Navigation Satellite," *Aerospace China,* No. 3, 2013, pp. 7-11.

[14] Fan, B.Y., "Satellite Navigation Systems and Their Important Roles in Aerospace Security," *Spacecraft Engineering,* Vol.3, No. 3, 2011, pp. 12-19.

[15] Hu, Z.G., "BeiDou Navigation Satellite System Performance Assessment Theory and Experimental Verification," Ph.D. Dissertation, Wuhan University, 2013.

[16] DFH-3. http://www.cast.cn/Item/Show.asp?m=1&d=2874, July, 2015.

[17] DFH-3a. http://www.cast.cn/Item/Show.asp?m=1&d=2875, July, 2015.

[18] Yang, Y.X., "Smart City and BDS," *The 8th China Smart City Development Technology Symposium*, Beijing, Oct. 2013.

[19] Liu J.Y., "Status and Development of the BeiDou Navigation Satellite System," *Journal of Telemetry Tracking and Command,* Vol. 34, No. 3, 2013, pp. 1-8.

[20] Yang, Y.X., "Chinese Geodetic Coordinate System 2000," *Chinese Science Bulletin*, Vol. 54, No. 15, 2009, pp. 2714-2721.

[21] Wei, Z.Q., " Chinese Geodetic Coordinate System 2000," *Journal of Geodesy and Geodynamics,* Vol. 28, No. 6, 2008, pp. 1-5.

[22] China Satellite Navigation Office, "Performance Requirements and Test Methods for BDS RDSS Unit," BD 420007-2015, Oct. 2015.

[23] Yang, Y. X., et al., "Preliminary Assessment of the Navigation and Positioning Performance of BeiDou Regional Navigation Satellite System," *Science China Earth Sciences*, Vol. 57, No. 1, 2014, pp. 144-152.

[24] Shen, J., "Development of BeiDou Navigation Satellite System (BDS): An Application Perspective," *The 10th Meeting of the International Committee on Global Navigation Satellite Systems,* November, 2015.

[25] Ren, J.T., "Capture Algorithm Research of Baseband Signal in Beidou Receiver," Master's Thesis, Hefei University of Technology, 2011.

[26] Yang, L., "Research and Design of Passive BeiDou System Timing Receiver," Master's Thesis, National University of Defense Technology, 2009.

[27] Jia, D.W., "Design and Implementation of Baseband Signal Processing of BeiDou System Receiver," Master's Thesis, Xidian University, 2011.

[28] China National Administration of GNSS and Applications (CNAGA), "COMPASS View on Compatibility and Interoperability," *ICG Working Group A Meeting on GNSS Interoperability*, July 2009, pp. 30-31.

[29] Tan, S.S., et al., "Studies of Compass Navigation Signals Design," *Scientia Sinica (Physica, Mechanica & Astronomica),* Vol. 40, No. 5, 2010, pp. 514-519.

[30] Lu, M.Q., "New Signal Structures for BeiDou Navigation Satellite System," *Stanford's PNT Challenges and Opportunities Symposium'2014*, Stanford, CA, 2014.

[31] Yao, Z., and Lu M.Q., *Design and Implementation of New Generation GNSS Signals*, Beijing, China: Publishing House of Electronics Industry, 2016.

[32] Yao, Z., Lu M.Q., and Feng Z.M., "Quadrature Multiplexed BOC Modulation for Interoperable GNSS Signals," *Electronics letters*, Vol. 46, No. 17, 2010, pp. 1234-1236.

[33] Yao, Z., and Lu M.Q., "Optimized Modulation for Compass B1-C Signal with Multiple Processing Modes," *Proceedings of ION GNSS Conference'2011*, Portland, OR, 2011, pp. 1234-1242.

[34] Tang, Z.P., et al., "TD-AltBOC: A New COMPASS B2 Modulation," *Science China Physics, Mechanics and Astronomy*, Vol. 54, No. 6, 2011, pp. 1014-1021.

[35] Yao, Z., and Lu M.Q., "Constant Envelope Combination for Components on Different Carrier Frequencies with Unequal Power Allocation," *Proceedings of ION ITM'2013*, San Diego, CA, 2013, pp. 629-637.

[36] Zhang, K., "Generalized Constant-Envelope DualQPSK and AltBOC Modulations for Modern GNSS Signals," *Electronics Letters*, Vol. 49, No. 21, 2013, pp. 1335-1337.

[37] Yan, T., et al., "Performance Analysis on Single Sideband of TD-AltBOC Modulation Signal," *Proceedings of China Satellite Navigation Conference (CSNC)'2013*, Springer, 2013, pp. 91-100.

[38] Liu, Y.X., et al, "Analysis for Cross Correlation in Multiplexing," *Proceedings* of *China Satellite Navigation Conference (CSNC)'2013*, Springer, 2013, pp. 81-90.

[39] Zhang, J.Y., Yao, Z., and Lu M.Q., "Applications and Low-complex Implementations of ACE-BOC Multiplexing," Proceedings of *ION ITM'2014*, San Diego, CA, 2014, pp. 781-791.

CHAPTER 7

Regional SATNAV Systems

Scott Feairheller and Brian Terrill

7.1 Quasi-Zenith Satellite System

7.1.1 Overview

The Quasi-Zenith Satellite System (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 providing high-elevation coverage to complement, augment, and be interoperable with the U.S. GPS (and potentially other GNSS constellations) over Japan. The high-elevation coverage is especially important in Japan where lower elevation GPS satellites are blocked by urban canyons and mountainous terrain. The first QZSS satellite is also providing experimental navigation and messaging services. By 2018, plans call for the QZSS constellation to expand to four satellites, and by 2023 the constellation is planned to consist of seven satellites that will provide independent regional capability in addition to complementing or augmenting other GNSS constellations [1, 2].

The QZSS Program was initiated in 2002 as a joint government-industry effort under the (then) Japanese Communications Research Laboratory contract. The Advanced Space Business Corporation (ASBC) team, including Mitsubishi Electric, Hitachi, and GNSS Technologies, worked on the concept until ASBC collapsed in 2007. In 2007, JAXA, the Satellite Positioning Research and Application Center (SPAC), and other organizations took over the work. The first QZSS satellite was launched in September 2010. In 2012, the Cabinet Office of Japan approved the launch of the next three satellites [1]. In 2015, the Cabinet Office of Japan approved the launch of three additional satellites by 2023 [2].

7.1.2 Space Segment

By December 2016, the constellation consisted of one satellite, QZS-1 or "Michibiki" (meaning guiding light) launched into inclined-elliptical-geostationary orbit (quasi-zenith) and placed over Japan. The figure-8 orbit is depicted in Figure 7.1.

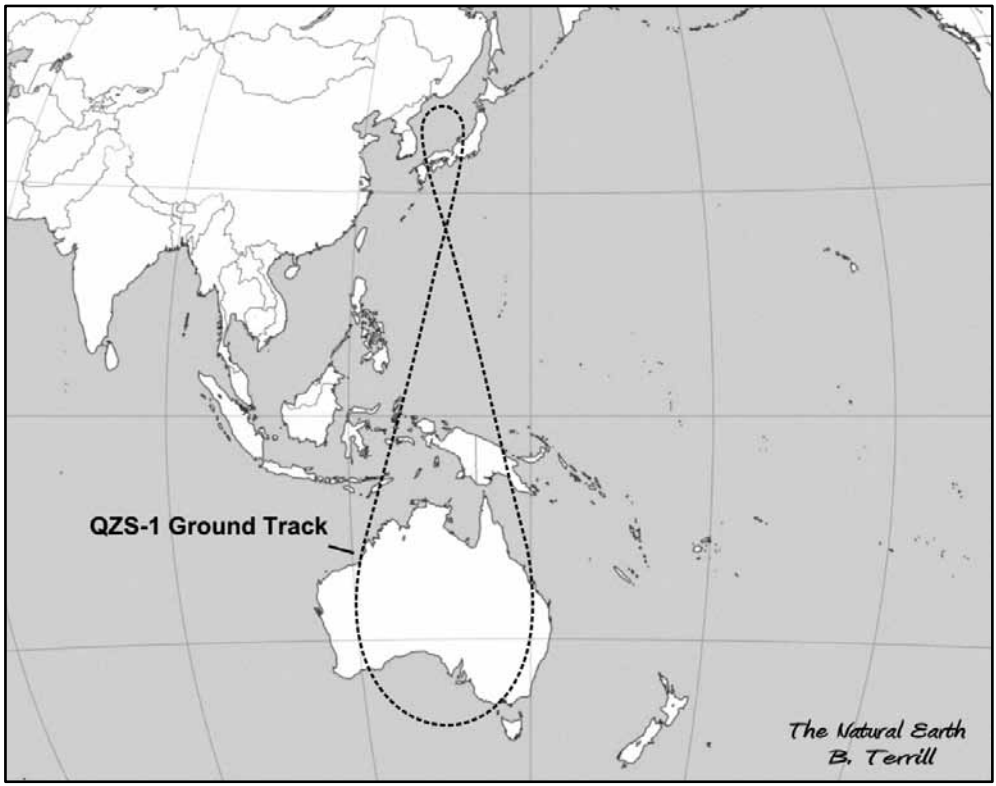


Figure 7.1 QZSS orbit [4]. (Courtesy of Brian Terrill.)

The orbit has a perigee of approximately 32,000 km, an apogee of approximately 40,000 km, and an orbital inclination of approximately 40° [3].

7.1.2.1 Constellation

As planned, by 2018, the constellation will consist of three satellites in quasi-zenith orbit and one in geostationary orbit (GSO). The three quasi-zenith orbit satellites will share the same figure-8 ground trace and the geostationary satellite will be placed at 127° longitude, offset from the figure-8 [5]. By 2023, the QZSS space segment will consists of seven satellites launched into quasi-zenith and geostationary orbits over Japan [5]. The orbits of the additional three satellites remain unspecified. Future satellites (QZS-2, QZS-3, and so forth) will also be nicknamed Michibiki with no additional numbers added to distinguish them [6]. The launch history and plans are listed in Table 7.1.

7.1.2.2 Spacecraft

QZS-1 was designed by Mitsubishi Electric Corporation (MELCO) Kamakura Works [7, 8]. The spacecraft design is based on the Japanese Engineering Test Satellite-8 (ETS-8) which uses the Mitsubishi DS2000 standardized bus [9, 10]. Follow-on spacecraft will use a similar design, but may carry additional payloads and capabilities. The Japanese plan to procure the follow-on spacecraft in two batches

Table 7.1 QZSS Launch History and Plans

Spacecraft	*Name*	*Launch Date*	*Orbit Type*	*Equatorial Crossing*
QZS-1	Michibiki	September 11, 2010	Inclined-elliptical-geosynchronous	Around 132° and 140°
QZS-2	Michibiki	Planned in 2017	Inclined-elliptical-geosynchronous	Around 132° and 140°
QZS-3	Michibiki	Planned in 2017	Inclined-elliptical-geosynchronous	Around 132° and 140°
QZS-4	Michibiki	Planned in 2018	Geostationary	127°
QZS-1R	Michibki	Planned in 2020	Inclined-elliptical-geosynchronous	Around 132° and 140°
QZS-5	Michibiki	Between 2020 and 2023	Inclined-elliptical-geosynchronous	?
QZS-6	Michibiki	Between 2020 and 2023	Inclined-elliptical-geosynchronous	?
QZS-7	Michibiki	Between 2020 and 2023	Geostationary	?

Source: [10, 11]. QZS-5 or QZS-6 could be placed in geostationary orbit instead of QZS-7.

of three satellites each allowing for incorporation of upgrades. The QZS-1 spacecraft configuration is shown in Figure 7.2 [3].

7.1.2.3 Bus

The DS2000 spacecraft bus is designed to boost the satellite to its final orbit, support the mission payloads, and maintain the satellite in the proper orbit. The DS2000 is a 3-axis stabilized spacecraft measuring 2.9m × 3.0m × 6.0m, with two solar panels measuring 25.3m tip-to-tip [11, 12]. The spacecraft used a central core design with reinforced carbon-fiber plastic panels to form the body. The DS2000 has a design lifetime of at least 10 years. Each spacecraft has a lift-off mass of 4,100 kg and dry mass of 1,800 kg, which supports a 320-kg navigation payload [12]. The DS2000 spacecraft bus consists of a number of subsystems.

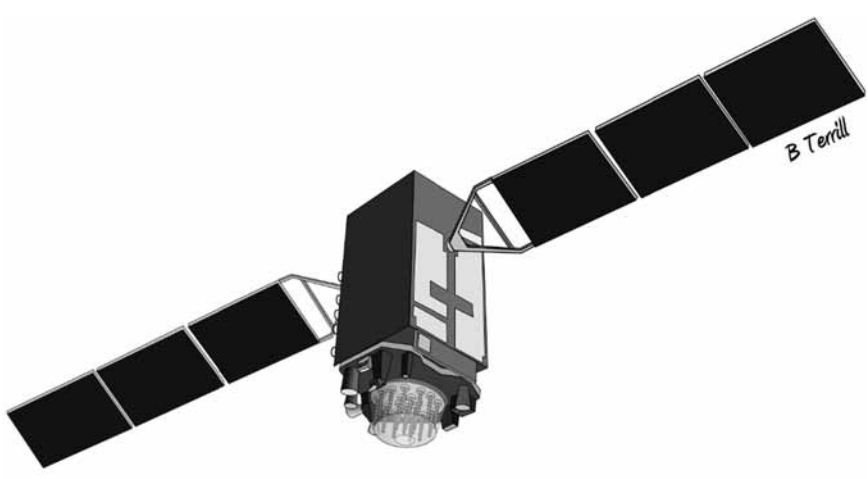


Figure 7.2 QZS-1 spacecraft [4]. (Courtesy of Brian Terrill.)

7.1.2.4 Electrical Power Subsystem

This provides 15 kW at the beginning of life and 5.3 kW at the end of life. The power system is composed of high-efficiency silicon, multijunction GaAs solar cells, Li-Ion batteries, and two independent electrical bus systems for redundancy [12].

7.1.2.5 Thermal Control Subsystem

This provides passive control of the satellite by radiating energy into space. The subsystem is composed of heat pipe-embedded payload panel, optical solar reflector (OSR), thermal blankets, and heater. Some of the heat pipes are dedicated to the rubidium clocks to maintain the temperature at 20°C ± 5°C and unaffected by solar equipment and other spacecraft equipment [12].

7.1.2.6 Propulsion Subsystem

This propels the satellite from transfer orbit to the quasi-zenith or geostationary orbits after their separation from the launch vehicle, maintains spacecraft attitude control and supports station-keeping. The subsystem consists of a 500-N apogee kick motor, attitude thrusters, and a bipropellant fuel system [12].

7.1.2.7 Onboard Control System

This provides satellite management, data handling, and attitude control for QZS-1 using an ARINC 1553B bus protocol. The subsystem must maintain the solar panels pointing at the Sun and the nadir deck (i.e., face of the satellite) pointing at the Earth while maintaining the thermal stability of the satellite. The attitude control part of the subsystem consists of two star trackers (STT), two redundant sets of three Fine Sun Sensor Heads (FSSH), two redundant Fine Sun Sensor Electronics (FSSE), two Earth Sensor Assemblies (ESA), one internally redundant inertial reference unit (IRU), and two redundant onboard computers called the satellite controller (SC). The subsystem has a pointing accuracy of $< \pm 0.05°$ in roll and $< \pm 0.15°$ in yaw. The satellite management parts of the subsystem autonomously maintain knowledge of the Sun's direction and generate commands to the attitude control components. The subsystem includes an onboard orbit propagator and attitude profile generator [12].

7.1.2.8 Telemetry Tracking and Command (TT&C)

This supports commanding and navigation messaging uploading to the satellite. Normally, the QZS-1 uses 4-Kbps C-band links to support the command and control (C2) and navigation payload. For launch and early orbit operation and emergency backup, the QZS-1 uses S-band C2 links [12].

7.1.2.9 Navigation Payload

The QZS-1 navigation payload is designed to generate and transmit six navigation signals: L1C/A (1,575.42 MHz), L1C (1,575.42 MHz), L2C (1,227.6 MHz), L5

(1,176.45 MHz), L1S (1,575.42 MHz), and L6 (1,278.75 MHz). (These signals are described in Section 7.1.6.) The navigation payload consists of three subsystems: the L-band signal transmission (LTS), time transfer subsystem (TTS), and laser retro-reflector assembly (LRS). The LTS consists of the onboard rubidium atomic clock, navigation onboard computer, various electronics, time-transfer comparison unit, L1S antenna, and L-band navigation helical array antenna. The LTS consists of two identical electronic chains. The TTS consists of the time comparison unit and Ku-band bidirectional comparison antenna. The payload allows the onboard clocks to be compared to ground clocks using two-way techniques [i.e., two-way satellite time and frequency transfer (TWSTFT)]. The LSR consists of 56-corner cube reflectors and allows the satellite to be tracked using two-way laser-ranging techniques under cloud-free conditions [13]. Additional electronics and signals will be added to future satellites.

In the future, GEO satellites will carry a navigation payload which will also transmit L1Sb (1,575.42 MHz) for an operational SBAS service and L5S (1,176.45 MHz) for experimental augmentation service. At the time of this writing, little information was available on the future QZS spacecraft designs [2].

7.1.3 Control Segment

The QZSS ground segment consists of master control stations (MCS), monitor stations (MS), satellite tracking control stations (TCS), laser-ranging stations (LRS), and time management stations (TMS). The ground network is shown in Figure 7.3. The ground system is detailed next.

7.1.3.1 Master Control Station (MCS)

In 2016, the MCS was located at the Tsukuba Space Center near Tokyo and was the focal point for the navigation and other satellite missions. The MCS determines and propagates the satellite orbit and satellite clock offset predictions, generates the navigation uploads, determines navigation integrity, plans navigation experiments, analyzes the system performance, and stores data associated with the system performance. For reliability, the MCS maintains a hot-redundant capability [12]. The Japanese plan to build two new MCSs at Hitachiota (near Tokyo) and at Kobe [14].

7.1.3.2 Tracking Control Station (TCS)

In 2016, the TCS was located in Okinawa and was responsible for uploading the ephemeris and clock corrections to the QZS-1 navigation computer, monitoring the spacecraft status and sending command and control signals to the satellite during the operational phase of QZS-1. During launch, a number of other TCS sites around the world supported QZS-1 until it reached its final orbit. The TCS maintains continuous contact with QZS-1 using the C-band uplinks and downlinks, because Okinawa is located near the equator and has visibility to the satellite throughout the entire orbit. In addition, about once per year, the TCS makes an orbit adjustment to maintain QZS-1 in the proper orbit. Additional, TCS sites are planned at Hitachiota (near Tokyo) and Tanegashima, Kumejima, Ishigakijima, and Miyakojima on the island chain in southern Japan [14].

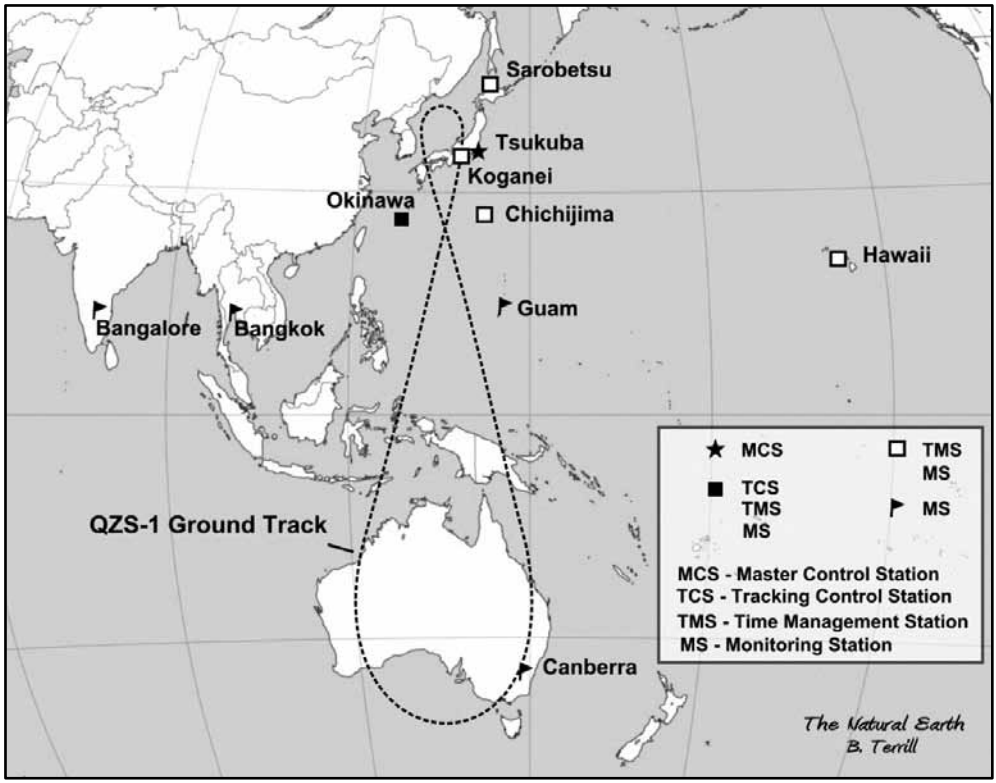


Figure 7.3 QZSS ground support network [4]. (Courtesy of Brian Terrill.)

7.1.3.3 Monitor Stations (MS)

QZSS currently maintains about 12 monitoring-stations located throughout the Asia-Pacific region. The MS stations collect QZS-1 and GPS navigation signals for precise estimation of QZS-1 orbit and satellite clock parameters to correlate with those of GPS. In addition, the sites also collect environmental data for forwarding to the MCS. The MS support several levels of capability. The MS sites at Okinawa, Sarobetsu (Soya), Koganei (Tokyo) and Chichijima (Ogasawara) in Japan and Bangalore (India), Canberra (Australia), Bangkok (Thailand), Guam (United States), and Hawaii (United States) monitor the QZS-1 and GPS satellites via the L-band navigation signals. In addition, Perth (Australia), Maspalomas, (Gran Canaria, Las Palmas, Spain), and Santiago (Chile) monitor only GPS satellites via navigation signals.

7.1.3.4 Time Management Station (TMS)

The TMS sites at Okinawa and Koganei (Tokyo) use the Ku-band transponder on the QZS-1 to perform two-way ranging to the satellite. The TMS sites at Okinawa, Sarobetsu (Soya), Koganei, and Chichijima (Ogasawara) and Hawaii use TWSTFT to perform time transfer between sites [15, 16].

7.1.3.5 Laser-Ranging Stations (LRS)

A number of international LRS sites support the QZS-1 mission by two-way ranging to the satellite under cloud-free weather conditions. Sites in Beijing, Changchum, Koganei (Tokyo), Mount Stromlo, Shanghai, Tamegashima, and Yarragadee routinely track the satellite under the protocol established by the International Earth Rotation Service.

7.1.4 Geodesy and Time Systems

The QZSS system provides navigation and positioning information in the Japanese Satellite Navigation Geodetic System (JGS). The JGS is based on the regional sites and measurements that support the Japanese contribution to the International Terrestrial Reference Frame (ITRF). The JGS is maintained to within less than 0.2m of WGS-84 [17].

Historically, Japan has used the Tokyo-1927 datum for mapping and surveying. Continued use of the older map products would require a corresponding transformation between JGS and Tokyo-1927 to be useful with QZSS user equipment. At the time of this writing, it was unclear how much of mapping and other documentation in Japan still relies on Tokyo-1927 datum.

The QZSS system uses QZSS time (QZSST). QZSST is maintained by the master clock at the MCS in Tsukuba Space Center. QZSST is similar to GPS time and operates continuously independent of leap seconds. By 2016, QZSST was 17 seconds ahead of UTC and 19 seconds behind international time (TAI). The slight offset between QZSST and GPS time is maintained to within 7 ns (or 2.0m 95%) [17].

7.1.5 Services

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 independent regional navigation capability independent of GPS and other GNSS constellations in addition to the current services. That is, there will be four or more QZSS satellites in view for users in the coverage region to obtain PVT information.

Currently, QZS-1 provides operational services which are being used for a variety of applications in Japan and experimental services which are being tested for future operational use. QZS-2 through QZS-4 will add new operational augmentation and experimental augmentation services. Satellites in GEO will provide SBAS corrections, experimental augmentation and S-Band messaging services. Satellites in the quasi-zenith orbit will also provide experimental and centimeter augmentation services. No information is available on navigation or messaging services to be provided by QZS-5 through QZS-7. The navigation and augmentation charges are offered free of any user fees. An overview of the services is provided in Table 7.2.

Table 7.2 Planned QZSS Services

Services	*Signal*	*Frequency*	*QZS-1*	*QZS-2*	*QZS-3*	*QZS-4*
Positioning complement to GPS	L1-C/A	1,575.42 MHz	X	X	X	X
Positioning complement to GPS	L1C		X	X	X	X
L1S submeter augmentation	L1S		X	X	X	X
Crisis Messaging Service	L1S		X	X	X	X
ICAO Standard SBAS	L1Sb					X
Positioning complement to GPS	L2C	1,227.60 MHz	X	X	X	X
Positioning complement to GPS	L5	1,176.45 MHz	X	X	X	X
Experimental augmentation	L5S			X	X	X
Centimeter-level augmentation Service	L6	1,278.75 MHz	X	X	X	X
Safety Messaging Service	S-band	2 GHz				X

7.1.5.1 Navigation Services

QZSS transmits a number of navigation signals to complement GPS; these include: L1-C/A (1,575.42 MHz), L1C (1,575.42 MHz), L2C (1,227.6 MHz), and L5 (1,176.45 MHz). The signals provide ranging error URE of 1.6m (95%) including the GPS-QZSS time and coordinate system biases. For single-frequency users (L1-C/A and QZSS L1-C/A), the horizontal accuracy is about 21.9m (95%). For dual-frequency users using L1 and L2, the horizontal accuracy is about 7.5m (95%). The services improve the reliability by providing failure monitoring and reporting system health problems.

7.1.5.2 Augmentation Services

QZSS currently transmits or will add in the future navigation signals designed to improve GPS accuracy. These include: L1S (1,575.42 MHz), which is designed to provide submeter corrections and be interoperable with GPS and other SBAS, and L6 (1,278.75 MHz), which is an experimental signal designed to provide high-precision service and compatible with the Galileo Commercial Service signal. The L1S provides wide-area differential correction data with positioning accuracy of 1-m horizontal RMS except in cases of large multipath error and large ionospheric disturbance. The L1S also provides submeter corrections but is under policy examination by the Japanese SPAC [1, 17, 18]. L6 is an experimental signal for high precision at 3-cm horizontal RMS level service. The policy on L6 is also under examination.

7.1.5.3 Messaging Services

QZSS currently transmits an experimental messaging service called the Satellite Report for Disaster and Crisis Management on the L1S signal. The service provides messages to users warning of disasters like earthquakes, tsunamis, volcanoes, fires, or explosion of factory or atomic power plant and warnings of rescue from terrorist attack or accidents. The Japanese are investigating extending these services to overseas users [19].

7.1.6 Signals

The QZSS satellites will transmit up to six navigation signals. The external characteristics of the signals are listed in Table 7.3.

The internal characteristics of the navigation signals were detailed in the QZSS Interface Control Document (ICD) designated as IS-QZSS version 1.6 in 2016. The most recent ICD can be found at the following link: http://qz-vision.jaxa.jp/USE/is-qzss/index_e.html [4].

7.1.6.1 QZS-L1-C/A, QZS-L1C, QZS-L2C, and QZS-L5

The QZSS satellites transmit signals very similar to (but not identical to) the modernized GPS L1-C/A, L1C, L2C, and L5 civil signals with additional QZSS related messages. The Japanese chose these signal designs and message structure to maximize interoperability with GPS. Details on these GPS signals are covered in Chapter 3 [4]. The additional message modifications to the QZSS signals are detailed in respective sections of the QZSS ICD and some details are covered next.

The QZSS satellites transmit PRNs in the same families of spreading codes as used by the GPS signals. The PRNs assigned to first five operational QZSS will be 193–197 for the QZS-L1C/A signals. PRN code numbers 198–202 are reserved for testing or maintenance. QZS-1 is using PRN 193.

7.1.6.2 QZS L1S

The QZS L1S is a submeter GNSS correction message designed to use the same data structure as used with the GPS-SBAS. L1S stands for L1-Submeter-class Augmentation with Integrity Function. It also improves the reliability of GNSS by providing system health and failure notifications [4, 20].

7.1.6.3 QZS L1S Signal Modulation

The L1S signal is modulated using BPSK-R similar to the GPS C/A code and depicted in in Section 3.7.1 [4].

Table 7.3 External Characteristics of the QZSS Navigation Signals

Signal	*Channel*	*Frequency*	*Bandwidth*	*Minimum Received Power*
QZS-L1C	L1CD	1,575.42 MHz	24 MHz	–163.0 dBW
	L1CP		24 MHz	–158.25 dBW
QZS-L1-C/A			24 MHz	–158.5 dBW
QZS-L1S			24 MHz	–161.0 dBW
QZS-L2C		1,227.60 MHz	24 MHz	–160.0 dBW
QZS-L5	L5I	1,176.45 MHz	25 MHz	–157.9 dBW
	L5Q		25 MHz	–157.9 dBW
QZS-L6		1,278.75 MHz	39 MHz	–155.7 dBW

Source: [4].

7.1.6.4 QZS L1S Code Properties

The QZS L1S transmit signals using PRNs 183–187 for the first five satellites with 188–192 as spares for additional satellites. The shift register design is the same as GPS C/A code depicted in Figure 3.36 [4].

7.1.6.5 QZS L1S Message Structure

The QZS L1-SAIF also called the Sub-meter Level Augmentation Service (SLAS) is transmitted in various message types formatted in 250-bit data message frames at one data frame per second. Each frame is composed of 8-bit preamble, 6-bit message type, 212-bit message type, and 24-bit CRC. The basic structure is depicted in Figure 7.4 [4].

The QZSS signal specification defines some of the messages types for QZS L1S. The message types currently defined are listed in Table 7.4 [4].

7.1.6.6 QZS L6

The QZS-L6, also called the Centimeter Level Augmentation Service (CLAS), is a high-data-rate 2-Kbps GNSS correction message for PPP and RTK applications. It is transmitted at 1,278.75 MHz and is compatible and interoperable with the Galileo E6 Commercial Service (CS) signal [4, 20].

7.1.6.7 QZS L6 Signal Modulation

The L6 signal is generated using BPSK-R spreading modulation with a chipping rate of 5.115 Mchips/s. The underlying data is modulated code shift keying [4].

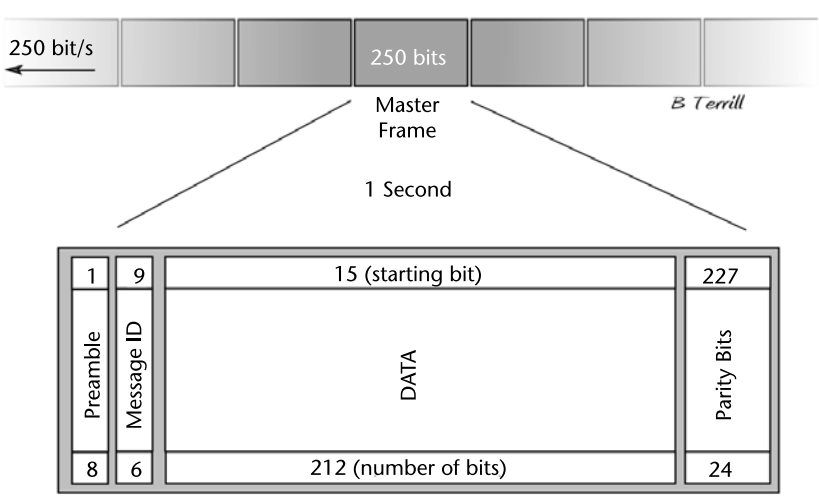


Figure 7.4 QZS L1S message structure [4]. (Courtesy of Brian Terrill.)

Table 7.4 Defined QZS-L1S Message Functions

Message Type	*Message Function*
TYPE 0	Test Mode
TYPE 1	PRN Mask
TYPE 2–5	Fast Correction and UDRE
TYPE 6	Integrity Data
TYPE 7	Fast Correction Degradation Factor
TYPE 10	Degradation Parameter
TYPE 12	Timing Information
TYPE 18	Ionospheric Grid Point Mask
TYPE 24	Fast Long-term Correction
TYPE 25	Long-term Correction
TYPE 26	Ionospheric Delay and GIVE
TYPE 28	Clock-Ephemeris Covariance
TYPE 40–51	Reserve for Demonstrations like L1-SAIF
TYPE 52	Tropospheric Grip Point Mask
TYPE 53	Tropospheric Delay Correction
TYPE 54–55	Undetermined Atmospheric Delay Information
TYPE 56	Intersignal Correction Bias Information
TYPE 57	Reserved for Undetermined Orbital Information
TYPE 58	QZS Ephemeris Data
TYPE 59	Undetermined QZSS Almanac Data
TYPE 60	Undetermined Regional Information/Maintenance Schedule
TYPE 62	Reserve for Inertial Test
TYPE 63	Null Message

Source: [4].

7.1.6.8 QZS L6 Code Properties

The L6 signal is generated by combining a short and long Kasami code. The short code is transmitted at 2.5575 MChip/s with a length of 10,230 chips and a 4-ms period and the long code is transmitted at 2.5575 MChip/s with a length of 1,048,575 chips and a 410-ms period. Figure 7.5 shows the shift register layout used to generate the L-band experiment (LEX) spreading codes [4].

7.1.6.9 QZS L6 Message Structure

The L6 signal is transmitted in various message types formatted in 2,000-bit data message frames at one data frame per second. Each frame is composed of 32-bit preamble, 8-bit PRN, 8-bit message type, 1-bit alert-flag, 1,695-bit data message, and 256-bit Reed-Solomon error correction code. The basic structure is depicted in Figure 7.6 [4].

The QZSS Signal Specification defines some of the messages types for QZS L6. The message types currently defined are listed in Table 7.5 [4].

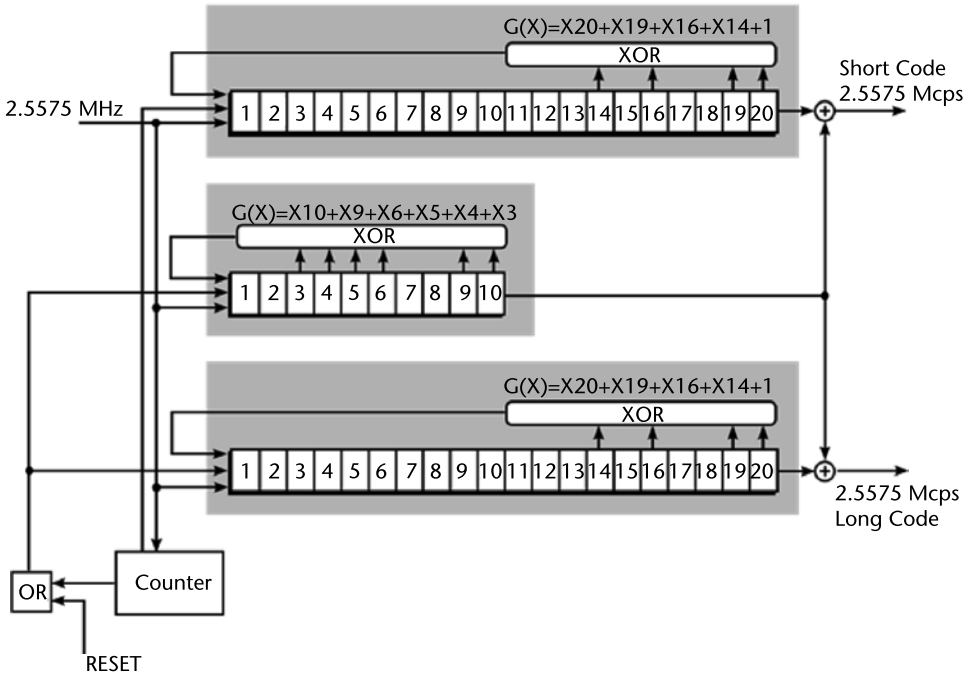


Figure 7.5 QZSS shift register design for SPS spreading code [21]. (Courtesy of Brian Terrill.)

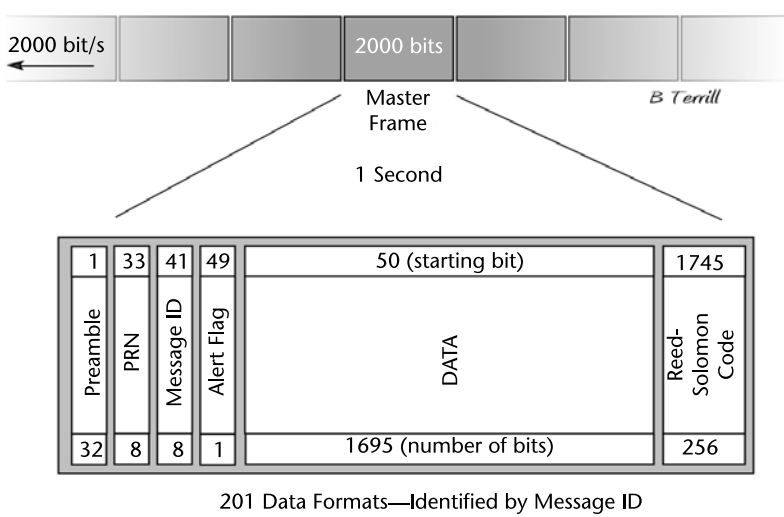


Figure 7.6 QZSS L6 navigation message structure [4]. (Courtesy of Brian Terrill.)

7.1.6.10 QZS Safety Messages

QZS also transmit a Crisis Short-Message Service on the L1S signal and Safety Messaging Service in the S-band at 2 GHz. However, additional details on the data structure are currently not available in the public version of the QZSS ICD.

Table 7.5 Defined QZS L6 Message Functions [4]

Message Type	*Message Function*
TYPE 0–9	Spare for System Use
TYPE 10	Signal health for 32 GPS satellites and 3 QZS satellites, ephemeris, and SV clocks
TYPE 11	Signal health for 32 GPS satellites and 3 QZS satellites, ephemeris, SV clocks, and ionospheric corrections
TYPE 12	Orbit and clock corrections, URA, and SV bias for MADOAC-LEX for PPP-AR
TYPE 13–19	Spare
TYPE 20–155	For experimental or application demonstration use by public sector (except JAXA)
TYPE 156–200	For experimental or application demonstration use by JAXA or the private sector

Source: [4].

7.1.6.11 QZS TT&C Signals

For normal TT&C operations, the QZS-1 uses C-band (5,000–5,010-MHz uplink and 5,010–5,030 MHz downlink) links to support the C2 and navigation payload. For LEOP, QZS-1 can also use S-band (2,025–2,110 MHz uplink and 2,200–2,290-MHz downlink) [4].

7.1.6.12 Applications and User Equipment

QZSS is designed to support a variety of civil applications. The proposed civil applications include enhanced navigation for car, rail, and bus navigation and positioning for mapping, construction work, monitoring services for children and senior citizens, personal navigation for disabled and aged persons, automatic control of agricultural machinery, detecting earthquakes and volcanic activities, weather forecasting, search and rescue, and many other applicable fields. At the time of this writing, over 234 companies and institutes were participating in the QZSS program and researching 105 themes of applications [22, 23].

7.2 Navigation with Indian Constellation (NavIC)

7.2.1 Overview

Navigation with Indian Constellation (NavIC), formerly known as the Indian Regional Navigation Satellite System (IRNSS), 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) [23–25]. IRNSS was renamed NavIC, meaning "boatman," in conjunction with the completion of the satellite constellation in April 2016 [26, 27]. NavIC is different than most other SATNAV systems in that it provides only regional coverage, and it transmits navigation signals in both the L5-band and S-band, while other navigation systems work primarily in the L-band [28]. The Chinese Beidou-2 is the only other system currently operating its regional component using navigation signals in the S-band [29].

The Indian government initiated the IRNSS program in 2006 [24, 30, 31]. The first NavIC satellite was launched in July 2013 [32]. By 2016, the NavIC system

consisted of seven geostationary and inclined geosynchronous satellites, ground support segment, and user equipment. The NavIC provides a position and navigation accuracy to better than 20m (2σ) horizontal accuracy, and timing accuracy support better than 20 ns (2σ) 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 1,500 km around India and covering the strategic important sea routes in the Arabian Sea, Indian Ocean, Bay of Bengal, and South China Sea [33–35].

The ISRO plans to conduct an internal review to evaluate the NavIC constellation once it is complete. After completion, the Indian government will consider options like upgrading the constellation from 7 to 11 spacecraft or initiating the development of the next-generation satellite [36], which will possibly add the L1-band [37, 38] navigation signals in addition to L5 and S-bands.

7.2.2 Space Segment

The NavIC space segment consists of seven satellites placed over India into geostationary orbit (GSO) and inclined geostationary orbit (IGSO). The three GSO satellites are located at 32.5° East, 83° East, and 131.5° East in the geostationary belt. Two IGSO satellites have a longitude crossing of 55° East and orbital inclination of 29°. The other two IGSO satellites have a longitude crossing of 111.75° East and orbital inclination of 29°. The arrangement of the constellation is shown in Figure 7.7.

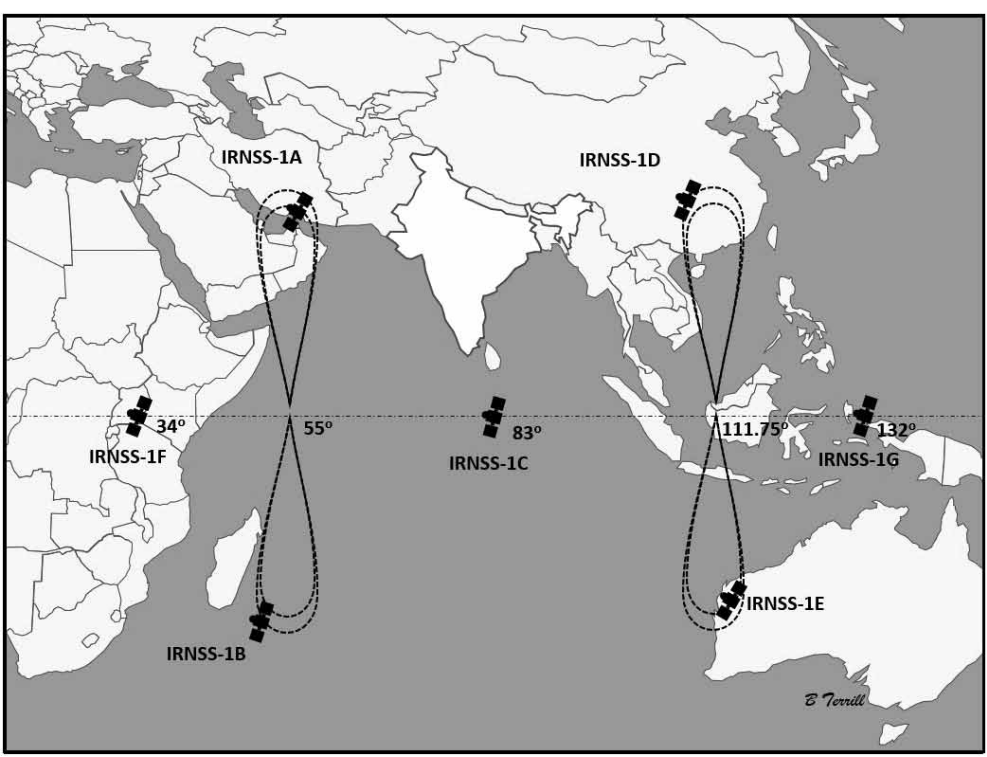


Figure 7.7 NavIC orbital constellation. (Courtesy of Brian Terrill.)

7.2.2.1 Spacecraft

All the NavIC satellites are identical and use the same I-1K spacecraft bus design, and carry the same mission payloads. Each satellite is designed with a 10-year lifetime [41]. The spacecraft configuration is shown in Figure 7.8.

7.2.2.2 Bus

The spacecraft bus is designed to boost the satellite in to its final orbit, support the mission payloads and maintain the satellite in the proper orbit. Each spacecraft has a lift-off mass of 1,425 kg and dry mass of 603 kg, which allows it to be launched onboard the Polar Space Launch Vehicle (PSLV), the smaller of the of available Indian indigenous launch vehicles. The I-1K is a 3-axis stabilized spacecraft, measuring 1.58m × 1.50m × 1.50m with two solar panels, Liquid Apogee Motor (LAM), and number of smaller thrusters [28, 42, 43].

The I-1K bus consists of a number of subsystems. The Attitude and Orbit Control System (AOCS) uses yaw steering to maintain the solar panels pointing at the Sun, thermal control of the satellite, and the navigation antenna pointing towards

Table 7.6 NavIC Launch History

Spacecraft	*Launch Date*	*Orbit Type*	*Equatorial Crossing*
IRNSS-1A	July 1, 2013	Inclined geosynchronous	55°
IRNSS-1B	April 4, 2014	Inclined geosynchronous	55°
IRNSS-1C	October 15, 2014	Geostationary	83°
IRNSS-1D	March 27, 2015	Inclined geosynchronous	111.75°
IRNSS-1E	January 20, 2016	Inclined geosynchronous	111.75°
IRNSS-1F	March 10, 2016	Geostatioary	32.5°
IRNSS-1G	April 28, 2016	Geostationary	131.5°

Source: [32, 39–41]. Note that while IRNSS was renamed NavIC, the satellites are still designated as IRNSS SVs. IRNSS-1A failed (clock failure) shortly after IRNSS-1G was launched. A replacement satellite is planned for launch in late 2017.

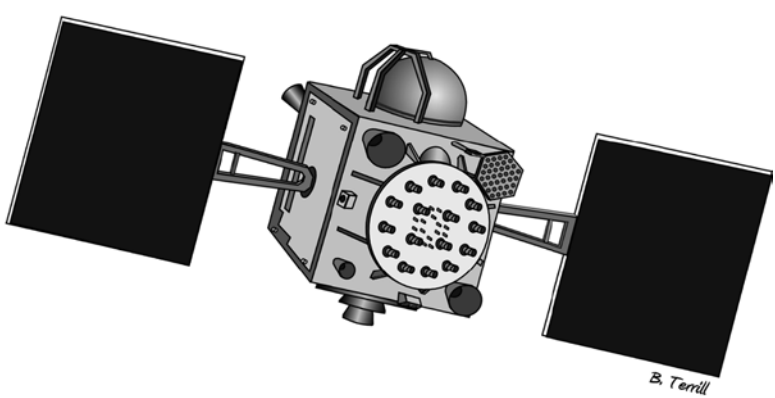


Figure 7.8 NavIC spacecraft. (Courtesy of Brian Terrill.)

the Earth. The AOCS consists of gyroscopes, reaction wheels, magnetic torquers, and solar and star sensors. The propulsion system consists of a 440-N LAM, twelve 22-N thrusters, and liquid fuel tanks. The power subsystem consists of two ultra-triple-junction solar cells and solar panels generating up to 1,660W of electricity and one 90-A/Hr lithium-ion battery [39, 42–44].

7.2.2.3 Payloads

The NavIC spacecraft carry several navigation and ranging payloads. The navigation payload transmits the restricted (military) and civil navigation services in both L5-band (1,176.45 MHz) and S-band (2,492.028 MHz) to users. The navigation payload consists of the Navigation Signal Generation Unit (NSGU), unspecified number of RAFS, and phased array antenna. The ranging payloads consist of a C-band transponder to support two-way CDMA ranging, and Corner Cube Retro Reflectors to support laser ranging [42, 44, 45].

7.2.3 NavIC Control Segment

The NavIC ground segment currently consists of 15 stations, with 6 more planned. All of the stations are located within India including the MCC at Hassan, Karnataka. The site distribution is shown in Figure 7.9 and the status as of 2016 is in Table 7.7 [41–43].

The NavIC spacecraft are maintained by two independent ground networks: the IRNSS Satellite Control Facility (IRSCF), which commands and controls the spacecraft and provides housekeeping functions, and the IRNSS Navigation Control Facility (IRNCF), which supports the navigation payload [40, 42].

7.2.3.1 IRNSS Satellite Control Facility (IRSCF)

The IRSCF consists of two IRNSS Satellite Control Centers (IRSCC) and nine IRNSS TT&C and Land Uplink Stations (INLUS) colocated at these sites. The IRSCC commands the NavIC satellites, and collects housekeeping telemetry from the satellites in the NavIC constellation via the INLUS antennas and support facilities. The IRSCC main station is located at Hassan, and a future backup station is being built at Bhopal. The main station has one 11-m antenna and four 7.2-m antennas for commanding the satellites. The main station supports Launch and Early Orbit Phase (LEOP), In-Orbit Test (IOT), and operational support to the constellation. When complete, the backup station will have one 11-m antenna and three 7.2-m antennas for control of NavIC [42, 43].

7.2.3.2 IRNSS Navigation Control Facility (IRNCF)

The IRNCF consists of two IRNSS navigation centers (INC), 12 IRNSS range and integrity monitoring stations (IRIMS), two network timing facilities (IRNWT), four IRNSS CDMA ranging stations (IRCDR), IRNSS laser ranging service (ILRS), and two data communication networks [42, 43].

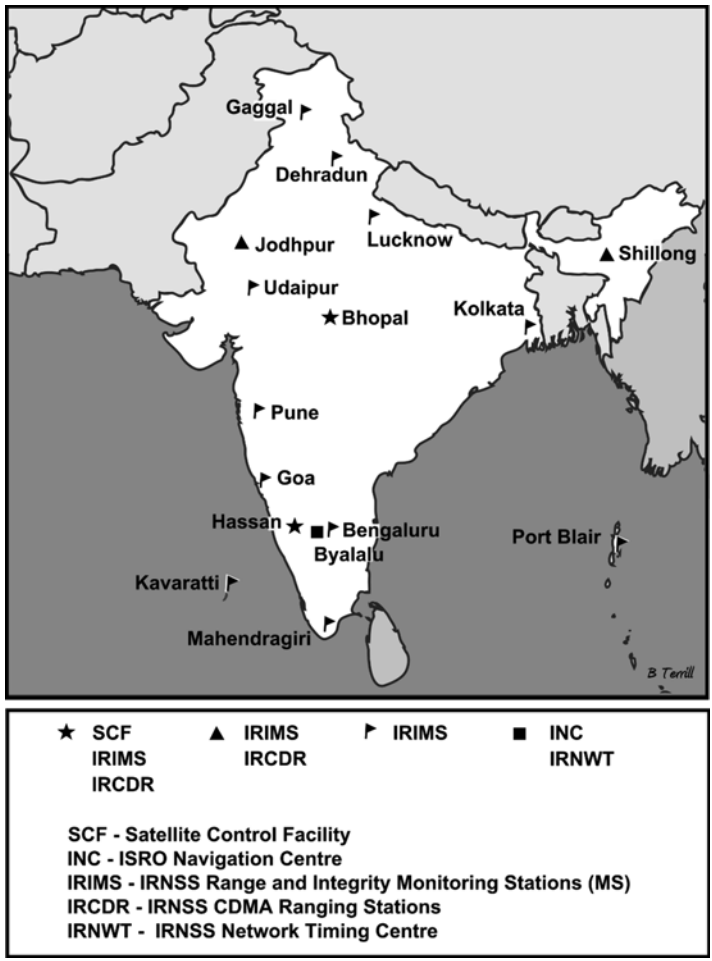


Figure 7.9 NavIC ground support segment [42]. (Courtesy of Brian Terrill.)

Table 7.7 NavIC Ground Site Status as of 2016

Type	*Current*	*Planned*
IRIMS	12	+3
IRNWT	1	+1
INC	1	+1
SCF	1	+1
Total sites	15	21

Source: [41]. Note that the ground control segment components are still designated as IRNSS components.

7.2.3.3 INC

The INC consists of two sites in India, which are mission support centers for navigation payloads. INC-1 is located in Byalalu, near Bengaluru, India, and has provided operational support to both the military and civil navigation payloads since

August 1, 2013. INC-2 was under development in 2015 and is located in Lucknow, India. When operational, the site will act as a backup for civil navigation services and house a backup timing facility [42, 43].

7.2.3.4 IRIMS

The IRIMS consists of 12 sites, which perform one-way ranging of the satellites and communicate the ranging measurements to the INC-1 in real time. Currently, 12 sites are operational and up to four more sites are planned [42, 43].

7.2.3.5 IRCDR

The IRCDR consists of four sites located at Bhopal, Hassan, Jodhpur, and Shillong, which perform two-way CDMA ranging and communicate the ranging measurements to the INC-1 in real time. The ranging measurements are communicated to the INC for processing [42, 43].

7.2.3.6 IRNWT

The IRNWT maintains the time scales for the IRNSS system with an ensemble of cesium and hydrogen maser clocks. The primary time facility is located with the INC-1 at Byalalu. The IRNSS time scale is maintained at a level of about 20 ns (2σ) with respect to UTC [42, 43].

7.2.3.7 IRDCN

The IRDCN provides dedicated communications support to the IRNSS network. The networks consist of terrestrial communication links between INC-1, the four IRCDR two-way ranging stations and the 12 one-way IRIMS. In the future, very-small-aperture terminal (VSAT) links will be added after receiving the necessary regulatory clearances [41–43].

7.2.3.8 IRLRS

The laser ranging service will track the IRNSS satellites using two-way laser ranging techniques and the retro-reflector onboard the satellites under cloud-free weather conditions. Currently, 10 international laser ranging stations under the International Laser Ranging Service (ILRS) provide limited experimental support to NavIC. See Figure 7.10 for locations. To date, the ILRS has undertaken two laser ranging campaigns to develop the capability to routinely track NavIC satellites in future [42, 43].

7.2.4 Geodesy and Time Systems

7.2.4.1 Geodesy

NavIC provides position coordinates in the U.S. WGS 84 Coordinate System as does the U.S. GPS. Details on WGS-84 can be found in Section 3.5

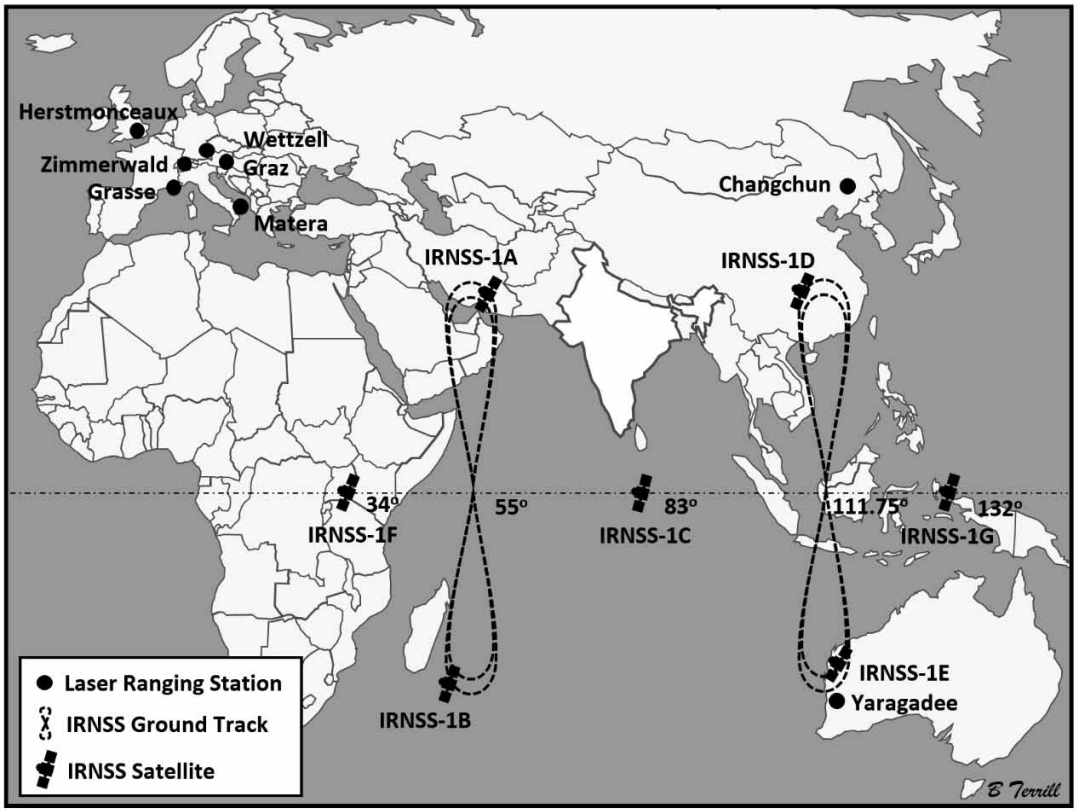


Figure 7.10 International laser ranging support for NavIC. (Courtesy of Brian Terrill.)

7.2.4.2 Time Systems

NavIC uses a time reference called IRNSS Network Time (IRNWT), which is a continuous time scale, meaning that it does not account for leap seconds, and is traceable to Indian National Time, designated UTC National Physics Laboratory of India (NPLI) or UTC (NPLI). IRNWT serves two primary functions:

- Navigation timekeeping: IRNWT supports the navigation mission by providing the time reference for orbit determination and time synchronization (OD&TS) of the NavIC constellation.
- Metrological timekeeping: IRNWT is steered towards International Atomic Time (TAI) and provides offsets between the IRNSS time and UTC (NPLI) and offsets used by other GNSS time scales such as GPS and GLONASS.

IRNWT will be maintained by two timing facilities and two parallel time scales; one is primary and the second is working in hot redundancy.

INC-1 at Byalalu maintains the primary time scale with an ensemble average of 3 (or 4) cesium clocks and 2 active hydrogen masers (AHMs) [46]. The main time scale is defined with a start time of 00:00 UTC (or UT) on Sunday, August 22, 1999 (midnight between August 21 and 22). At the start epoch, IRNSS system time was ahead of UTC by 13 leap seconds (i.e., IRNSS time, August 22, 1999, 00:00:00 corresponds to UTC time August 21, 1999, 23:59:47).

INC-2 (under development) in Lucknow will maintain the backup time scale. IRNWT will also be maintained by 3 (or 4) Cs AFSs and 2 active AHMs. The definition of epoch for IRNWT-II will be similar to the Galileo start of epoch and GPS rollover. IRNWT start epoch 00:00:00 (WN= 0, TOW= 0) shall be 1999-08-21 23:59:47 UTC. The time format is similar to GPS (i.e., week number and time of week modulo 604800). At the start epoch, TAI shall be ahead of IRNWT-II by 32 leap seconds [47].

7.2.4.3 Indian National Time

NPL maintains UTC (NPLI). The IRNSS time scale is traceable to UTC (NPLI). Currently, time is transferred between NPL and INC-1 and INC-2 using an all-in-view GPS P3 receiver. The Indians are planning to supplement GPS time transfer (possibly from the USNO) with a dedicated TWSTFT link [28, 33].

7.2.4.4 Future Upgrade Plans

Upgrades to IRNWT are planned to provide additional robustness to the IRNSS system time scale. The concept is to exploit both ground and space clocks. An algorithm will allow for the addition or deletion of clocks at any time without significantly affecting system performance and while providing automatic error detection and correction [33].

7.2.5 Navigation Services

NavIC will provide two levels of service, a public Standard Positioning Service (SPS) and an encrypted Restricted Service (RS); both will be available on both L5-band (1,176.45 MHz) and S-band (2,492.028 MHz) [28, 41]. 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 [48].

- The expected position accuracy for single-frequency navigation accuracy was unspecified at the time of this writing. The broadcast ionospheric correction model is based on a grid of 80 points and should support accurate positioning under nominal ionospheric condition.
- The expected position accuracy for dual frequency receivers is projected to be 20m (2σ) horizontal for users within the Indian Ocean Region (approximately 1,500-km beyond India's borders) and less than 10m (2σ) horizontal accuracy over India. The NavIC navigation signals are transmitted on both L5-band and S-band using on a common oscillator, thus allowing the receiver to measure the ionospheric delay in real time and allowing the user equipment to apply corrections.
- The expected time accuracy is projected to be 20 ns (2σ) of UTC (NPLI) when using the broadcast corrections in navigation message.
- The operators of NavIC have not specified the expected the RS navigation accuracy.

At the time of this writing, the seven-satellite NavIC constellation was just completed and performance had not yet been measured by ISRO. However, the ISRO measured NavIC position accuracy (longitude, latitude, and altitude) using the four-satellite-constellation on April 30, 2015. Based on the measurements, NavIC provided better than 15m (2σ) horizontal position accuracy during periods of satellites visibility (18 hours per day) using a dual-frequency receiver.

7.2.6 Signals

The external and internal characteristics of the civil modulation of the navigation signals were detailed in the IRNSS Interface Control Document (ICD) version 1.0 in 2016. The ICD can be found at the following link: http://irnss.isro.gov.in. ISRO has not disclosed information on the military RS modulation [48].

7.2.6.1 NavIC Navigation Signal Frequencies

NavIC broadcasts the SPS and the encrypted RS on both the L5-band and the S-band. The L5 radio frequency is centered at 1,176.45 MHz with a bandwidth of 24 MHz (1,164.45 to 1,188.45 MHz). The S-band signal is centered at 2,492.028 MHz with a bandwidth of 16.5 MHz (2,483.50 to 2,500.00 MHz). All the NavIC navigation signals are right-hand circularly polarized. The NavIC will provide a minimum received power of –159 dBW for the L5 SPS navigation signal and –162.3 dBW for the S-band navigation SPS signal. The maximum received power is –154 dBW for the L5 SPS navigation signal and –157.3 dBW for the S-band navigation SPS signal. The RS service power levels are unspecified [28].

7.2.6.2 NavIC Navigation Signal Modulations

The SPS signal uses BPSK-R(1) modulation. The RS uses BOC(5,2) modulation for each of two components, a data channel and pilot [48]. Interplexing is used to provide a constant power envelope (see Section 2.4.3).

7.2.6.3 NavIC Code Properties

The NavIC SPS uses Gold codes similar to the GPS SPS. NavIC shares the same code length of 1,023 chips and code chipping rate of 1.023 Mcps as GPS. The code is generated with the G1 and G2 polynomials as defined as:

$$\text{G1: } X10 + X3 + 1 \text{ and G2: } X10 + X9 + X8 + X6 + X3 + X2 + 1$$

The G1 and G2 polynomials are similar to those defined for the GPS C/A signal. Details on the GPS Gold codes are found in Section 3.7. The polynomials are generated are by a 10-bit maximum length shift registers XOR'ed to create a 1,023-chip long PRN sequence. Figure 7.11 shows the shift registry layout used to generate the NavIC SPS spreading codes [28].

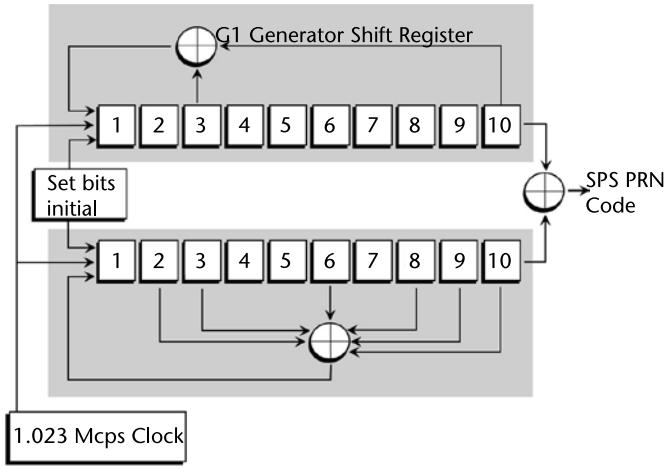


Figure 7.11 NavIC shift register design for SPS spreading code [28]. (Courtesy of Brian Terrill.)

7.2.6.4 NavIC Navigation Message

The NavIC navigation message is defined as a 2,400-bit-symbol master frame broken down into four 600-bit subframes transmitted at 50 bps. Each subframe consists of a 16-bit sync word followed by 584 bits of interleaved data. The 584 bits contain the navigation data interleaved with FEC, resulting in 292 data bits for the navigation message. Subframes 1 and 2 contain the primary navigation message needed for computing a navigation fix and subframes 3 and 4 contain the secondary navigation information like ionospheric grid correction parameters, text messages, and differential corrections [49]. The primary navigation message is composed of a telemetry word (TLM), time of week count (TOWC), alert, autonav, subframe identification (ID), spare bit, navigation data, CRC, and tail bits. The secondary navigation message is composed of TLM, TOWC, alert, autonav, subframe ID, spare bit, navigation data, CRC, tail bits, additional message ID, and additional PRN ID. Many of the navigation data elements, including the ephemeris and clock correction parameters, are similar to those used for GPS (see Section 3.7.4). Figure 7.12 shows the layout of the NavIC navigation message [28].

Additional details on the data structure, message types, message functions, and data algorithms are provided in the IRNSS SPS ICD and corresponding future updates [28].

7.2.7 Applications and NavIC User Equipment

As stated above, ISRO has designed NavIC to support both civil and military applications. The proposed civil applications include terrestrial, aerial, and marine navigation, disaster management, vehicle tracking and fleet management, integration with mobile phones, precise timing, collection of mapping and geodetic data, and visual and voice navigation for drivers [50]. At the time of this writing, no specific information was available on planned military applications.

Also, at the time of this writing, there was little information available on NavIC user equipment, likely because the constellation was just completed. Under available ISRO plans, ISRO is sponsoring development of NavIC and NavIC-GPS

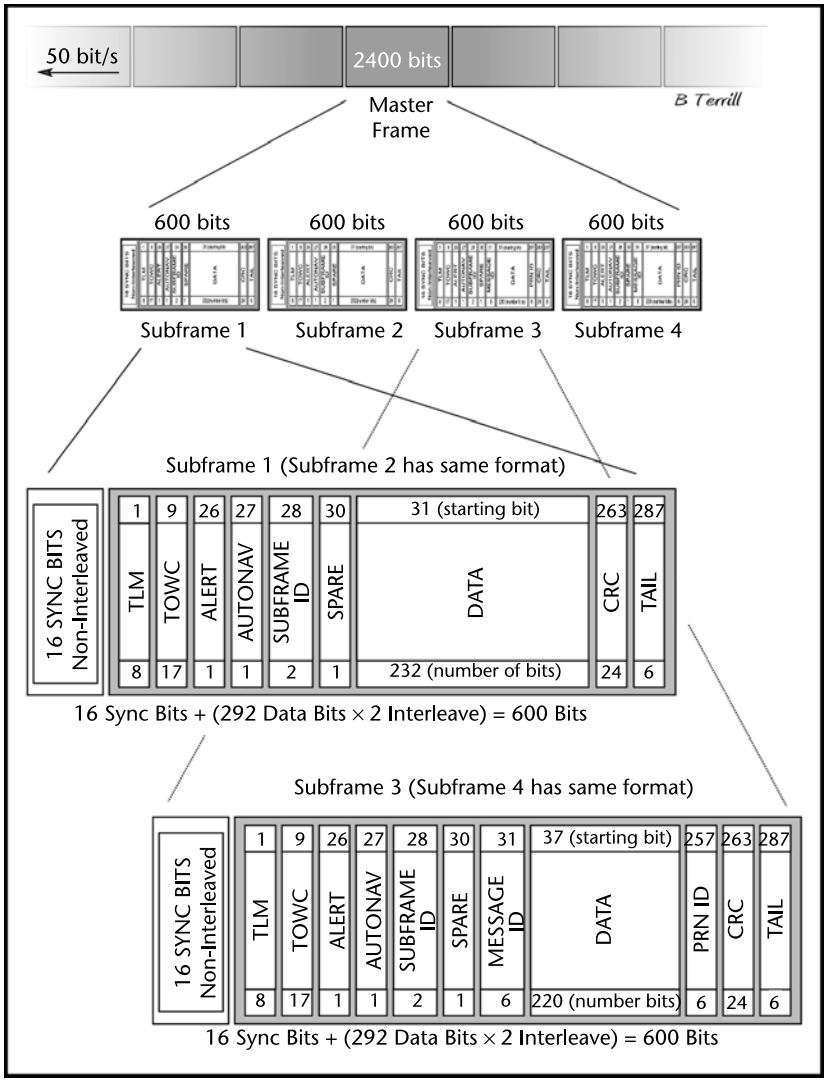


Figure 7.12 NavIC navigation message structure [28]. (Courtesy of Brian Terrill.)

receivers and NavIC-related chip sets for embedded applications. Under current plans, ISRO will sponsor development of single frequency designs independently for both the L5-band and S-band for both the civil BPSK and restricted (military) BOC navigation services. In addition, ISRO plans to develop dual-frequency (S-/L5-bands) receivers using either the civil or military navigation services [21, 51].

Internationally, only a few companies had announced plans to develop NavIC capable receivers, antennas or chipsets at the time of this writing. These included JAVAD GNSS, IFEN GmbH, and Hemisphere GPS LLC [52–54]. This is expected to dramatically change once NavIC is fully deployed and continuously available. Single-band L5 NavIC-GPS receiver designs should be easy to modify because only minor software upgrades are required to use the NavIC L5 navigation signal along with GPS L5.

References

[1] Matsumoto, A., "Quasi-Zenith Satellite System," *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, 2016.

[2] Moriyama, H., "Status Update on the Quasi-Zenith Satellite System," *ICG-10*, Boulder, CO, 2015. www.unoosa.org/pdf/icg/2015/icg10/06.pdf. Accessed January 1, 2016.

[3] JAXA, "Overview of the First Quasi-Zenith Satellite 'MICHIBIKI'," http://global.jaxa.jp/countdown/f18/overview/michibiki_e.html. Accessed January 1, 2016.

[4] IS-QZSS, Version 1.6, p. 8. qz-vision.jaxa.jp. Accessed January 1, 2016.

[5] Murai, Y., "Project Overview of the Quasi-Zenith Satellite System," *2015 PNT Advisory Board*, www.gps.gov/governance/advisory/meetings/2015-10/murai.pdf. Accessed January 1, 2016.

[6] "Overview of the Quasi-Zenith Satellite System (QZSS)," Office of Space Policy, Cabinet Office, Government of Japan. http://qzss.go.jp/en/overview/services/sv01_what.html. Accessed January 1, 2016.

[7] "Mitsubishi Electric Completes Expansion of Satellite Production Facility." http://www.mitsubishielectric.com/news/2013/0322-a.html. Accessed January 1, 2016.

[8] "Mitsubishi Electric Completes Expansion of Satellite Production Facility," http://www.businesswire.com/news/home/20130321005486/en/Mitsubishi-Electric-Completes-Expansion-Satellite-Production-Facility. Accessed January 1, 2016.

[9] "DS 2000 A Proven Commercial Platform," March 1, 2012. worldspaceriskforum.com/2012/wp-content/uploads/.../35TORU1.pdf, Accessed January 1, 2016.

[10] "Satellite Platform DS2000 - Mitsubishi Electric," www.mitsubishielectric.com/bu/space/satellite_platform/. Accessed January 1, 2016.

[11] QZSS (Quasi Zenith Satellite System), "eoPortal Directory," https://directory.eoportal.org/web/eoportal/satellite-missions/q/qzss. Accessed January 1, 2016.

[12] International GNSS Service, "QZSS Constellation Status Information," https://igscb.jpl.nasa.gov/projects/mgex/Status_QZSS.htm. Accessed January 1, 2016.

[13] Sawamura, T. T., et al., "Performance of QZSS (Quasi-Zenith Satellite System) & L-Band Navigation Payload," *Proceedings of the 2012 International Technical Meeting of The Institute of Navigation*, January 30–February 1, 2012, pp. 1228–1254.

[14] "New Ground Antennas and Satellite Control System for QZSS," http://www.mitsubishielectric.com/bu/space/ground/control/qzss/index.html. Accessed January 1, 2016.

[15] Hama, S., et al., "Development of the Time Management System," *Proceedings of the 22nd International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2009)*, Savannah, GA, September 2009, pp. 3338–3343.

[16] "R&D for Satellite Navigation," NICT Presentation, October 23, 2009. http://www2.nict.go.jp/aeri/sts/2009TrainingProgram/Reserch%20Activities%20of%20NICT/Training_Hama_2009Oct.pdf.

[17] Kubo, N., "Japanese GPS Argumentation System—QZSS," *2010 Stanford PNT Symposium*, November 9, 2010, http://scpnt.stanford.edu/pnt/PNT10/presentation_slides/8-PNT_Symposium_Kubo.pdf. Accessed January 1, 2016.

[18] Sakai, T., "The L1-SAIF Signal: How Was It Designed to Be Used?" *ION GNSS 2012*, Nashville, TN, September 17–21, 2012, www.enri.go.jp/~sakai/pub/gnss2012_saif.ppt. Accessed January 1, 2016.

[19] Murai, Y., "2014 Project Overview Quasi-Zenith Satellite System," *2014 QZS System Services (QSS)*, February 17, 2014, www.unoosa.org/pdf/pres/stsc2014/2014gnss-05E.pdf. Accessed January 1, 2016.

[20] Kogure, S., and M. Kishimoto, "Evaluation of QZS-1 LEX Signal," *International Committee on GNSS (ICG) Working Group B*, Beijing, China, November 4–9, 2012, http://www.unoosa.org/pdf/icg/2012/icg-7/wg/wgb3-2.pdf.

[21] "Current status of Quasi-Zenith Satellite System," International Committee on GNSS, Pasadena, CA, December 8, 2008, www.unoosa.org/pdf/icg/2008/icg3/08-0.pdf

[22] Asari, K., "Application Demonstrations," *Proceedings of the 25th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2012)*, Nashville, TN, September 2012, pp. 3278–3294.

[23] Terada, K., "Current Status of Quasi-Zenith Satellite System (QZSS)," *Munich Navigation Congress*, March 2011.

[24] Vithiyapathy, P., "India's Strategic Guardian of the Sky," Chennai Centre for China Studies, Occasional Paper 001-2015, August 25, 2015, p. 4. http://www.c3sindia.org/strategicissues/5201.

[25] Chander, S. A., "IRNSS Is Important for 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".

[26] "A Gift to People from Scientists: India's GPS Named 'NAVIC'," *Hindustan Times*, New Delhi, April 29, 2016.

[27] "IRNSS Launch: PM Modi Names New Navigation System as 'NAVIC'," *NDTV Press Trust of India*, April 29, 2016.

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

[29] Mateu, I., et al., "A Search for Spectrum: GNSS Signals in S-Band Part 1," *Inside GNSS Magazine*, Vol. 6, No. 5, 2010, pp. 67–69.

[30] Suresh, V. K., "Indian SATNAV Programme - Challenges and Opportunities," *First ICG Meeting, UN OOSA*, Vienna, November 1–2, 2006, www.unoosa.org/pdf/sap/2006/icg/09-1.pdf. Accessed January 1, 2016.

[31] "Indian Space Programme - Major Events During 2006," December 27, 2006, http://www.isro.gov.in/update/27-dec-2006/indian-space-programme-major-events-during-2006. Accessed January 1, 2016.

[32] ISRO, "Satellites for Navigation," http://www.isro.gov.in/applications/satellites-navigation. Accessed January 1, 2016.

[33] Kumari, A., A. Kartik, and S. C. Rathnakara, "IRNSS Composite Clock: IRNSS New Time Scale," *ISAC/ISRO, India, Precise Time and Time Interval Meeting*, Monterey, CA, January 25–28, 2016, https://www.ion.org/ptti/abstracts.cfm?paperID. Accessed January 1, 2016.

[34] ISRO, "Indian Regional Navigation Satellite System," http://irnss.isro.gov.in/. Accessed January 1, 2016.

[35] Ganeshan, A. S., et al., "First Position Fix with IRNSS, Successful Proof-of-Concept Demonstration," *InsideGNSS Magazine*, July/August 2015, http://www.insidegnss.com/node/4545.

[36] Private conversation with Kaushikkumar Suryakant Parikh, *10th International Committee on GNSS (ICG-10) after presentation Status Update on the Indian Regional Navigation Satellite Sytem (IRNSS) and the GPS-Aided GEO-Augmented Navigation System (GAGAN)*, Kaushikkumar Suryakant Parikh, Indian Space Research Organization, India, ICG-10, Boulder, CO, November 1–6, 2015.

[37] "IRNSS Signal Plan," *Navipedia*, http://www.navipedia.net/index.php/IRNSS_Signal_Plan.

[38] Jain, P. K., "Indian Satellite Navigation Programme: An Update," *ISRO HQ, India, 45th Session of S&T Subcommittee of UN-COPUOS*, Vienna, February 11–22, 2008, www.unoosa.org/pdf/icg/providersforum/02/pres04.pdf.

[39] Indian Regional Navigational Satellite System (IRNSS), "eoPortal," January 1, 2016, https://directory.eoportal.org/web/eoportal/satellite-missions/i/irnss.

[40] Indian Regional Navigation Satellite System (IRNSS), "ISRO Satellite Centre, Bengaluru," http://www.isac.gov.in/navigation/irnss.jsp. Accessed January 1, 2016.

[41] Seelin, S. A., and K. Kumar, "IRNSS," Chapter 8.6 in *From Fishing Hamlets to Red Planet, India's Space Journey*, Rao, P., et al., (eds.), New York: HarperCollins, 2015, p. 606.

[42] "ISRO PSLV-C22/IRNSS-1A Mission Brochure," http://www.isro.gov.in/pslvc22-brochure, September 2010, page 1. Accessed January 1, 2016.

[43] Government of India, "Department of Space Annual Report 2014-2015," January 2015, http://www.isro.gov.in/sites/default/files/article-files/right-to-information/AR2014-15.pdf. Accessed January 1, 2016.

[44] Indian Regional Navigation Satellite System (IRNSS), "Salient Features of IRNSS 1E," http://www.isac.gov.in/navigation/html/irnss-1e.jsp. Accessed January 29, 2015.

[45] Parikh, K. S., "Update on Indian Regional Navigation Satellite System (IRNSS) and GPS Aided Geo Augmented Navigation (GAGAN)," *CG-10*, Boulder, CO, November 1–6, 2015, www.unoosa.org/pdf/icg/2015/icg10/05.pdf.

[46] Neelakantan, N., "Overview of the Timing System Planned for IRNSS," November 2010, www.unoosa.org/pdf/icg/2010/ICG5/timing-session/06.pdf.

[47] LEAPSECS, "Welcome to a New Practical Time Scale," October 2014, https://pairlist6.pair.net/pipermail/leapsecs/2014-October/005238.html, refers to Sat Oct 18 11:28:47 EDT 2014, According to http://www.isro.org/Tender/Istrac/ISTRAC-05-2013-14.pdf, section 3.1.14, but the original Indian Tender is unavailable. Accessed January 1, 2016.

[50] ISRO, "Indian Regional Navigation Satellite System," http://www.isro.gov.in/irnss-programme. Accessed January 1, 2016.

[51] "Request for Proposal for the Development of IRNSS SPS-GPS User Receivers," March 2014, Space Applications Center Indian Space Research Organization, SAC/SNTD/SNAA/IRNSS/RFP1/BulkRx/12/2013_Ver1, https://eprocure.isro.gov.in/tnduploads/sac/pressnotices/PRSN1.pdf.

[52] "We Track IRNSS – JAVAD GNSS," October 1, 2014, https://www.javad.com/jgnss/javad/news/pr20141001.html.

[53] "GPS World Receiver Survey," *GPS World Magazine*, July 2015, gpsworld.com/resources/gps-world-receiver-survey/.

[54] "RF (Including GNSS) Signal," U.S. Patent US8897407, www.google.ch/patents/US8897407.

CHAPTER 8

GNSS Receivers

Phillp W. Ward

8.1 Overview

Numerous GNSS receiver designs have evolved since the inception of GPS, which was the first satellite navigation system that used direct sequence spread spectrum (DSSS) technology. This evolution continues as navigation satellite constellations and electronic technology advance in response to worldwide position, velocity and timing (PVT) services market demands. Receivers designed for different markets often have significantly different form factors and features as a result of performance trade-offs. Some major performance limitations include signal blockage, including signal attenuation due to physical objects, heavy foliage or dense particulate-bearing smoke, and noise interference, including natural interference (scintillation), intentional interference (jamming), adjacent-band interference and multipath (see Chapter 9). Some receiver designs are more robust to these limitations than others. Receiver design trade-offs differ depending on the intended GNSS applications as well as the constellation and technology eras, again resulting in different form factors and features. However, there are many basic design principles that are applicable to all GNSS receivers. This chapter describes these basic design principles beginning with the functional block diagram of Figure 8.1 that depicts a generic, multifrequency GNSS receiver architecture.

All GNSS receivers that operate continuously in real time with live satellite signals require the functions shown in Figure 8.1, namely, one or more antenna elements and associated antenna electronics, one or more front ends, multiple digital receiver channels (including a search engine), a receiver control and processing function, a navigation control and processing function, and other essential functions such as the reference oscillator, frequency synthesizer, power supply with DC power regulators, appropriate user and/or external interfaces, and possibly an alternate receiver control interface for a GNSS receiver designed to be integrated with and controlled by another navigation system. There is usually an internal rechargeable battery that provides standby power to keep the reference oscillator and the timekeeping function alive when the receiver is operating in standby mode (and sometimes even when the receiver is actually powered off). The prime power

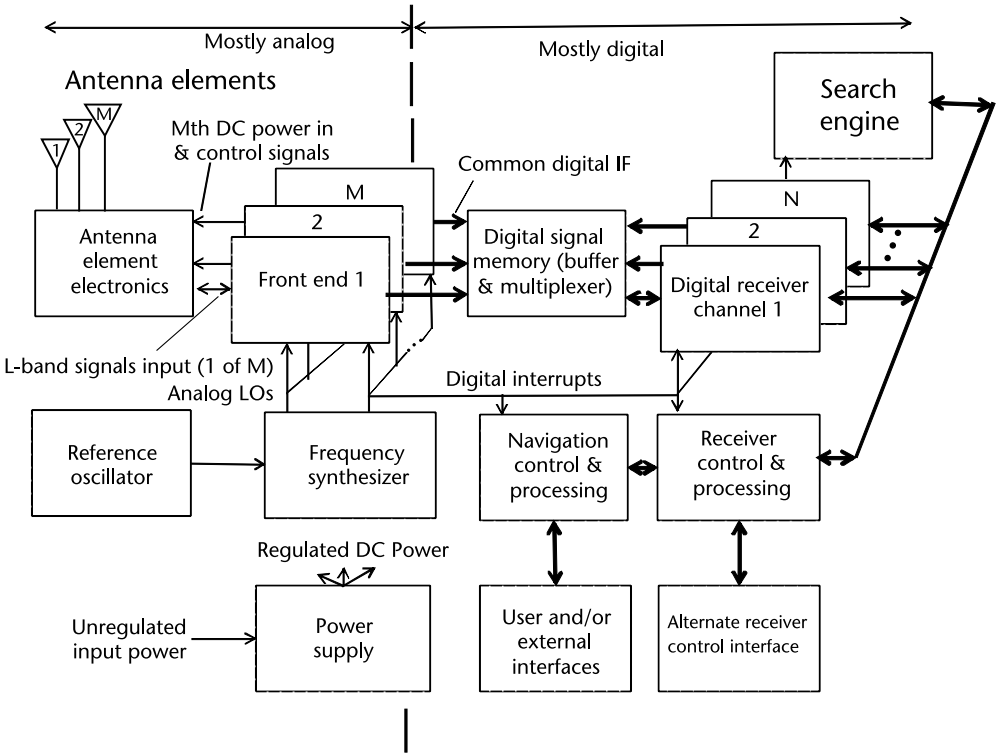


Figure 8.1 Generic GNSS receiver block diagram.

may be an internal replaceable or rechargeable battery or provided by external (phantom) AC or DC power.

Not shown in Figure 8.1, but included in many GNSS receiver designs are augmentation features such as an inertial measurement unit (IMU) that provides velocity aiding. Such augmentation, often called the flywheel effect in the receiver velocity measurement dimension, can significantly improve the dynamic stress performance of the receiver as well as operate-thru navigation performance when the receiver loses track, plus it provides independent velocity aiding that speeds up the reacquisition process. Another optional but beneficial augmentation is an atomic clock, such as a chip-scale atomic clock (CSAC) phase-locked with a crystal oscillator. When a CSAC replaces the reference oscillator, this provides the flywheel effect in the receiver time dimension. During receiver outages, a CSAC maintains precise time, thereby reducing reacquisition time (and subsequent acquisition time if kept operational when the receiver is turned off). Externally provided correction signals, such as those derived by an independent, worldwide, ground-based monitoring system that uplinks the corrections to geostationary (GEO) SVs that retransmit them, are received and processed along with the GNSS signals. These correction signals (see Chapter 12) improve the real time PVT repeatability and accuracy to centimeter-level or higher precision. Ranging from signals of opportunity such as cell phone towers can also add robustness to the GNSS receiver even for indoor operation (see Chapter 13).

The mostly analog portion of the receiver includes the antenna, front end, reference oscillator, frequency synthesizer, and power supply, although there may be some digital technology in any of these primarily analog designs. The remaining functions are mostly digital in modernized receivers. Although there is a growing trend toward all-microprocessor implementations, the higher speed functions that operate at the analog-to-digital converter (ADC) sample rate are often implemented as application-specific integrated circuits (ASICs) for high-volume production or with field programmable gate arrays (FPGAs) for low-volume production or experimental receivers, while the remaining slower rate functions are implemented in software. Oftentimes the software programs, once debugged, are implemented in firmware (i.e., the programs are stored in nonvolatile memory), with data stored in both nonvolatile memory (if constant) and volatile memory (if variable). These programs run in microprocessors and/or more specialized digital signal processors. Each processor is typically asynchronous in the sense that it contains its own ultrahigh-speed (analog) oscillator that determines its clock speed. Digital communications are typically conducted via digital memory using either direct (parallel bit) access or high-speed serial data transfer techniques. The processes are synchronized by prioritized interrupts, the hardware ones provided by the frequency synthesizer and the slower software ones provided by real-time operating system(s) in the processor(s) based on the hardware interrupts. To the extent that part of the receiver design can be implemented using digital signal processing (DSP) techniques running in one or more microprocessors, then that portion could be called software-defined, but it is impossible to synthesize a complete GNSS receiver as a software-defined receiver (SDR), although this misnomer is often used in the literature. However, the term SDR does readily apply to receivers that do not operate in real time, but are either used to verify or enhance a real-time receiver design in progress (or in planning) or to perform postmission processing using prerecorded front-end data. It is possible to synthesize a configurable GNSS receiver (CGR), wherein the operational architecture of the receiver is defined by the reconfigurable modular functions shown in Figure 8.1. In the CGR context, all software-defined functions would run under a real-time operating system that supports multiple hardware (and associated software) configurations and interfaces. In practice, most modernized GNSS receivers have some ability to be software-defined but are usually limited in configurability.

8.1.1 Antenna Elements and Electronics

Referring to Figure 8.1, each antenna element converts the received GNSS signals (plus noise) from electromagnetic fields in space into electrical voltages. Unlike the era when only the GPS L1 C/A code was available for conventional civil applications, a modernized civil receiver will require a multiple-frequency antenna design involving a single wideband or multiple narrowband antenna elements. Figure 8.1 depicts M single-frequency antenna elements, one element per supported L-band frequency. There could also be one wideband antenna element with passive signal splitter circuitry provided by the antenna element electronics that matches the load impedance and provides multiple supported L-band frequencies. If the antenna is remotely located, the antenna element electronics may also provide a low noise amplifier (LNA) to compensate for losses of the connecting coaxial transmission cable.

In this case the signal splitting shown in Figure 8.1 would be performed in the receiver. A GNSS antenna with an LNA is referred to as an active antenna, whereas an antenna without an LNA is referred to as a passive antenna. The antenna element electronics may also play a role in ensuring that each single frequency antenna element is receptive to right-hand circularly polarized (RHCP) electromagnetic waves.

8.1.2 Front End

The front end provides signal conditioning consisting of analog bandpass filtering, amplification, antialiasing filtering, and downconversion to a common intermediate frequency (IF) or to an in-phase (I) and quadrature (Q) baseband signal, followed by analog-to-digital conversion (ADC). Historic adjacent-band interference threats [1] should provide lessons learned that one front end per GNSS signal carrier frequency or closely spaced grouping of carrier frequencies is a prudent choice for providing adequate stopband protection. This front-end architecture is depicted in Figure 8.1. Regardless of the source L-band frequency, the received GNSS signals are normally all converted to a common IF. Therefore, as a minimum, there will be one front end for each GNSS signal center frequency or closely spaced grouping. As a maximum, there could be a need for unique front-end bandwidth optimization for each of the several GNSS signals on the same carrier frequency with significantly different bandwidths. In general, only one front end is provided for all signals on the same carrier frequency with a bandwidth defined by the requirements of the widest bandwidth signal. The digital receiver channel (described later) provides any additional bandwidth reduction required (usually by decimation).

8.1.3 Digital Memory (Buffer and Multiplexer) and Digital Receiver Channels

The digital memory (buffer and multiplexer) stores blocks of real-time data that are processed by the digital receiver channels. One block of digital data is typically being written by the front end while the previous block is being read by as many of the N receiver channels as are assigned to different SVs on that same carrier frequency. The real-time multiplexing scheme must be fast enough that all digital receiver channels have read and processed their block of the digital memory before that block is overwritten by the source front end. Note that each digital data stream contains the signals of all visible SVs transmitting on that one L-band carrier frequency. Also note that at this point all of the GNSS signals are buried in noise (i.e., the signal-to-noise ratio is highly negative in decibels, so each of these digital data streams would appear to be band-limited white noise if its analog counterpart was monitored with an oscilloscope).

Each digital receiver channel performs the faster carrier and code wipe-off processes. Each channel detects and tracks one SV signal under the receiver control and processing direction by selecting the appropriate front-end sampled data stream. The search engine may be the collective use of all available receiver channels to search in the time domain or a dedicated search engine operating in the frequency domain to acquire the first four SVs during a cold start. Cold start means that the receiver has little or no information about time or position and possibly inaccurate almanac information about the GNSS constellations, thereby making the two-dimensional search space (code range and carrier Doppler uncertainty) very large.

After the first four SVs have been acquired and the measurements incorporated, the search uncertainties become very small unless the almanac data are obsolete. When the total search uncertainty is small, then the remaining receiver channels can be allocated to do their own short search and signal acquisition process of additional SVs and at other L-bands. If a dedicated search engine performed the cold start search, then it is powered down after all sources of search uncertainties have become small. The navigation and control process is responsible for the determination of search uncertainty.

8.1.4 Receiver Control and Processing and Navigation Control and Processing

The receiver control and processing function performs the slower and smarter state control and baseband processing for all of the high-speed receiver channels. It directs navigation aiding (when available) to the digital receiver channels to speed up the acquisition/reacquisition process and provide robustness to the tracking process. It extracts measurements and navigation message data and passes these to the navigation control and processing function that incorporates these to produce PVT for the intended application. This navigation control and processing function provides the highest level of control and aiding to the receiver control and processing function.

8.1.5 Reference Oscillator and Frequency Synthesizer

The reference oscillator in combination with its frequency synthesizer supports analog signal downconversion, sample timing for the analog-to-digital conversion processes, and time synchronization of digital signal processing with respect to the real-time signals by means of discrete time interrupts to the digital processes.

8.1.6 User and/or External Interfaces

The user and/or external interfaces adapt the receiver to its operating environment. User interfaces include the control display unit (CDU) designed for human interaction. External interfaces include standard electronic interfaces to open systems or nonstandard interfaces for proprietary closed systems. These interfaces provide real-time PVT information to these systems and sometimes control and data information from them. For examples, augmentation systems, such as inertial navigation systems, eLoran, optical pattern recognition and Doppler radar, have been integrated into GNSS receivers via external interfaces in a loosely coupled (asynchronous) manner.

The most popular augmentation scheme integrates an IMU with the GNSS receiver (not shown in Figure 8.1) using internal interfaces. One of these internal interfaces is the reference oscillator, thereby enabling the IMU observables to be synchronous with the GNSS observables. The other interface is with the navigation and control processor that incorporates the observables of both the GNSS and IMU signals in a synergistic tightly coupled manner. This synergism combines the drift-free but RF interference-vulnerable GNSS receiver with the drift-prone but RF interference-immune IMU that significantly enhances the navigation performance of both systems.

Another optional function not shown in Figure 8.1 is built-in provision for data recording and storage, but this can also be an external interface. There are numerous GNSS applications that require built-in data storage for postmission processing purposes. If this requires the storage of raw observables, this becomes massive data storage. Solid-state digital data storage technology has made such immense progress in high density at low cost that it can now be used for this purpose.

8.1.7 Alternate Receiver Control Interface

If the receiver is the augmentation system for another navigation system, then the alternate receiver control interface provides access to the real-time observables and responds to smart external control. The most sophisticated version of this integration process would disable the navigation control and processing and the user and/or external interfaces functions. Since these are all real-time operations, there should also be an interface provision for synchronous operation between system clocks or the provision of an external interface to the GNSS receiver reference oscillator.

8.1.8 Power Supply

The role of the power supply is to provide regulated DC power to the extent required by all of the functions in the receiver. It is becoming more practical for there to be a common DC power for all circuits, but it is still preferable that the analog and digital power (and even the ground paths) be separate because of the inevitable cross-talk through the power supply lines, especially that caused by the digital circuits. Portable receiver applications require built-in battery operation but often have a provision for battery recharging while operating and some have a provision for hot-swapping the battery pack without disturbing the operation.

8.1.9 Summary

The functions in Figure 8.1 have all been described (along with a few related functions) to provide a comprehensive overview at a high level of a generic GNSS receiver and to lay 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, and then to 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.

8.2 Antennas

The GNSS receiver antenna receives the RHCP electromagnetic waves emanating from the GNSS SVs plus all unwanted electromagnetic emissions within the reception bandwidth, and converts them into electrical radio frequency (RF) voltages.

GNSS antennas consist of a radiating element (or elements) and associated microwave electronics. The microwave circuits are usually implemented physically below the radiating element as a multilayered printed circuit board containing a symmetric stripline transmission line layout between two ground planes. The ground plane suppresses the antenna gain pattern that would otherwise be in that direction, keeping the primary gain pattern hemispherical. The radiating element and microwave circuit are usually protected by a cover that is designed to be transparent to radio waves, which is referred to as a radome.

8.2.1 Desired Attributes

Ideal GNSS antenna attributes include low cost, a form factor acceptable for the intended platform, high phase center accuracy with small phase center variation, good gain characteristics (described below), and a perfectly matched output impedance [i.e., a 1:1 voltage standing wave ratio (VSWR)].

Gain is a measure of how receptive an antenna is to an electromagnetic wave as a function of the incident direction (azimuth and elevation angle) of the wave. Gain is measured with respect to an idealized, lossless antenna that receives electromagnetic waves equally well in all directions. This idealized reference is referred to as an isotropic antenna. Gain of an arbitrary antenna, as a function of azimuth and elevation angle, is formally defined to be the ratio of power out of the actual antenna to that out of the isotropic reference antenna in the presence of the same incident electromagnetic wave. Decibel units, referred to as dBi (decibels with respect to an isotropic antenna), are normally employed for antenna gains for a generic electromagnetic wave polarization. For gain towards a linearly or circularly polarized source, units of dBil and dBic are used, respectively. The physics of antennas is such that in order to increase the gain in some directions, it is necessary to reduce gain in others. The span of azimuth angles or elevation angles with appreciable gain (e.g., within 3 dB of the peak gain) is referred to as the antenna beamwidth. To increase peak gain, beamwidth must be reduced. Conversely, to increase beamwidth, peak gain is reduced.

GNSS antenna gain is ideally greater than unity with respect to an isotropic RHCP (0 dBic) antenna with constant RHCP gain over a hemispherical view angle from zenith down to a cutoff elevation angle (typically between 15° to 5° above the horizon) with sharp gain roll-off below cutoff, uniform (flat) gain over bandwidth of interest (resulting in constant group delay). An ideal GNSS antenna would also be minimally receptive to left-hand circular polarization (LHCP) below its cutoff elevation angle to reject specular ground reflected multipath. There are exceptions to these guidelines. For instance, a GNSS antenna in a mobile device would ideally have appreciable gain over a wider range of elevation angles since the device may be held in different orientations (e.g., with the screen vertical or horizontal).

Other typical GNSS antenna attributes include 50-ohm output impedance, lightning protection and, for remotely located antennas, a built-in low noise (1- to 3-dB noise figure) LNA with 15 to 50 dB of gain, but only for the desired L-band frequencies and with a very wide dynamic range.

Form factor is usually the most significant driver, second only to cost, for every GNSS antenna application, but even if both of these attributes are eliminated from the ideal attribute list, such an antenna design is never fully realizable. The

radiating element design determines the fundamental GNSS antenna attributes, namely its form factor and bandwidth. The antenna electronics, ground plane, and radome determine the final gain pattern, phase center, and output impedance attributes. The antenna phase center (i.e., the location being navigated by the GNSS receiver) is not a physical location. It is an electrical location that is not necessarily bounded by the physical body of the antenna. For precision GNSS applications, the precise location (accuracy) of the phase center with respect to a marked physical location is critical. Phase center variation as a function of elevation angle to any and all SVs must be extremely small and symmetrical (i.e., repeatable for all azimuth and elevation angles above the cutoff observation region). Achieving millimeter-level precision in this attribute results in the highest antenna cost and also imposes the most severe limitations on achieving a low-profile form factor. Fortunately, most GNSS applications do not require this level of precision.

8.2.2 Antenna Designs

Ultraprecise phase center accuracy and stability attributes are best achieved if the radiating element is a well-designed spiral (or variant of a spiral) antenna such as a conical spiral antenna mounted on a ground plane. This is because the RHCP characteristic is natural and the bandwidth is essentially frequency-independent for the spiral design, resulting in a highly uniform gain pattern that sustains its RHCP as well as its LHCP rejection at lower elevation angles [2].

The protective radome plays an important role in shedding signal contaminants such as water, snow, and ice buildup, but it also slightly alters the gain pattern, including phase center location and variation, so calibration should always include the radome and the radome material should possess uniform dielectric properties.

It is noteworthy that the antenna arrays on many GNSS satellites use long cylindrical (helical) spiral antenna elements that have directional beam properties, but conical spiral elements (that are much shorter) have been successfully demonstrated for this purpose. The lowest-profile version of this family is the flat spiral antenna, which sacrifices uniform gain versus elevation attributes of the conical spiral design, but largely retains the other desirable attributes. This or a variant of this design, sometimes with slight curvature, is a popular choice for precision GNSS applications requiring a low-profile antenna.

Another naturally RHCP antenna design is the quadrifilar helix design [3], that when fabricated on a ceramic dielectric can be both wideband and physically small. Numerous other low-profile GNSS antenna designs that are not natural RHCP, such as cross-dipole, patch, patch-on-dielectric (ceramic patch) antennas, sacrifice various performance features because of the limitations of their form factor and the manner in which they are converted into RHCP antennas. Another such antenna that is almost universally used in mobile phones is the Planar Inverted-F Antenna (PIFA), so called because the shape of its stacked patch elements on ceramic substrates look similar to the inverted letter "F" [4].

When the radiating element is naturally linearly polarized, it is converted into RHCP by a polarizing microwave (typically stripline) circuit that also performs impedance matching. For example, since patch antenna elements are naturally linearly polarized, they may be artificially converted to become circularly polarized by using quadrature excitation of its two linearly polarized ports. However, their

circular polarization always diminishes near the horizon, just where it is needed most for multipath rejection. The overall quality of the circular polarization depends largely on the feeding network and is often sensitive to the fabrication process. Using a ground plane causes the radiation pattern to be nearly hemispherical and also provides shielding at lower elevation angles.

Table 8.1 lists a representative mix of popular GNSS antenna element design types along with the typical applications that drive the selection of that antenna element type, including the most critical antenna parameter(s) demanded by the application, followed by the typical size, cost category, RHCP formation, and polarized bandwidth that results from this type of antenna element design. From this table, it should be apparent that antenna element design choice is one of the most challenging tasks in the design of GNSS receivers, especially for future small and low-cost applications that require receiving more than one GNSS frequency and sometimes frequency bands other than for GNSS signals. Table 8.2 is basically a continuation of Table 8.1 showing the most common performance specifications of GNSS antenna elements. These performance specifications typically include the antenna's axial ratio and VSWR, described next. Reference [2] provides a comprehensive treatment of a wide variety of currently available GNSS antennas.

8.2.3 Axial Ratio

The polarization characteristics of a GNSS antenna are important contributors to its overall performance. All of the navigation signals broadcast by GNSS satellites today are RHCP. Many interference sources are linearly polarized. The terms linear

Table 8.1 Typical Application Design-Driven GNSS Antenna Types with Key Features

Design Type	*Design Driving Applications*	*Critical Design Parameter(s)*	*Size (inches)*[1]	*Cost Category*	*RHCP Formation*	*Polarized Bandwidth*
Conical spiral	First-order positioning and timing	Highest precision	9.5 D × 5.5 H[2]	Highest	Natural	All GNSS bands
Low-profile spiral (or variant)	Highest precision land, marine, air, space	Highest precision low profile	7.5 D × 2.5 H	High	Natural	All GNSS bands
Cross dipole	High-precision land, marine, air, space	High precision, low profile	5.0 D × 2.5 H	Medium high	Artificial	All GNSS bands
Patch	Avionics, PVT	Best precision in lowest profile	2.7 D × 0.9 H	Medium low	Artificial	Multiple 20-MHz bands
Ceramic quadrifilar helix	Handheld GNSS	Best for man-portable	0.5 D × 0.9 H	Low	Natural	Wideband
PIFA[3]	Mobile phone	Lowest cost, smallest size and weight	1.5 L × 1.0 W × 0.33 H	Lowest	Artificial (typically poor)	GPS/GLONASS L1 C/A band + multiple mobile bands
Ceramic patch	Automotive	Lowest cost, smallest size	1.0 L × 1.0 W × 0.2 H	Lowest	Artificial (typically poor)	GPS/GLONASS L1 C/A band + radio bands

Note 1: D = diameter, H = height, L = length, W = width. Note 2: Dimensions are much larger if a multipath mitigation ground plane (such as a choke ring or resistive plane) plus an overall protective radome are required. Note 3: Planar Inverted-F Antenna (PIFA), so called because the shape of its stacked patch elements on ceramic substrates looks similar to the inverted letter "F" [4].

Table 8.2 Typical Performance Specifications for GNSS Antennas

Type	*Useful Beamwidth (Degrees)*	*Gain (dBic at Degrees)*	*Phase Center Variation (mm)*	*Axial Ratio (dB at Degrees)*	*VSWR[3] (dB at Center Frequency)*
Conical, spiral	160	>2 at 90, >−4 at 10	<2 accuracy[2], <1 stability[2]	<0.2 at 90, <0.5 at 10	<2:1
Low profile spiral	150	>5 at 90, >−3 at 15	<10 accuracy, <5 stability	<0.2 at 90, <2.0 at 10	<2:1
Cross-dipole	140	>3 at 90, >−2 at 20	Seldom specified	<1.0 at 90, <3.0 at 20	<2:1
Patch	160	>5 at 90, >−0.5 at 10[4]	Seldom specified	<1.0 at 90, <3.0 at 10[4]	<1.5:1
Ceramic quadri-filar helix	120	>3 at 90, >−6 at 30	Seldom specified	Seldom specified	<2.3:1
PIFA[1]	140	>−3 at 90, >−6 at 10	Seldom specified	Seldom specified	Seldom specified
Ceramic patch	140	>−3 at 90, >−6 at 10	Seldom specified	Seldom specified	Seldom specified

Note 1: Planar Inverted-F Antenna (PIFA) – so called because the shape of its stacked patch elements on ceramic substrates look similar to the inverted letter "F" [4]. Note 2: These accuracies do not include the inaccuracy caused by the near-field effect [5]. Note 3: The specification of VSWR only at the center frequency of the antenna is typical, but variations in VSWR are typically much larger for narrowband antenna elements than for wideband antenna elements. Note 4: For very low-profile patch antennas, gain and axial ratio performance at low elevation angles suffer. Representative specifications are –4 dBic at 10° for gain and < 15 dB axial ratio at 10°.

polarization and circular polarization refer to the characteristics of the electromagnetic waves. Far from the source, and when traveling through simple media such as a vacuum or air, the electric (E–) and magnetic (B–) field vectors that constitute a radio wave are always perpendicular to each other and also perpendicular to the direction of signal travel. In a linearly polarized wave, both the E-field and B-field oscillate in amplitude but always remain pointing in the same directions as the wave travels. Many terrestrial communication systems use one of two special types of linear polarization: vertical polarization, where the E-field is perpendicular to the surface of the Earth, or horizontal polarization, where the E-field is parallel to the surface of the Earth. In a circularly polarized wave, the E-field and B-field do not point in constant directions, but rather rotate 360° per wavelength as the wave travels. From the transmitter's perspective, the E– and B field vectors can rotate clockwise or counterclockwise. Clockwise rotation from the transmitter's perspective is by convention referred to as RHCP, and counterclockwise is referred to as LHCP.

For a linearly polarized wave incident from an arbitrary direction, an ideal RHCP antenna would provide constant gain as a function of the E-field orientation. For instance, if an antenna was perfectly RHCP towards the horizon, the gain of the antenna towards a horizontally or vertically polarized source at the horizon would be equal. Real antennas may exhibit a change in gain if the incident wave E-field orientation is varied over all possible directions (e.g., all directions perpendicular to the direction from the source to the antenna). The extent of this variation is referred to as the antenna *axial ratio*, *AR*, expressed in units of decibels. For a GNSS antenna, axial ratio can be used to determine the loss of RHCP antenna gain, L_a, where

$$L_a = 10\log_{10}\left(\frac{\left(1+AR_l\right)^2}{2\left(1+AR_l^2\right)}\right)\ (\text{dB}) \tag{8.1}$$

and $AR_l = 10^{AR/20}$ is the axial ratio in linear units. Thus, the loss of RHCP antenna gain due to an ideal axial ratio of 0 dB is 0 dB. Table 8.3 computes L_a for a range of AR, including the typically specified values that appear in Table 8.2, but those values are typically met only at the antenna zenith where there is usually the strongest signal, most antenna gain and least multipath. Note that an infinite axial ratio results in a bounded 3-dB loss of RHCP gain.

Figure 8.2 illustrates the axial ratio of a typical GPS L1 patch antenna gain pattern measured in an anechoic antenna chamber. This type of gain pattern results when the antenna under test at the receiving end is rotated 360° while being radiated at the L1 frequency by a calibrated linear antenna that is rotating in a right-hand circular manner at the transmitting end. This is a desirable test for two reasons. First, the user's antenna test equipment, including the calibrated reference linear antenna, is traceable by metrology to the BIPM (Bureau International des Poids et Measures), the international bureau of weights and measures, via a six-level "traceability pyramid" [6]. This pyramid begins with the *User's Test Equipment* at the base, followed by *General Purpose Calibration Laboratories, Working Metrology Laboratories, Reference Laboratories, National Metrology Institutions (NMIs),* and *BIPM* at the top of the pyramid. Second, the axial ratio of the antenna under test can be determined by inspection of the amplitude gain excursions (i.e., the axial ratio is defined as the peak-to-peak swing in decibels generated by the rotating linearly polarized antenna). The loss, L_a, can then be determined using (8.1). The resulting RHCP antenna gain, G_a, is then computed (typically in discrete elevation angle increments) simply as

$$G_a = G_p - L_a \quad (\text{dBic}) \tag{8.2}$$

where G_p is the peak gain envelope from the plot. Figure 8.3 illustrates a plot of the peak gain excursions (G_p) obtained from Figure 8.2. For example, referring to the notes in the top right corner of Figure 8.2, AR = 0.5 dB at 0° (zenith), resulting in $L_a = 0.004$ dB. Observe that $G_p = 5.3$ dBi at zenith, so $G_a = 5.3 - 0.004 \approx 5.3$ dBic at zenith.

At the elevation angles of 265° and 85° points in the Figure 8.2 plot, AR = 4.5 dB resulting in $L_a = 0.27$ dB. These angles correspond to the antenna elevation angles of 85° (below zenith) on both sides in Figure 8.3 plot (that, in turn, correspond to 5° user elevation angles above the horizon). Referring to Figure 8.3, $G_p = 0$ dBi at 85° on the left side, but $G_p = -1$ dBi at 85° on the right side, so the RHCP antenna gain, $G_a = -0.27$ dBic on the left side and $G_a = -1.27$ dBic on the right side. In other words, even though the axial ratio appears to be uniform for a

Table 8.3 Effect of Axial Ratio (*AR*) on RHCP Antenna Gain Loss (L_a)

AR (dB)	0	0.5	1	1.5	3	3.5	4	4.5	Infinite
LA (dB)	0.0	0.004	0.01	0.03	0.13	0.17	0.22	0.27	3.0

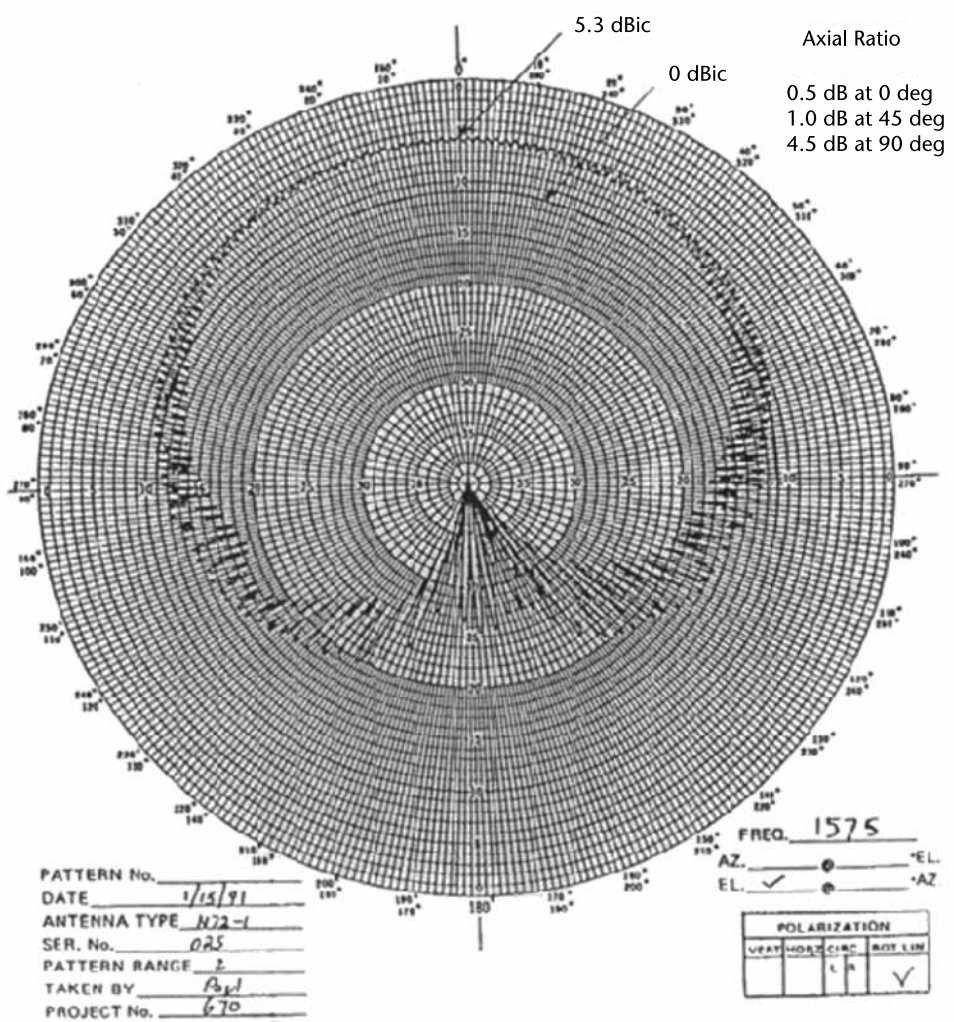


Figure 8.2 Patch antenna gain pattern including axial ratio.

given elevation angle, the actual RHCP antenna gain has at least 1-dB fluctuation in gain around the lower user elevation angle of 5°. Referring to Figure 8.2, note that this patch antenna is rapidly losing its RHCP characteristic right where it is needed most, namely at the lower user elevation angles as evidenced by the increasing axial ratio (the peak-to-peak excursions) as the elevation angle approaches the user horizon. This is characteristic of all antenna elements that do not have natural RHCP and is also characteristic of all antenna elements that have a low profile even if they do have natural RHCP.

Calibrated antenna gain patterns with axial ratios provided by antenna vendors are the exception rather than the rule because of the expense involved. Even high-end antennas specifications typically come with only a single representative axial ratio plot that cannot be accurately measured by the technique described above. Numerous such calibrated gain patterns are required in order to truly calibrate a GNSS antenna. This will require access to an anechoic antenna chamber with calibration traceable to BIPM along with antenna measurement expertise to

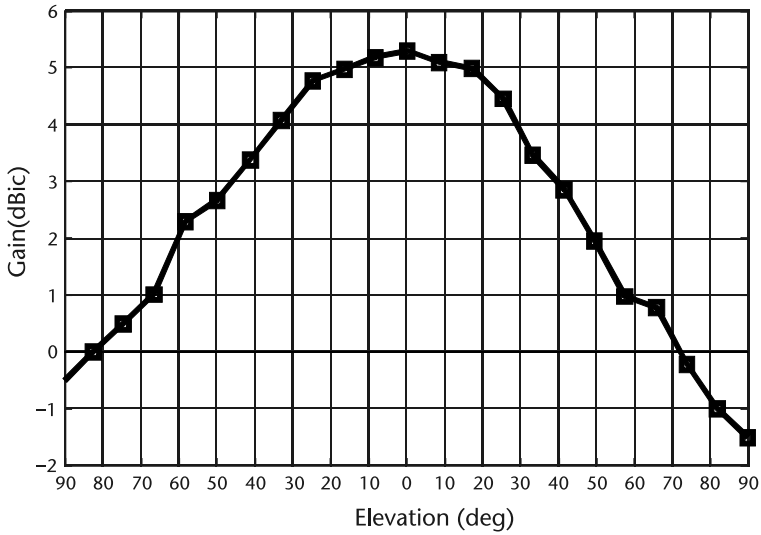


Figure 8.3 Patch antenna peak gain plot versus elevation angle (zenith = 0°).

conduct the antenna calibration measurements. An even more time-consuming and sophisticated antenna test setup in an anechoic chamber is required to determine the antenna phase center location and variations.

8.2.4 VSWR

VSWR is a measure of how well the antenna signal transmission line impedance is matched to its load, which is typically 50 ohms. The ideal VSWR would be expressed as 1:1 (i.e., no standing wave). The VSWR equation in terms of the reflection coefficient (a dimensionless ratio of complex voltage amplitudes or impedances) is

$$\begin{aligned} VSWR &= \frac{1+|\Gamma|}{1-|\Gamma|}:1 \\ \text{where:}\Gamma &= \frac{V_{reflected}}{V_{incident}} = \frac{Z_{load}-Z_0}{Z_{load}+Z_0}\ (\text{dimensionless}) \end{aligned} \tag{8.3}$$

and $V_{reflected}$ is the voltage amplitude reflected by the antenna load, $V_{incident}$ is the voltage amplitude of the signal from the antenna output, Z_0 is the complex output impedance of the antenna and Z_{load} is the complex input impedance of the antenna load. The inverse computation for measured or known VSWR is

$$|\Gamma| = \left|\frac{VSWR-1}{VSWR+1}\right|\ (\text{dimensionless}) \tag{8.4}$$

VSWR can also be determined by an antenna impedance matching performance factor called return loss (RL) in units of decibels as

$$VSWR = \frac{10^{\frac{RL}{10}} + 1}{10^{\frac{RL}{10}} - 1} : 1 \ \text{(dB)} \tag{8.5}$$

where

$$RL = -20\log_{10}|\Gamma| = 10\log_{10}\frac{P_{incident}}{P_{reflected}} \ \text{(dB)} \tag{8.6}$$

$P_{incident}$ is the incident power into the antenna load and $P_{reflected}$ is the reflected power from the load. RL is another measure of how well the antenna is matched to its load. The inverse equation where VSWR is measured or known is

$$RL = 20\log_{10}\left(\frac{VSWR + 1}{VSWR - 1}\right) \ \text{(dB)} \tag{8.7}$$

Inspection of the above equations shows that the best VSWR is achieved when Γ is low and RL is high. For a case example, assume VSWR = 1.01:1, then Γ = 0.00498 and RL = 46.1 dB. An often-used antenna bandwidth specification, referred to as the return loss bandwidth, is the span of frequencies for which RL is within 10 dB of its value at the antenna center frequency. To compare this metric to the more intuitive VSWR metric, assume the VSWR is an impressive 1.5:1 at the antenna center frequency. Using (8.7), the corresponding RL would be 14 dB. If RL is then deteriorated by 10 dB to 4 dB to define its bandwidth, then using (8.5) shows the VSWR has deteriorated to an unimpressive 4.42:1.

8.2.5 Antenna Noise

The value of T_{ant}, the antenna noise temperature measured in units of Kelvin (K) is required to analyze the signal to noise ratio of the received signal. Perhaps the most commonly misunderstood performance parameter in a GNSS antenna is the use of temperature units in degrees K to describe its noise contribution, especially since this essential receiver noise analysis parameter is never included in the antenna specification. This is because the antenna noise temperature is not its physical temperature, but is due to the temperature, $T(\theta, \phi)$, in the spherical direction that the antenna gain pattern, $G_{ant}(\theta, \phi)$, sees in its solid angle of gain coverage. This solid angle includes its side lobes and back lobes. The parameter ϕ is the azimuth and the parameter θ is the elevation in units of radians from the antenna boresight whose origin is at the nominal antenna phase center. The associated equation that defines this is [7]

$$T_{ant} = \frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi} G(\theta,\phi)\cdot T(\theta,\phi)\, d\theta\, d\phi \ \text{(K)} \tag{8.8}$$

The precise solution to this equation is seldom practical because of uncertainties in the direction of the antenna boresight during real time operation of the antenna (even if the antenna gain pattern and the surrounding temperature pattern are known accurately), but there are practical approximations. For example, if only a dark (cold) sky is covered by the entire gain pattern, T_{ant} could be as low as the 10K sky temperature at GNSS frequencies, but there are also hot spots included in the antenna gain pattern. These would include the Sun and stars, but the usual dominant factor is the (hot) Earth's surface at around 300K. So if there were half dark sky and half Earth coverage, then the antenna temperature would be approximately 150K. Since most of a typical GNSS antenna gain coverage is toward the sky during normal operation, then a reasonable assumption would be that the antenna noise temperature is about 100K, so this is the value assumed for the computational examples used in this chapter and in Chapter 9.

However, the designer should keep the possible variations in mind. Some higher GNSS T_{ant} field of view environments include: (1) inside buildings where the GNSS signals are received indirectly through windows or from a re-radiator; (2) a location surrounded by high mountains or looking into an urban canyon or under heavy foliage; (3) user equipment operated in a vehicle that blocks part of the antenna view or the antenna is tilted so that a substantial part of its field of view includes the Earth; (4) on a missile or artillery shell where the antenna field of view is skyward at launch but then tilts during the trajectory where the field of view includes the Earth; and (5) on a satellite in space where the antenna array is looking almost entirely at the hot Earth.

If the receiver is connected to a GNSS simulator, the antenna will be disconnected. In this case, the objective is to change the GNSS simulator signal (received signal power) to compensate for the difference in noise level. The equation for that change is [8]

$$\left(\Delta C_S\right)_{dB} = 10\log_{10}\left(\frac{T_{sim} + T_{receiver}}{T_{ant} + T_{receiver}}\right) \text{ (dB)} \tag{8.9}$$

where

T_{sim} = simulator noise temperature (assumed standard room temperature of 290K);

$T_{receiver} = 290\left(10^{\frac{(N_f)_{dB}}{10}} - 1\right)$ = receiver system noise temperature in K;

$(N_f)_{dB}$ = receiver noise figure at 290K;

T_{ant} = antenna noise temperature (K).

Note that this only compensates for the noise difference when connected to the GNSS simulator. The compensation for the combination of receiver antenna gain pattern and the (multiple space and time variable) received GNSS signal gain patterns is far more complex. Only very sophisticated GNSS simulators support such an elaborate gain compensation feature.

8.2.6 Passive Antenna

A passive antenna contains no powered components. A passive antenna is the most reliable antenna because active components have much higher failure rates than passive components (assuming uniform quality for both classes of components). A passive antenna can be used when there is minimal insertion loss (very short physical distance) between it and the first preamplifier, called the low noise amplifier (LNA), in the receiver front end. As will be observed later, a low noise figure LNA with sufficient gain keeps the overall receiver noise figure approximately equal to the noise figure (in decibels) of the LNA plus the minimum insertion loss (in decibels) prior to the LNA. There may be several passive narrowband antenna elements (one for each L-band frequency used by the receiver) or one passive wideband antenna element (that spans the entire range of L-band signals used by the receiver) or some combination of narrowband and wideband elements. In any case, the passive components must be combined in a manner that matches the composite antenna impedance to the coaxial cable impedance when it is terminated at the LNA input. For example, a 2-element antenna requires a passive 50-ohm diplexer to combine the separate RF signals into a 50-ohm coaxial cable that conducts the composite wideband RF signal to the 50-ohm input impedance of the LNA. There may also be passive L-band high-Q, low-insertion-loss, bandpass filters, such as cavity filters, to provide RF interference suppression in desired stopbands. The passive insertion losses prior to the first LNA, including coaxial cable and connector insertion losses, increase the receiver noise figure. This, in turn, reduces the carrier-to-noise power ratio in a 1-Hz bandwidth (C/N_0). This ratio in units of Hz is an excellent measure of GNSS signal quality since it is the same anywhere in the GNSS receiver except for implementation losses along the way. It is usually expressed in units of dB-Hz and defined by

$$\left(C/N_0\right)_{dB} = 10\log_{10}\left(C/N_0\right) \ \text{(dB-Hz)} \tag{8.10}$$

The piecewise equations for computing unjammed $(C/N_0)_{dB}$ are presented in (9.20) in Chapter 9.

8.2.7 Active Antenna

An active antenna means there are one or more LNAs (active components) inside the antenna housing that require external DC power. This power is provided by the receiver power supply via the center conductor of the coaxial cable. An active antenna is essential if the antenna is remotely located (such as a rooftop or mast mounted antenna). Every decibel of passive loss before the first LNA effectively reduces $(C/N_0)_{dB}$ by about the same amount, so the antenna LNA gain stage prior to cable passive loss keeps the overall receiver noise figure low. If a large dynamic range is not required by the GNSS receiver to achieve a high in-band RF interference (RFI) tolerance, then a single wideband antenna element that covers all RF signals of interest and a single wideband LNA can be used. Otherwise, a more elaborate active antenna scheme is required and is described in more detail in Section 8.3.

8.2.8 Smart Antenna

A smart antenna contains the entire GNSS receiver within the antenna enclosure thereby eliminating coaxial cables and connectors as well as other RF components. All of the output signals from a smart antenna are capable of driving long signal cable lines or are wireless. Modern semiconductor technology supports this very high level of integration in a very small enclosure that includes the antenna and radome. This integration not only significantly reduces insertion losses but also eliminates combining and later splitting the RF signals for each front end. The smart antenna may be battery-operated or require phantom power from the host platform. It may contain hot-swap, small, and massive memory storage units that provide many hours of data recording for postmission processing. The name "smart antenna" usually connotes that it is also a high-precision, self-contained GNSS receiver, including the reception and processing of an independent GNSS-correction signal, thereby achieving centimeter or decimeter level positioning accuracy in real time. The correction signal may be provided locally in the classic differential mode of operation or globally via geostationary (GEO) satellites (see Chapter 12). Since an assisted GPS/GNSS receiver [9] does not operate independently or continuously, it is not considered to be a smart antenna and is not described further herein.

8.2.9 Military Antennas

The military classifies GNSS antennas as either a fixed reception pattern antenna (FRPA) or controlled reception pattern antenna (CRPA). The FRPA is either a single wideband element covering all frequency bands of interest or it may contain a multiple number of narrowband antenna elements that, when combined, receive all military frequency bands of interest. In either case, the antenna provides a fixed gain pattern.

The CRPA is a phased array antenna (see [2, 3]) composed of multiple antenna elements per carrier frequency that is capable of varying its gain pattern using digital signal processing techniques. The most popular CRPA technique is an N-element CRPA that can steer deep gain nulls toward $N - 1$ jammers plus provide a small amount of gain toward multiple SVs in view that are not in line with the jammers. The typical military CRPA has 7 dual-frequency (L1 and L2) elements for large military platforms (such as aircraft). There are also smaller CRPAs with fewer elements used in smart weapons. Another CRPA technique uses a massive number of antenna elements with an equally massive digital signal processing capability to steer narrow beams of gain toward a selected number of SVs, none of which are in line with a jammer. Beam-steered CRPAs are the most effective, but least practical of the two designs for most military applications. Since baseband signal processing techniques cannot decrease the noise level increase caused by band limited white noise (BLWN) interference that is in-band to the GNSS receiver, the CRPA (or equivalent antenna selectivity technique) is the only remaining means of mitigating this threat beyond the FRPA-based receiver robustness to that same wideband interference. The CRPA robustness improvement is the sum of the null depth achieved (as a positive number of decibels) plus the antenna gain increase as compared to the FRPA counterpart for the same interference source and SV. This robustness improvement can be from 30 to 50 dB depending on the sophistication

of the CRPA technology. That sophistication is severely limited unless the GNSS receiver has the ability to provide direction cosines for the lines of sight to the SVs (or the equivalent) to the CRPA signal processing function. That ability requires IMU-aiding to the receiver for mobile platforms. The use of CRPAs and other techniques to mitigate interference and other forms of GNSS signal disruptions are discussed further in Chapter 9.

8.3 Front End

The fundamental goal of the front end is to amplify the L-band signals (plus noise) received by the antenna to an appropriate amplitude level while also downconverting them to a lower frequency so that they can be digitized for subsequent digital signal processing. Figure 8.4 illustrates a high dynamic range analog front-end design that accomplishes this fundamental goal and much more. There is usually one front end for each L-band carrier frequency being supported by the GNSS receiver.

Since there are multiple front ends in a multiband GNSS receiver, each front end is designed so that it is relatively simple to adapt the same basic design to each L-band center frequency that is incorporated into the receiver design. Each front-end has some unique components [e.g., bandpass filters and first local oscillator frequency (LO_1) tailored to the center frequency]. However, the design intent is to maximize commonality of parts to the extent possible (e.g., use a common IF and all associated parts thereafter). Other design goals are to achieve a low receiver noise figure and high dynamic range. Later it will be described how this front-end design, in combination with the receiver channels described in Section 8.4, can also accommodate the FDMA signals of the GLONASS constellation with substantial design commonality.

The front end is characterized by its gain plan, frequency plan, frequency downconversion scheme, and type of digital output signal. Referring to Figure 8.4, all amplifier gain and mixer stages should be much wider than the multiple bandpass filters so that the filters play the dominate role in establishing bandwidth, B_{fe}, passband flatness and group delay. These filters also determine the stopband

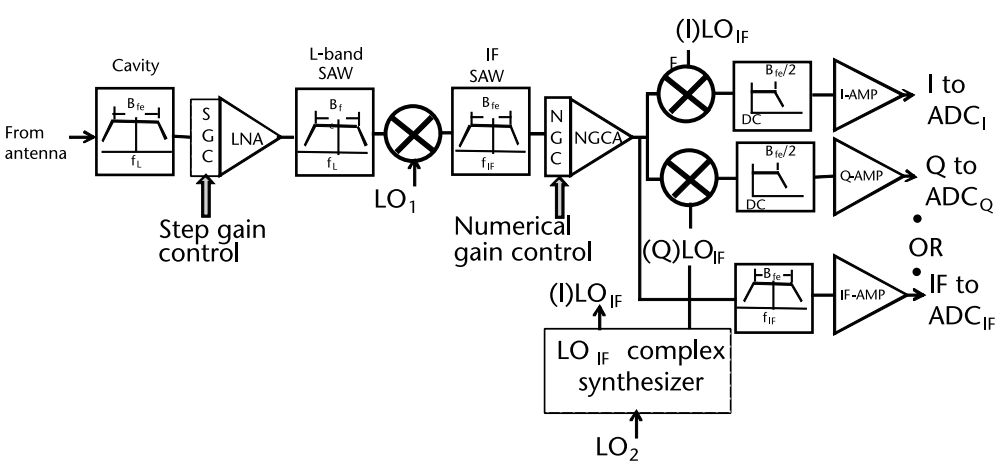


Figure 8.4 High dynamic range analog front-end block diagram.

rejection levels by their combined attenuations at frequencies above and below the passband. The front end is further characterized by its performance features, such as noise figure and dynamic range. These front-end characteristics are described beginning with the functional description.

8.3.1 Functional Description

Referring to Figure 8.4, there is a cavity filter (a high-Q, low insertion loss, passive L-band bandpass prefilter) shown as the first stage that protects the first active stage by minimizing out-of-band (particularly adjacent band) RF interference. There are usually nonlinear protective circuits (not shown) such as back-to-back PIN diodes to clamp any excessive (damaging) RF signal to ground. Because cavity filters tend to be physically large, this filter might be moved to the antenna assembly or replaced with a smaller prefilter or even eliminated if stopband rejection prior to the first active stage is an acceptable trade-off.

The first active stage is the LNA that provides gain to the antenna RF signals. The role that the LNA plays in setting the receiver noise figure is described later. If the antenna is remote, another LNA must be added there with this prefilter ahead of it for stopband protection. The gain of the local LNA must be adjusted accordingly, but the step gain control (SGC) dynamic range must be preserved. The LNA signals are bandpass filtered by the L-band surface acoustic wave (SAW) filter, then downconverted to an intermediate frequency (IF), f_{IF}, using the first local oscillator signal mixing frequency of $LO_1 = f_L - f_{IF}$, where f_L is the L-band frequency of interest. The frequency synthesizer (implemented in the frequency synthesizer function shown in Figure 8.1, but shown in more detail in a later figure) provides all of the required LOs and these are phase-locked to the reference oscillator. These frequencies are chosen based on the frequency plan of the receiver design. One LO per downconverter stage is required.

The LO signal mixing process generates both upper and lower sidebands of the SV signals (plus noise and leak-through signals). The IF SAW bandpass filter selects the lower sideband that is the difference between the input L-band signal and LO_1 [i.e., $f_L - (f_L - f_{IF}) = f_{IF}$]. The upper sidebands and leak-through signals are rejected in this design by the postmixer bandpass IF SAW filter. Special care must be taken with both the frequency plan and the SAW filter stage to remove all potential sources of image signals (i.e., unwanted signals that, when mixed with the first LO, end up in the IF band). The signal Dopplers and the ranging codes (buried in noise) are preserved after the mixing process produces the IF. Only the carrier frequency is lowered, but the Doppler shift (frequency offset from the carrier) of each SV signal remains referenced to its original L-band signal. The IF signal is fed to the numerical gain control amplifier (NGCA), usually called the automatic gain control (AGC), but this design specifically uses numerical gain control (NGC).

The numerical gain control (NGC) digital signal to the NGCA originates outside of the Figure 8.4 functional block diagram, but is shown in more detail in a later figure. The NGC digital signal discretely and precisely controls the NGCA gain, thereby avoiding analog control drift. This design technique has demonstrated a 60-dB dynamic range [10] plus it supports an interference situational awareness feature [10–13] described later.

Referring to Figure 8.4, the upper IF path feeds two mixers that, in combination with an in-phase IF local oscillator frequency, (I)LO_{IF}, and quadra-phase local oscillator frequency, (Q)LO_{IF}, convert the real IF signal into complex baseband in-phase (I) and quadra-phase (Q) components. The upper bands and leak-through signals out of the mixing process are rejected by their respective lowpass filters that also serve as antialiasing filters, each with half of the front-end bandwidth. As indicated in the filter diagrams, there is no carrier frequency remaining in these signals because their origin is now DC. The filtered signals are amplified and fed to two baseband ADCs.

In Figure 8.4, the lower path remains real at IF and is passed to its ADC via an antialiasing IF bandpass filter and amplifier at full front-end bandwidth. Note that typically only one path or the other is actually used. The choice usually depends on the ADC technology that is available and/or the digital signal processing power that is available to the designer, but there are clear performance advantages to the lower real IF implementation that are described later. Since the choice may be different for different GNSS signals and there is substantial investment in the development of a production front end, both options may be retained in a design with the provision that one or the other path can be powered off.

8.3.2 Gain

An estimate of the required amount of front-end voltage gain, $(G_{fe})_{dB,}$ can be calculated based on $(N_0)_{dB}$, the receiver thermal noise power in a 1-Hz bandwidth, B_{fe}, the front-end bandwidth (assumed as 30 MHz), the antenna load (assumed as 50 ohms), and the maximum ADC input voltage, assumed as 2V peak to peak. The computation sequence is shown in Table 8.4 along with the relevant equations, using the value for $(N_0)_{dB}$ that is computed in Chapter 9, but recomputed in the table, making the same assumptions (i.e., receiver noise figure, $(N_f)_{dB}$ = 2 dB and antenna temperature, T_{ant} = 100K). Recall that the GNSS signals of interest are buried in noise, so it is assumed that only thermal noise is present (i.e., no in-band interference is present and the added power from all in-band and in-view GNSS signals contribute a negligible amount of additional power to the thermal noise). The assumption that the ADC has a 1-V peak limit is based on a modern high performance, very high sample rate, and wide bandwidth 16-bit ADC (see, e.g., [14]) that limits the input voltage range to 2-V, peak-to-peak, full-scale input. So a 1-V peak or 0.7071-V RMS input to the ADC is assumed. For these assumptions, Table 8.4 shows that the maximum net front-end gain is about 110 dB. The net qualification is important because even more total gain is required to overcome all insertion losses in the front end. Note that the antenna VSWR was assumed to be a perfect 1:1 (which it never is), so this loss must also be overcome by gain.

If this front-end bandwidth is reduced to 1.7 MHz for a simple L1 C/A code receiver design, the thermal noise in that bandwidth is reduced to about –142 dBW (i.e., about 12 dB lower than the wideband case). The gain increases to about 122 dB, so about 12 dB more gain is required than the wideband case example.

In any case, all things else being equal, the actual gain is determined by the thermal noise in the front-end bandwidth, B_{fe}. Since that bandwidth does not change for a specific front-end design, the only thing that changes the gain is in-band interference (and small variations in gain due to component variations with

Table 8.4 Maximum Net Front-End Voltage Gain Computation

Symbol	*Units*	*Equation*	*Value*	*Parameter*
$(N_0)_{dB}$	dBW/Hz	$10\log_{10}[k(T_{ant}+T_{receiver})]$	−204.3	Thermal noise power in 1-Hz bandwidth
k	J/K	Constant	1.38E-23	Boltzmann's constant
T_{ant}	K	See (8.8)	100	Antenna temperature
T_{amp}	K	$290\left(10^{\frac{(N_f)_{dB}}{10}}-1\right)$	169.6	Receiver temperature
$(N_f)_{dB}$	dB		2	Receiver noise figure
B_{fe}	Hz		3.0E07	Front-end bandwidth
$(N)_{dB}$	dBW	$(N)_{dB}=(N_0)_{dB}-10\log_{10}(1/B_{fe})$	−129.5	Thermal noise power (dBW) in bandwidth B_{fe}
N	W	$N=10^{\frac{(N_{dB})}{10}}$	1.1E-13	Thermal noise power (W) in bandwidth B_{fe}
V_N	Volts (RMS)	$V_N=\sqrt{N\times 50}$	2.36E-06	Thermal noise volts RMS across 50 ohms
V_{ADC}	Volts (RMS)	$V_{RMS}=\frac{\sqrt{2}}{2}V_{PEAK}$	0.707	Maximum ADC RMS input voltage assuming 1-V peak ADC input
$(G_{fe})_{dB}$	dB	$(G_{fe})_{dB}=20\log_{10}(V_{ADC}/V_N)$	109.5	Maximum (net) front-end voltage gain (dB)

temperature, age, and so forth). As in-band interference increases, the front-end gain must decrease accordingly. The front-end design of Figure 8.4 can accommodate a very large range of gain attenuation.

8.3.3 Downconversion Scheme

The choice of downconversion scheme depends significantly on the analog microwave technology available to the designer. Monolithic microwave integrated circuit (MMIC) technology and specialized microwave components continue to improve, including reduced noise figure, feature size, and power. This technology has also increased stage isolation to unwanted conducted or radiated sneak paths. This helps to reduce the number of isolating downconversion stages that provide the enormous amount of passband gain and stopband rejection required before the analog signals are digitized. Since the inception of the first precorrelation ADC receiver, the all-digital receiver [15], these technology advances have enabled generational reductions in number of downconversion stages from triple to double and now single downconversion front-end designs. Even direct L-band digital sampling and digitization front ends have been proposed and fielded. In spite of the aforementioned technology advances, the leakage paths from so many same-frequency, high-gain stages prior to direct L-band sampling invite instability (oscillation) in the front end. However, downconversion to IF after only one LNA gain stage at L-band significantly reduces the leakage path feedback. In the design shown in Figure 8.4, the gains are distributed between two separate frequencies, L-band and IF, with most of the gain at the lower frequency. This design also permits the use of an identical, smaller, lower-cost, higher-Q and lower insertion loss SAW filter at IF as

compared to the ones used at L-band, where each SAW must necessarily be different to match its respective front-end L-band frequency. These IF gain and filter stages significantly improve the overall front-end stopband rejection performance. These two critical features (enhanced stability and stopband rejection) plus identical IF components for different front-end L-bands are the principal reasons against the use of direct L-band sampling.

8.3.4 Output to ADC

Note in Figure 8.4 that the upper signal path of the front-end output to the ADC is a complex baseband (I and Q) signal. There are both advantages and disadvantages to this scheme. The clear advantage is that the signal spectrum origin has now shifted from IF to DC and the bandwidth has been halved. In practice, the underlying GNSS signal carriers have Doppler shifts, so these offsets remain. Also, the complex IF downconversion (mixing) signal has some frequency error inherited from the reference oscillator, so this common mode offset remains. When the underlying real GNSS signals at IF are converted to a complex signal by the process shown in Figure 8.4, it is impossible for this composite signal to be a true complex representation of each underlying real signal. This is because of varying amounts of Doppler in each underlying signal combined with the reference oscillator frequency offset, and the imperfect analog 90° analog phase shift circuits. This baseband design survives if these are all small errors, but it is impossible to totally eliminate these errors. Another disadvantage is that the baseband signals can no longer be AC-coupled because the origin of the baseband spectrum is DC and analog DC paths from gain stages are subject to drift (i.e., the inclusion of DC drift in the ADC process can cause analog bias problems). However, this was the original digital receiver scheme when ADC speeds could not support the (real) analog IF signal bandwidth and it will continue to be used in some designs for the same reason. Various techniques have been developed to minimize the DC bias problem at the ADC input as well as the imperfections in the complex baseband signals.

The lower path analog IF (real) signal does not have the DC bias problems because it is AC coupled to the ADC. Typically, a DC bias is required for the modern unipolar ADC input, but this bias circuit is not subject to active gain stage drift (i.e., it is as stable as the reference voltage and resistors used in the bias circuit).

In the lower path scheme, each digital receiver channel performs the conversion of the digitized real IF signal into complex baseband components. More detail on the ADC process that converts the real analog IF signal to digital IF is presented in Section 8.3.8 and the digital receiver process that uses it is presented in Section 8.4, but the advantages of digital signal processing using the real IF signal are presented here. Since underlying GNSS signals at digital IF are processed and detected digitally, there is first a by-product of the digital sampling process in the ADC that moves the IF signal to a lower image frequency (described in Section 8.3.8). The key by-product is that each digital receiver channel essentially extracts one SV signal out of the noise (and separates it from the signals of all other SVs in view at the same carrier frequency) by a replica carrier wipe-off process followed by the replica code wipe-off process. These wipe-off processes are either open-loop when searching or closed-loop when tracking the SV (described in Section 8.7). The intent here is to focus on the advantages of using the digital carrier wipe-off process

to also perform the complex conversion on one SV signal (not all in view) with digital replica I and Q carrier signals that are assured to have a perfect 90° phase-shift. When that SV signal is found and being tracked in phase-lock by the assigned digital receiver channel, these replica signals are also essentially perfectly aligned with that one SV signal, meaning that these replica signals also contain the exact Doppler shift, the exact common-mode reference oscillator frequency and phase offsets, as well as the exact carrier frequency of the image IF. (In Section 8.3.8, it is explained how the actual IF signal is folded in frequency to a lower image IF by the ADC using undersampling.)

Since near-perfection is achieved by the digital baseband complex conversion process, the real analog IF signal is the preferred front-end output to the ADC, assuming that the ADC design can operate with IF as its i-nput and provide the required number of bits at the 2 times increase in ADC sample rate. There are also some simplifications in the digital receiver channel carrier wipe-off process that will be described later.

8.3.5 ADC, Digital Gain Control, and Analog Frequency Synthesizer Functions

The functional block diagram of the ADC that is part of each front end is shown in Figure 8.5. Both ADC options required by Figure 8.4 are illustrated in Figure 8.5. Referring to Figure 8.5, there is a pair of baseband ADCs in the upper part of the figure that are used if a complex digital baseband signal output is implemented and a single ADC in the lower part of the figure if a real digital IF signal output is implemented, but only one combination is used. In either case, the same digital gain control feature is implemented, except the signal detector is either complex or real. Note that the sampling clock for the complex signal ADCs is (approximately) half the rate for the real signal ADC. This is because the complex signal bandwidth is (approximately) half that of the real signal. Digital gain control is also shown for both ADC implementations with the J/N meter situational awareness as

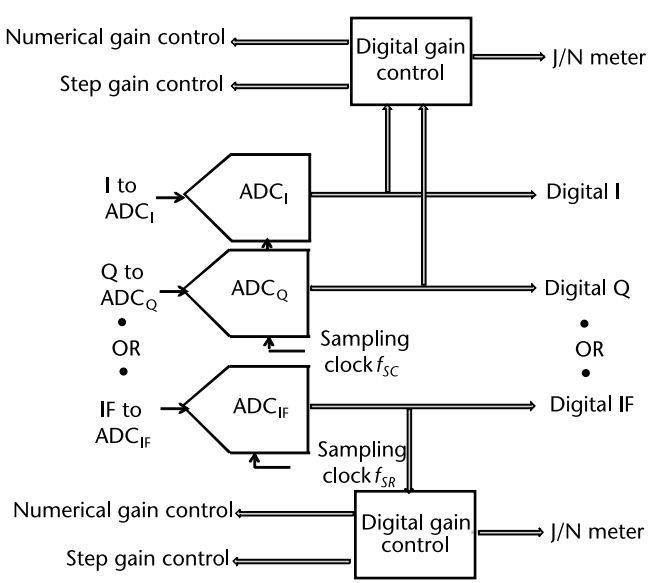


Figure 8.5 Front-end ADC options with digital gain control features.

a by-product. The detailed digital gain control functions are shown in a closed-loop form in Figure 8.6 for the real signal path. The digital gain control scheme [10] has the same functions (detector, lowpass filter, comparator, AGC gain, and error integrator) as its analog counterpart but this scheme has the advantages of precise and easy tuning, high dynamic range, and drift-free integration that make the step gain control (and its J/N meter by-product) feasible. Other situational awareness features such as interference characterization could also be implemented as part of the digital gain control design, but are not shown here.

The functional block diagram of the analog frequency synthesizers that service all M front ends is shown in Figure 8.7. It provides a unique LO_1 to each front end so that a common IF is produced. If a complex baseband signal is synthesized by the front end, then the frequency synthesizer provides a common LO_2 to each front end, typically 2 to 4 times the LO_{IF} that is produced by the complex synthesizer in Figure 8.4. All synthesized frequencies are phase-locked to the reference oscillator.

The digital frequency synthesizers (shown at a high level in Figure 8.1) are also phase-locked to the reference oscillator and provide the ADC sample rate(s) to all M front ends and also provide the set-time interrupts to all receiver channels. It is prudent to provide separate power supply regulators for the analog and digital circuits of the receiver to keep them free from ground loops by providing separate ground paths joined only at one return point near their source power and to keep them as physically separate as practical to minimize radiated cross-talk. This is further emphasized by illustrating the ADC diagram as functionally separate from the analog part of the front-end even though every front end requires its own ADC.

8.3.6 ADC Implementation Loss and a Design Example

The finite quantization level of the ADC causes implementation loss in the GNSS receiver. This loss is described in this section and a flash analog-to-digital converter design example is presented. Historically, all signal processing was performed in the time domain because GNSS receivers operate in real time. Time-domain baseband processing permits the number of ADC bits to be quite small, thereby greatly simplifying the design and increasing its sampling rate for the technology level existing at the time. Frequency-domain processing in the search engine has become popular

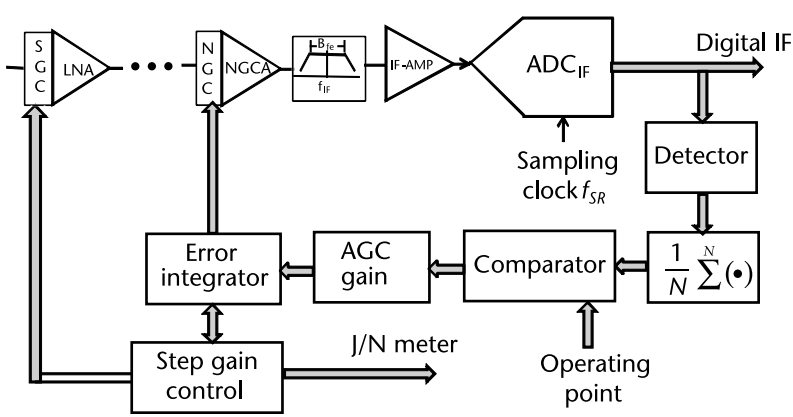


Figure 8.6 Front-end expanded digital gain control features in the closed loop.

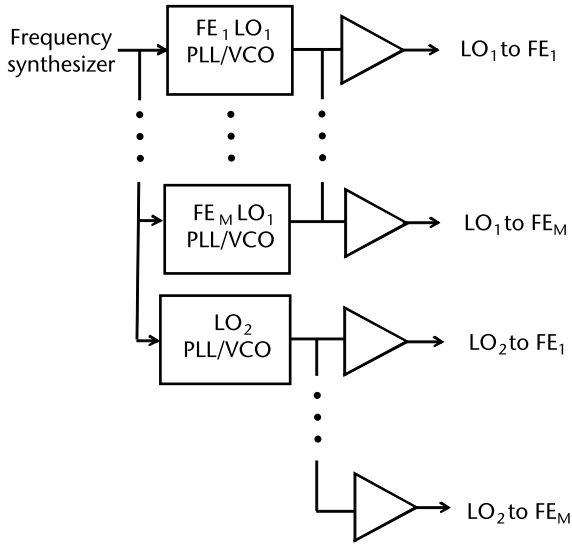


Figure 8.7 Front-end analog local oscillator frequency synthesizers.

since current ADC and signal processing technology support the required 12-bit or higher ADC resolution. When such ADCs are used the implementation loss issue becomes a moot point. However, there are real-time techniques that can perform the search engine functions effectively, so the implementation loss issue remains important.

First generation GPS C/A-code digital receivers used 1-bit and 2-bit ADCs with theoretical implementation losses of 1.96 dB and 0.5495 dB, respectively [16]. Theoretically, the 1-bit ADC requires no automatic gain control, but practically some form of assured minimum and maximum amplitude is required to achieve uniform decision performance by the one analog comparator with enough threshold hysteresis to avoid oscillation. Modern high-performance time-domain GNSS receivers use only 3-bit or 4-bit ADCs. As will be seen, there are diminishing returns in ADC implementation loss reduction with more bits, but the automatic gain control must adjust the input RMS (one-sigma) analog noise level optimally for both quantization and clipping noise in the ADC. In choosing an ADC with an acceptable implementation loss specifically for a GNSS receiver that only operates in the time domain, [17] provides the most comprehensive and accurate results for a wide range of quantization levels as well as the optimum one-sigma amplitude for the analog input noise. The implementation losses due to aliasing are removed in [17] because an ideal antialiasing filter is used in the analytical model. This is beneficial with respect to the selection of an ADC based solely on its contribution to implementation loss, but the sampling rates used in the analytical model would be misleading with respect to the rejection of aliasing since it is impossible to synthesize an ideal antialiasing filter. This issue in the context of ADC design is addressed in Section 8.3.7.

The effective signal power losses in decibels for several ADC quantization levels from [17] are shown in Figure 8.8. Certain quantization levels, n, have been labeled with the associated number of ADC bits, beginning with the 1-bit label for $n = 2$, 1.5-bit label for $n = 3$, 2-bit label for $n = 4$, and so forth as shown in Figure 8.8. Each level is plotted as a function of the ratio of maximum threshold to one-sigma

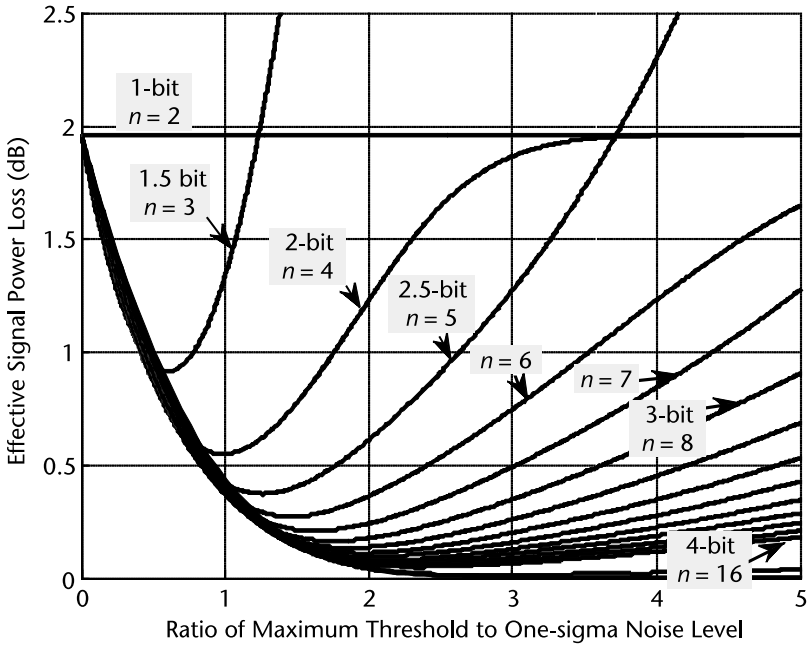


Figure 8.8 GNSS receiver ADC implementation loss for several quantization levels.

noise level. The maximum threshold corresponds to the peak input voltage level of the ADC (as opposed to the peak-to-peak input voltage level). Table 8.5 provides the exact ratio of maximum threshold to one-sigma noise level (shown as T in the table) that could be approximated from Figure 8.8. The table also includes the optimum quantization level, Q, not shown in Figure 8.8, but is the basis for the determination of T in Table 8.5. Reference [17] defined this as $T = \frac{(n-2)Q}{2}$, where n is the number of ADC quantization levels (as depicted in Figure 8.8).

The optimum peak-to-peak ADC reference voltage, V_{REF}, and the corresponding optimum one-sigma (RMS) noise input for a particular ADC (except the 1-bit ADC) are shown in Table 8.6. The first three columns of Table 8.5 are repeated for reference convenience. The optimum peak-to-peak ADC reference voltage column is computed using $V_{REF} = n \cdot Q$ where Q is taken from Table 8.5 for the same n. The optimum one-sigma column is computed using $V_{RMS} = (V_{REF}/2)/T$ where T is taken from Table 8.5 for the same n.

Inspection of Figure 8.8 reveals that a value of 1 on the abscissa corresponds to a one-sigma (RMS) noise level that produces ADC clipping noise when the optimum ratio of $(V_{REF}/2)$/one-sigma is in this region. This happens with the lower ADC quantization levels with larger quantization noise so that some clipping noise is beneficial in optimizing implementation loss. Clearly, the optimum one-sigma levels are very gain sensitive for these lower ADC quantization levels. In contrast, for the ADCs with higher numbers of quantization levels (e.g., 3 bits and more), there is virtually no clipping noise for the optimum one-sigma region and there is considerably less gain sensitivity (i.e., the curve is flatter in the optimum region). So, for the higher quantization levels, the ADC reference voltage does not have to

Table 8.5 Minimum ADC Implementation Loss (L) at Optimum *Q* and *T* Value

N (bits)	*n (levels)*	*L (dB)*	*Q (volts)*	*T (ratio)*
1	2	1.961	N/A	N/A
1.5	3	0.916	1.224	0.612
2	4	0.549	0.996	0.996
2.5	5	0.372	0.843	1.265
3	8	0.166	0.586	1.758
4	16	0.05	0.335	2.345
5	32	0.015	0.188	2.82
6	64	0.005	0.104	3.224
7	128	0.001	0.057	3.591

Table 8.6 Optimum ADC Reference Voltage and One-Sigma Input Noise Level

N (Bits)	*n (levels)*	*L (dB)*	V_{REF} *(volts P to P)*	*One-sigma (volts RMS)*
1	2	1.961	N/A	N/A
1.5	3	0.916	3.672	3
2	4	0.549	3.984	2
2.5	5	0.372	4.215	1.666
3	8	0.166	4.688	1.333
4	16	0.05	5.36	1.143
5	32	0.015	6.016	1.067
6	64	0.005	6.656	1.032
7	128	0.001	7.296	1.016

be the optimum reference voltage to three decimal places, but the peak-to-peak ADC reference voltage should be approximately $n{\cdot}Q$.

Note that the Table 8.5 implementation losses for the 1-bit and 2-bit ADCs are in agreement with those from [16]. Also note that there are diminishing returns in reduced implementation losses for quantization levels higher than for 3-bit or 4-bit ADCs.

There are numerous ADC designs, each with unique performance advantages and disadvantages. Reference [18] is an excellent (downloadable) book that provides insight into all aspects of data conversion, including its interesting history. The flash ADC design [19] is a popular choice for low-bit ADC applications because every possible analog quantization level (except zero, which is detected by default) is continuously detected using an analog comparator for each level (except zero) whose discrete output is fed to a digital flip-flop (e.g., 7 analog comparators feeding 7 digital flip-flops for a 3-bit ADC).

Figure 8.9 is a schematic of a 3-bit (8-level) analog-to-digital flash converter [19]. The analog input is connected to the positive side of all 7 comparators. The negative sides are each connected to a resistor string that receives a constant current (from the reference voltage). Each resistor junction in the string provides the negative side of each comparator with a reference voltage that is one least significant bit

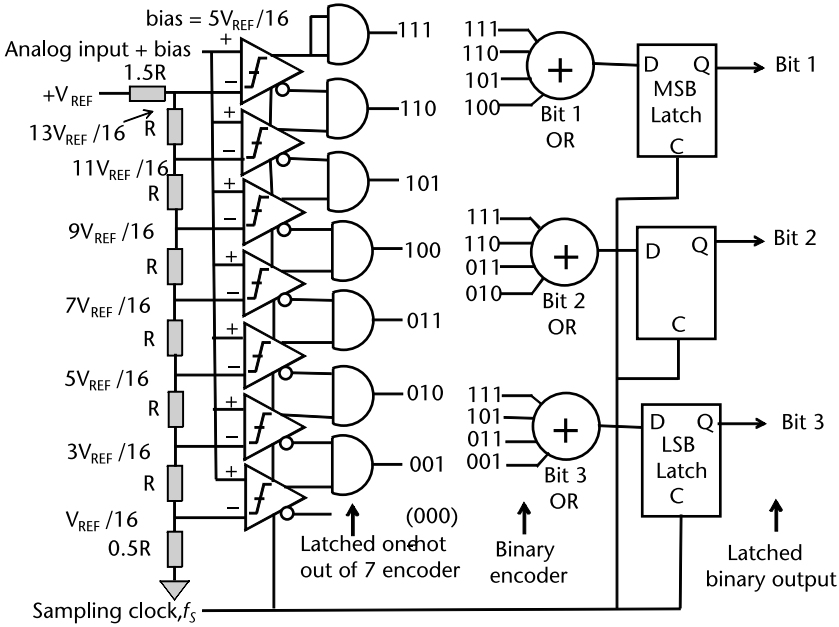


Figure 8.9 Schematic of 3-bit (8-level) analog-to-digital flash converter.

(LSB) higher than the one below it, except the LSB comparator that has a voltage reference of 0.5-LSB. The output of this comparator string is comparable to a mercury thermometer (i.e., when the analog input voltage rises, the number of 1s are rising from bottom to top proportionally, one LSB value at a time, and the number of 0s appear discretely from top to bottom as the analog input voltage decreases). However, the AND gates between comparator differential outputs provide a "one-hot out of 7" result at their outputs (i.e., only the AND gate at the highest input level produces a 1 output). The sampling clock momentarily samples and holds the current ADC decision allowing the appropriate "one-hot out of 7" AND gate (or none) to produce a 1 in accordance with current analog input level, the binary decoder OR gates convert that level into its corresponding binary state as input to the 3 flip-flops, and this state is latched into flip-flops. There is a precise and small amount (less than 0.5 LSB) of hysteresis in each comparator that prevents it from oscillating near its threshold. The symbol on each comparator signifies that there is hysteresis in the decision process. This binary conversion process is clearly shown in Figure 8.9. Figure 8.10 shows the input to output transfer function of this 3-bit analog-to-digital flash converter (including hysteresis) using a unipolar $V_{REF} = 4$V ($Q = 0.5$V) with a DC bias of 1.25V to accommodate the bipolar analog input. The computed optimum one-sigma ADC input level is $2/T = 1.138$V RMS. Note from Tables 8.5 and 8.6 that the optimum $Q = 0.586$V ($V_{REF} = 4.688$V) and optimum one-sigma ADC input level is 1.333-V RMS. The optimum DC bias for this unipolar 3-bit ADC would be the reference voltage at the 011 comparator or 1.466V (i.e., using the bias equation shown in Figure 8.9). There is a negligible implementation loss penalty for the practical choices made for this design example.

The flash ADC design closely fulfills the ideal requirements of an ADC. (1) The analog signal must be sampled idealistically with zero aperture time, but realistically sampled in a time width that is short enough that the highest frequency

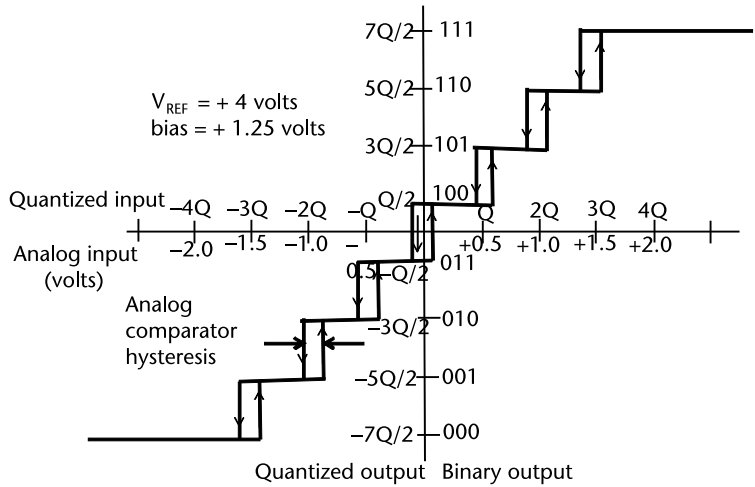


Figure 8.10 Input to output relationship for 3-bit analog-to-digital flash converter.

present (f_c) in the analog signal will change in amplitude by less than half the ADC LSB decision level. (2) The sampled analog signal must be converted into its digital representation idealistically with zero delay, but realistically in a time width shorter than the sampling clock period. (3) The sampled analog signal should be quantized idealistically instantaneously, but realistically must be held with no more error than half the ADC LSB until it is quantized.

8.3.7 ADC Sampling Rate and Antialiasing

There is a sampling rate guideline that is unique to applications involving spreading codes such as GNSS [20]. This guideline is that the ADC sample rate should not be synchronous with the GNSS carrier frequency or the GNSS signal spreading code rate (sometimes called commensurate sampling) (i.e., $f_s \neq f_{carrier}/k$ where k is any integer, which is equivalent to $f_s \neq R_c k$, where R_c is the spreading code chipping rate). This is because the digital correlation envelope becomes distorted into a symmetrical staircase if these frequencies become synchronous, instead of the required symmetrical triangle within the code correlator envelope [20]. In other words, the spreading code symbols need to be sampled at various places in time over the spreading code period. It is not uncommon for commensurate sampling to happen (with some protection provided by the effect of code Doppler frequency offset) because of the common misunderstanding of the Nyquist theorem that leads to the false conclusion that the minimum sample rate can be twice the spreading code rate. Since all frequencies synthesized by the frequency synthesizers are phase-locked (synchronous) to the reference oscillator, it is not an uncommon practice that f_S happens to be synchronous with integer multiples of 1.023 MHz in some GNSS receiver designs. When any sample frequency epochs consistently align with the code transition boundaries of the incoming signal the code tracking loop discriminator becomes nonlinear (i.e., does not consistently produce true code tracking error).

The Nyquist theorem (also known as the sampling theorem) assures that the digital signal output of the ADC is a faithful reproduction of the sampled and quantized analog signal. Consider the case for a signal centered at DC (i.e., a baseband

signal). Although the underlying theory is very complex, the Nyquist theorem can be stated very simply for this case: The sampling frequency, f_S, must be equal to twice the highest frequency component, f_C, present in the analog signal (i.e., $f_S = 2 f_C$). A practical difficulty with the Nyquist theorem is that it requires f_C to be at the stopband frequency where the signal attenuation is infinite, but the reality is that infinite signal plus noise rejection is impossible by any analog filtering technique. To the extent that the antialiasing filter fails to remove the higher frequency signals, these will all be folded back on integer boundaries of $0.5 f_S$ [21]. The consequence of the presence of any signals above f_C is that these signals are aliased back into the digitized signal and cannot be removed by any further digital signal process. Therefore, the compromise is that some aliasing must be accepted. Most theoretical models assume that there is an ideal filter that perfectly satisfies the Nyquist theorem. This can lead the designer into the false assumption that the same sampling frequency used in the theoretical model can be used in the ADC design.

The acceptable compromise in baseband ADC aliasing when the sampling frequency equals $2f_C$ is illustrated in Figure 8.11 [22]. The ADC input signal in this figure is the I or Q output of the analog front-end depicted in the top right-hand corner of Figure 8.4. The compromise is the dynamic range (DR) of signal attenuation in the transition band between the signals of interest ending at $f_B = B_{fe}/2$ and at the stopband frequency, f_C. For a given compromise DR, the width of the transition band depends on the rolloff rate of the lowpass antialiasing filter. Note that f_S is always greater than 2 times the bandwidth of the signals of interest.

Referring to Figure 8.11, the digitized baseband signal occupies not only its original spectrum zone but also multiple image zones, each called a Nyquist zone (NZ) with folding boundaries at intervals of $0.5\ f_S$ [22]. The baseband image occupies NZ(1) with the spectrum ordered the same as it was in the analog world and in all higher odd zones thereafter. The spectrum is reversed in NZ(2) and all even zones thereafter. So the high frequency end of an even zone is adjacent to the high frequency end of the next lower odd zone. Any spectrum overlap from one zone to another produces aliasing. Note that if the unwanted signals plus noise have not been filtered at the input of an infinite bandwidth sampler, then any frequency component that falls outside the Nyquist bandwidth in any Nyquist zone is aliased back into NZ(1). A finite bandwidth (real-world) sampler will have an attenuating effect at frequencies above its bandwidth, but the residue will be aliased into NZ(1).

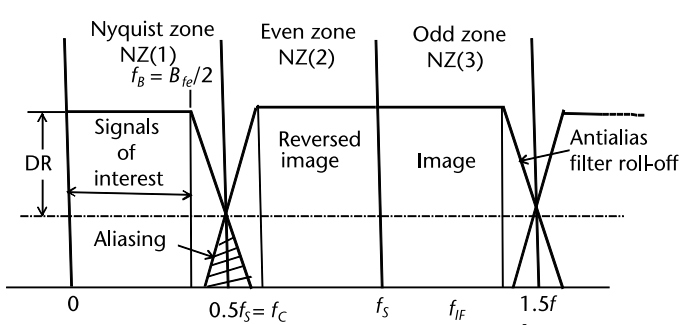


Figure 8.11 Baseband ADC aliasing when sample frequency equals $2f_c$.

It should now be clear from Figure 8.11 that in the real world of GNSS analog signals (buried in noise) being sampled and digitized at baseband, an antialias filter that completely removes all unwanted signals above f_C is unrealizable. The objective of a practical antialiasing filter is to filter the unwanted signals (plus noise) to a level where the aliasing results in negligible aliasing noise to the GNSS digital signal processing that follows. The amount of antialiasing filtering required is $(\mathrm{DR})_{\mathrm{dB}} = 20\log_{10}(\mathrm{DR})$, the attenuation required between the passband occupied by the signals of interest ending at f_B and the stopband frequency, beginning at f_C for all higher frequencies. The antialiasing compromise at DR is related to the dynamic range of the ADC. Since DR defines the required stopband attenuation and the number of ADC bits defines the required fidelity, then DR should be greater than the ADC dynamic range. The dynamic range of an N-bit ADC (as a ratio) is $(2^N - 1)/1$, so

$$\left(ADC_{DR}\right)_{dB} = 20\log_{10}\left(2^N - 1\right)\ \text{(dB)} \tag{8.11}$$

A typical compromise is to set the stopband attenuation to half the LSB weight of the ADC using

$$\left(DR\right)_{dB} \le -20\log_{10}\left(\frac{2^N - 1}{0.5}\right) = -20\log_{10}\left(2^{N+1} - 2\right)\ \text{(dB)} \tag{8.12}$$

This obtains $(ADC_{DR})_{dB} + (DR)_{dB} = -6$ dB of aliasing attenuation below the ADC dynamic range.

For a given DR the actual amount of aliasing noise is determined by the margin of NZ separation (i.e., the margin provided by f_S exceeding $2f_c$). Note in Figure 8.11 that the shaded area below the DR line is the aliasing that takes place when $f_S = 2f_C$. By inspection, if the sampling rate is decreased, the aliasing will be higher than required by DR. Similarly, if the sampling rate is increased, the aliasing will be lower than required by DR.

Baseband sampling implies that the analog signal that was sampled lies in the first Nyquist zone, NZ(1), as shown in Figure 8.11 that includes DC on its left boundary, so a lowpass antialiasing filter is required. A popular lowpass antialiasing filter choice for time-domain receivers is the Butterworth lowpass filter, characterized as a maximally flat analog filter with linear phase response. It has an almost perfect flat response from DC to near the corner frequency f_B. The rolloff at this corner frequency is typically specified at the –3-dB point (i.e., the typical lowpass filter bandwidth is defined by where it is down by 3 dB). Starting at the corner frequency, the rolloff rate is 6 dB per octave per pole used in the filter design.

For example, assume that only time-domain processing is used and a 3-bit ADC is selected. Using (8.12), $(DR)_{\mathrm{dB}}$ is about –23 dB. Further assume that the GPS L5 signal (with a 10.23-Mcps spreading code rate) is downconverted to the complex I and Q baseband signals shown in the top right corner of Figure 8.4. Since there are nulls (no signal energy) in the L5 signal at exactly ±10.23 MHz away from the carrier, it can be shown that there is less than 0.2 dB of additional signal power loss, compared to the signal power loss for a 20.46-MHz bandwidth, if B_{fe} = 17 MHz. Therefore, assume that the single-sideband corner frequency of the antialias filter is f_B = 8.5 MHz. If the acceptable stopband frequency is 1 octave

away at $f_C = 17$ MHz and $f_S = 34$ MHz, then the Butterworth filter would require a minimum of 4 poles for –24 dB of stopband attenuation [i.e., a fairly simple lowpass filter design that provides 1-dB margin to the acceptable $(DR)_{dB}$]. Keep in mind that some antialias filtering has already been performed in the earlier stages of the front end, so this could provide additional margin. Also, the GNSS signals are well below the thermal noise level at this point. However, the thermal noise is actually beneficial to the operation of the ADC because it provides dithering to the quantization process [23].

Antialias filters become more complex as the transition band becomes sharper, all other things being equal. The combination of sharper transition bands and higher quantization bit ADCs typically requires another filter type such as an elliptic filter that has the desired attributes of in-band flatness, a sharp transition band, and linear phase response. The antialias bandpass filters for the ADCs that quantize the IF signal, described in Section 8.3.8, can use SAW filters (or similar technologies) that have bandpass flatness, very sharp transition bands, and acceptable phase response linearity. Typically, all of the front-end filters are purchased from specialists in these designs, but it is essential to understand how to specify and verify them.

8.3.8 ADC Undersampling

ADC undersampling can be thought of as digital signal downconversion of the real IF signal that is later converted into complex digital I and Q components by each digital receiver channel. This eliminates the need for the IF demodulator shown in the top right corner of Figure 8.4 and replaces it with a precise, drift-free, all-digital process [22]. The demodulator performance benefits of this scheme were described earlier in Section 8.3.4. Figure 8.12 shows a case example of the digital frequency spectrum for an undersampled ADC with the IF centered in the third Nyquist zone, NZ(3). Higher odd-numbered Nyquist zones can be used if the front-end IF is higher and/or the sampling rate is lower. Recall that the spectrum is reversed in even-numbered Nyquist zones, so the IF should never be placed in even-numbered Nyquist zones for GNSS receiver designs. Undersampling the IF signal above the first Nyquist zone is advantageous because it aliases the IF signal from the target Nyquist zone, in this case example NZ(3), down to the first Nyquist zone NZ(1) where the origin is DC and the bandpass center frequency is lowered to 0.25 f_S.

This technique requires only one ADC (as shown at the bottom of Figure 8.5), but this is no ordinary ADC. This ADC design must operate compatibly with the

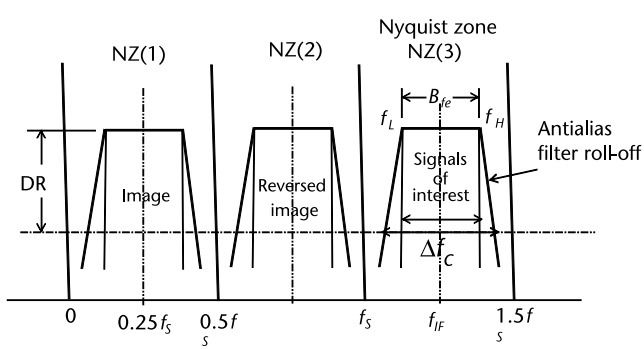


Figure 8.12 Undersampled ADC with the IF centered in third Nyquist zone.

front-end IF that is much higher than its baseband ADC counterpart and with twice the signals of interest bandwidth. As the IFs become higher, the dynamic performance requirements on the ADC become more critical [22]. ADCs designed to operate in the first Nyquist zone (i.e., at baseband) will not be adequate for undersampling applications because they cannot maintain dynamic performance into the higher order Nyquist zones. However, current ADC technology supports even 16-bit operation for GNSS IF signals in odd Nyquist zones [14].

There are two equations required to select the sampling frequency for the undersampled ADC. The first is the practical Nyquist criteria restated for the nonbaseband case as

$$f_S \geq 2\Delta f_C \quad \text{(Hz)} \tag{8.13}$$

where Δf_C is the bandwidth of the bandpass antialias filter at DR. Recall that the antialiasing filter provides a stopband that has theoretically (according to the original Nyquist theorem) reached an infinite attenuation level, but practically has reached the level defined by DR. Figure 8.12 illustrates how DR defines the bandwidth, Δf_C, where all signals outside of that bandwidth have been filtered down to an acceptable aliasing noise level.

The second is the Nyquist Zone (NZ) equation stated in terms of the ADC sampling rate as

$$f_S = \frac{4f_{IF}}{2NZ - 1} \quad \text{(Hz)} \tag{8.14}$$

where f_{IF} is the front-end carrier frequency that has been downconverted to IF and NZ = 1, 3, 5, 7… corresponding to the odd Nyquist zone into which the digitized IF falls. Recall that if an even NZ is chosen, the spectrum becomes reversed in NZ(1), which is the zone that is used by the receiver channels. Thus NZ is chosen to be an odd zone. The sharper the antialiasing filter bandpass roll-off, the narrower will be Δf_c at the required DR, thereby lowering the minimum required sampling rate and enabling NZ to be larger as determined by (8.14). There can be tradeoffs between the complexity of the antialiasing filter by choosing smaller values of NZ and a higher sampling rate.

For this case example, assume that the IF is centered at 140 MHz and B_{fe} = 17 MHz (or wider) using a 3-bit ADC. Note that this is the same front-end bandwidth and ADC fidelity used for the baseband case example, but the ADC input bandwidth is twice as wide as the complex baseband ADC input bandwidth and has a carrier frequency of 140 MHz instead of DC. The DR for a 3-bit ADC requires that all frequencies on either side of Δf_C be attenuated by 23 dB or more. A commercial off-the-shelf (COTS) 140-MHz center frequency inductor-capacitor (L-C) bandpass filter with less than 1-dB insertion loss and a 20-MHz passband at the –3 dB points provides a minimum of 40 dB of attenuation at Δf_C = 50 MHz. This will take care of the antialiasing problem even for a 6-bit ADC. To compute the sampling rate and the Nyquist zone, first determine the minimum sampling frequency using (8.13). So assume f_S = 100 MHz and f_{IF} = 140 MHz to solve (8.14) for NZ, which yields NZ = 3.3, but NZ must be an integer and odd, so set NZ = 3. Substituting NZ = 3 and f_{IF} = 140 MHz in (8.14) increases f_S to 112 MHz, which passes

the requirement for (8.13), that is, slightly oversampled. As shown in Figure 8.10, NZ(3) is aliased into NZ(1) where the center frequency is 0.25 f_S = 28 MHz. Since this alias originates from NZ(3), the center frequency can also be calculated as $f_{IF} - f_S$ = 28 MHz because f_S happens to be the left boundary of that zone. Note that when the first usually noninteger solution for NZ in (8.14) lands in an even zone, it will be necessary to move down to a lower odd zone in order to comply with (8.13), thereby increasing the sampling rate. In this case example, it landed in an odd zone, but when that noninteger was rounded down to its integer odd value, the sampling rate increased when (8.14) was recomputed with the integer zone value. As the last step, always verify that (8.13) remains true after determining the final sample rate from (8.14).

Sampling signals above the first Nyquist zone is beneficial because it reduces the sampling frequency. For example, recall that f_S was 34 MHz for the baseband case example, making the NZ(3) case example higher by a factor of 3.3, but this is because the design had substantial antialiasing design margin. With the margin cut down to the minimum DR required by the 3-bit ADC, the same IF antialiasing bandpass filter provides –24-dB attenuation for Δf_C = 30 MHz with $f_S \geq 60$ MHz. This places the digital signal in NZ(5) with f_S = 62.22 MHz. The NZ(5) case example is higher by only a factor of 1.8. The rule-of-thumb ratio would be 2 since there is a factor-of-2 wider bandwidth for the undersampled design. This clearly illustrates the benefits of better antialiasing filter technology and moving higher in the NZ zone number. For example, SAW bandpass filter technology is excellent in the 140-MHz region and is a good candidate when high ADC bit levels are required.

Undersampling is widely used in communications applications because the process eliminates the need for an analog IF demodulator and filters and replaces it with a superior digital signal processing technique. It can also be used effectively for GNSS receiver front-end design to move digital technology further into the analog domain with superior results. Clearly, the undersampling technique becomes even more relevant if frequency-domain processing is used in the digital receivers for fast signal acquisition. Frequency-domain conversion of time-domain signals requires a much larger number of bits in the ADC with subsequent higher DR for the antialiasing filter, but SAW bandpass filter technology has evolved with excellent performance at low cost that can play a critical role in this area at IF. The 16-bit ADC in [14] operates at IF frequencies beyond 200 MHz at sampling rates up to 125 Msps, demonstrating that ADC technology is available to support the undersampling techniques described above.

8.3.9 Noise Figure

The noise figure of the receiver in decibels is determined from the equation

$$\left(N_f\right)_{dB} = 10\log_{10} N_f \quad \text{(dB)} \tag{8.15}$$

where N_f is the dimensionless noise factor (a ratio) determined by the Friis formula that calculates the cascade of all front-end stages, each stage with its own noise factor and gain as follows

$$N_f = N_{f1} + \frac{N_{f2} - 1}{G_1} + \frac{N_{f3} - 1}{G_1 G_2} + \ldots + \frac{N_{fn} - 1}{G_1 G_2 G_3 \ldots G_{n-1}} \quad \text{(ratio)} \tag{8.16}$$

where N_{fi} and G_i are the noise factor and power gain, respectively, of the ith stage (and assuming that the impedances are matched at each stage). Note that both values are in units of ratio, not decibels. Since the noise factors and power gains of the components are usually specified in decibels, their conversions into ratios for use in the Friis formula are

$$N_{fi} = 10^{\frac{(N_{fi})_{dB}}{10}} \quad \text{(ratio)} \tag{8.17}$$

and

$$G_i = 10^{\frac{(G_i)_{dB}}{10}} \quad \text{(ratio)} \tag{8.18}$$

These equations will be used in a case example later. The GNSS receiver noise figure is approximately the noise figure of the first LNA if there is negligible insertion loss prior to the LNA, the LNA has a low noise figure and there is sufficient gain in the LNA. In the case example front-end design, the noise figure is essentially determined by the sum of the insertion loss of the cavity filter and the LNA noise figure.

8.3.10 Dynamic Range, Situational Awareness, and Effects on Noise Figure

Note that there is a step gain control feature shown with the LNA. This feature is provided along with a numerical gain control (NGC) feature in the NGCA that provides not just automatic gain control, but precise discrete front-end gain management of the front end. The combination of 10, 28, and 46-dB step gain control (36-dB range) on the LNA plus 60-dB dynamic range of the NGCA for each step provides a front-end dynamic range of more than 96 dB. The precise measure of gain change provides a precise measure of in-band interference change above the thermal noise level. This receiver situational awareness design feature is called a jamming to (thermal) noise power ratio meter (i.e., a J/N meter [10–13]). The synergism between dynamic range robustness to in-band interference power in the front end and situational awareness in the receiver control function supports optimization of the operational states of the receiver channels to work in harsh environments. Some of these strategies are described in Sections 8.5, 8.6, and 8.7.

It should now be clear that the front end provides enough gain to bring the input thermal noise level up to an RMS voltage level that is optimum for the ADC input. Recall that the GNSS signals arrive well below the thermal noise level. The NGC of the NGCA optimizes the signal level at the ADC input based on the input noise level, so in the absence of interference or jamming, this level is determined by thermal noise. Note that when in-band interference and/or jamming are present, then the NGCA reduces gain proportionally, so the highest front-end gain occurs in the presence of only thermal noise. As in-band interference increases (and this means any unwanted signal power that gets past the front-end bandpass filters

including adjacent band interference), the NGCA in combination with the SGC systematically reduce the gain to maintain the optimum ADC RMS voltage level. The gain control function also maintains an accounting of the $(J/N)_{dB}$ level, where $(N)_{dB}$ is associated with the gain for the unjammed thermal noise power level and $(J)_{dB} - (N)_{dB}$ is associated with the amount of gain reduction caused by the in-band jamming.

Inevitably, there is a step increase in front-end noise figure when the SDC function steps the LNA gain down to prevent it from gain compression due to increased in-band interference. However, low-noise figure is not as important under severe interference conditions as otherwise, so this is an excellent receiver performance trade-off and it can be well managed with excellent LNA stepped gain design. A case example demonstrates this effect using straw-man stage values and the Friis formula as shown in Table 8.7.

Table 8.7 shows the assumed insertion losses and noise figures for the eight stages of the front end, including the signal path via the ADC used by the IF output signal. These values are used to compute the corresponding ratios (factors) used by the Friis formula to compute each stage ratio. The ratios computed at each stage shown in the bottom row must be summed and converted into decibels to obtain the total noise figure, in this case for the highest LNA gain setting (46 dB). Table 8.8 shows this sum and the consequent noise factor and noise figure in decibels for each of the three LNA gain steps assuming the LNA retains its noise figure of 1.5 dB and all other stages retain their same noise figure and gain. The total gain for the Table 8.7 values at input of the ADC is 135 dB and the total insertion loss is –30 dB, with a net gain of 105 dB.

Several things should now become apparent. The L-band cavity filter insertion loss of 1 dB is a major contributor to the total front-end noise figure, so this is the reason why passive losses ahead of the LNA are minimized to the extent practical. The LNA noise figure sets the total noise figure of the front end with negligible contributions from later higher noise figure stages if the LNA gain is high enough (neglecting the passive loses prior to the LNA that increase this noise figure). The noise figure of this front-end design is remarkably good considering that it provides

Table 8.7 Using the Friis Formula to Analyze Front-End Noise Factor

Stage		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Symbol	Device Units	Cavity	LNA	SAW	Mixer	SAW	NGCA	Amp	ADC
$(N_{fi})_{dB}$	dB	1	1.5	4	10	2	5	5	30
$(G_i)_{dB}$	dB	0	46	0	9	0	60	55	0
N_{fi}	ratio	1.25893	1.41254	2.51189	10	1.58492	3.16228	3.16228	1000
G_i	ratio	1	39,810.7	1	7.94328	1	10,000	10,000	1
$\frac{N_{fn}-1}{G_1 \quad G_{n-1}}$	ratio	1.25893	0.41254	3.80E-05	2.26 E-04	1.85E-06	6.84E-06	6.84E-10	3.16E-11

Note 1: The Friis formula is extremely sensitive to round-off error. Some of the above ratio entries were rounded-off to fit legibly in the table, but all computations were performed at full spreadsheet precision.

Note 2: The determination of the ADC noise figure is based on the Analog Devices tutorial in [24].

Table 8.8 Front-End Noise Factors and Noise Figures for Three LNA 18-dB Step Gains

LNA Gain (dB)	N_f *(ratio)*	$(N_f)_{dB}$ *(dB)*
46	1.671735691	2.231676146
28	1.688671351	2.275451356
10	2.757239279	4.404744568

substantial and sharp stopband protection prior to the LNA as well as an additional 36 dB of step gain dynamic range to attenuate in-band interference.

This case example is not intended to define the actual gains and noise values that would result in an actual front-end design since there are numerous design and technology factors that alter the final frequency and gain plan of a GNSS front end as the design evolves, but this case example does illustrate the methodology that goes into an actual design to track the front-end noise figure while the design is evolving. During that evolving design, both an operational (discrete component) breadboard and spreadsheet should be maintained to track the impact of any design change. The spreadsheet should account for the minimum, typical and maximum values that can occur in production of the front-end product.

8.3.11 Compatibility with GLONASS FDMA Signals

The fundamental difference between GLONASS FDMA and CDMA used by other GNSS constellations is that all GLONASS FDMA signals are spread by the same PRN sequence, so they must be separated by differences in carrier frequency in order to demodulate them without cross-interference between the GLONASS SVs. What FDMA and CDMA have in common is that both use DSSS modulation and demodulation techniques, so the objective is to pursue an architecture that takes advantage of this commonality.

The baseline front-end design of Figure 8.4 is for CDMA operation. The bandwidth of the front end that provides signal conditioning for GNSS CDMA signals is determined by the spreading rate of the PRN code in the signal and the way the common carrier frequency is modulated. CDMA signals with the same PRN code design can occupy the same spectrum bandwidth because each SV has a unique PRN sequence. Since the GLONASS PRN code sequence is identical, their signals cannot use the same carrier frequency. There are 14 different GLONASS L1 carrier frequencies used, each one separated from its neighbor by 0.5625 MHz for a total bandwidth occupation of 7.875 MHz. The code length is 511 chips, spreading code rate 0.511 Mcps, code period 1 ms, data period 10 ms, and modulation type BPSK. The classical GLONASS FDMA receiver maximizes analog and minimizes digital technology, so one front end is required for each SV being tracked and each front end is tuned to the carrier frequency of the specific SV to be tracked; for example, there are 14 possible center frequencies of $1{,}602 + 0.5625N_G$ MHz, where $N_G = -7, -6, \ldots, 0, +1, +2, \ldots, +6$.

The front-end design of Figure 8.4 can be adapted to signal condition the entire L1 band of GLONASS FDMA signals. That allows one front end to provide signal conditioning for all 14 L1 GLONASS carrier frequencies using an L1 center

frequency at the middle of the GLONASS L1 band (1,601.71875 MHz) with a minimum passband of B_{fe} = 7.875 MHz. The IF would be the same as for the CDMA signals, so 140 MHz is assumed based on earlier case examples. The undersampling technique is also assumed so that the digitized GLONASS band is folded into NZ(1) (i.e., at a low IF for use by the receiver channel baseband processing functions). Table 8.9 summarizes the case example design parameters for the GLONASS L1 band. Note that the passband must be flat over the GLONASS 7.8750-MHz bandwidth so the filters used should be wide enough to ensure that there is no signal roll-off at either end of the passband.

Each digital receiver channel assigned to track GLONASS SVs would select the digitized carrier frequency signal of interest by synthesizing a complex replica carrier frequency (plus Doppler) corresponding to the L1 GLONASS signal of interest. This replica is multiplied (in the time domain) with the incoming digital IF signal at the ADC sampling rate. This is called carrier wipe-off. All 14 of the carrier frequencies have been downconverted to digital IF plus or minus their respective frequency offsets and they remain separated by 562.5 kHz from their nearest neighbor (the same frequency separation as at L-band). So the frequency-domain convolution process that takes place rearranges them with the same frequency separation, but the selected frequency is placed at baseband converted to a complex signal (when the replica matches the incoming signal). Then the code wipe-off process is performed on the complex baseband signal using the replica GLONASS L1 PRN code (plus code Doppler) synthesized at its 0.511Mcps spreading code rate (plus code Doppler). This time-domain correlation process (also at the ADC sample rate) results in despreading (wiping off) the code (when they match) at baseband. There is some correlation possible with the other SV signals but these remain separated in the frequency domain by 562.5 kHz. There is a lowpass integrate-and-dump filtering process that takes place following the two wipe-off processes (i.e., integrate at the ADC sampling rate and dump the complex results for signal detection and subsequent tracking after a time period shorter than or equal to the data bit period of 10 ms). If the receiver knows the GLONASS time, then it can align the integrate-and-dump periods with the data transitions and use the maximum 10 ms for the integrate and dump to improve the baseband signal-to-noise ratio. These receiver channel synthesis techniques are described in more detail in Section 8.4; however, it should be apparent that all of the GLONASS signals are being multiplied in the time domain but with separation in the frequency domain. There may or may

Table 8.9 Case Example of Front-End Design Parameters for GLONASS L1 Band

Parameter	*Symbol*	*Units*	*Value*
Passband center frequency	f_G	MHz	1,601.71875
Passband bandwidth	B_{fe}	MHz	>7.8750
First local oscillator frequency	LO_1	MHz	146.1719
Intermediate frequency	f_{IF}	MHz	140.000
SAW DR bandwidth for 4-bit ADC $(DR)_{dB} = -30$ dB	Δf_c	MHz	>15.750
Sampling rate	f_S	MHz	32.941
Nyquist zone	NZ	(Index = odd)	NZ(9)

not be sufficient filtering for the subsequent signal processing to reject unwanted GLONASS bands. If not then digital bandpass filtering prior to the wipe-off processes will be required at the ADC sample rates; for example, a lowpass followed by a bandpass finite impulse response (FIR) filter must perform all computations in less than one ADC sample interval, with time left over for the remaining processes that must also complete their computations in that interval [25, 26]. This adds a significant burden on the DSP throughput capacity of the receiver channel that is not required when the front end has provided ample signal attenuation of the unwanted carrier frequencies by analog filtering. However, digital filtering is vastly superior to its analog counterpart [25].

In summary, the two main advantages of this architectural scheme are one front end for the entire GLONASS L1 band and commonality of receiver channels to track either CDMA or FDMA signals. Each receiver channel could be designed to be reconfigurable under receiver control to operate either with GLONASS FDMA or with any of the multiple CDMA signals. Basically, the GLONASS L1 FDMA signal acquisition and tracking scheme becomes functionally identical to the CDMA scheme. The design similarity will become more apparent in Section 8.4 when the digital receiver baseband processes are described in more detail. The primary difference from the classical GLONASS receiver design is this analog front end does not provide any frequency rejection of unwanted GLONASS signals. This may place an excessive throughput burden on the receiver channels if additional digital bandpass filtering is required prior to the carrier and code wipe-off processes, but adding DSP sophistication requiring more real time throughput capability is a well-established precedent in GNSS digital channel design. This is because DSP technology continues to rapidly increase in both sophistication and speed.

8.4 Digital Channels

The digital channels acquire and track SV signals received from an assigned front end. At this point, the digitized signals are ready to be processed by each of the N digital channels shown in Figure 8.1. No signal detection has taken place in the front end, only signal gain and conditioning plus digital conversion.

The foregoing functional description of a typical GNSS digital channel is presented functionally in top-down order of the real time digital signal processing flow: first the extremely high-speed functions that operate at the same rate as the ADC sampling clock, and then the extremely low-speed functions that operate after massive integration by the high-speed functions. The digital channel architecture is therefore partitioned into the following two categories: fast functions and slow functions. The description of these functions is simplified by describing one digital channel tracking one SV signal, so the functions are depicted in closed loop. Most of the digital channel functions are active during closed loop operation. Since there are two types of ADC inputs, complex baseband and real IF, two versions of the fast functions block diagram are illustrated. The slow functions block diagram is identical for both input types. After the closed loop functions have been described, the time-domain search functions are presented, revealing that some fast functions have been expanded, some slow functions have become idle and some new slow search functions have been added during the digital channel search process. The

real-time search engine architecture is then introduced and described. It requires not only architectural changes necessary for one digital channel to search (as in re-acquisition), but also a massive number of fast and slow search functions using all available digital channels focused on the acquisition of one SV at a time. After these functions have been described at a high level, the detailed processes associated with them are presented in the following order: acquisition in Section 8.5 (that will also describe the frequency domain version of a search engine); carrier tracking in Section 8.6, code tracking in Section 8.7 and loop filters in Section 8.8.

8.4.1 Fast Functions

Figure 8.13 is a block diagram of all fast functions of one digital channel using complex baseband signals, I_n and Q_n, as inputs, where n is the sample number. The fast functions depicted in this figure are shown in closed loop operation, tracking a BPSK modulated signal such as the GPS L1 C/A code signal. All functions in this figure must operate at the baseband ADC sampling clock rate, f_S. Recall that baseband (complex) ADC design guidelines require f_S to be faster than the spreading code rate.

Figure 8.14 is a similar block diagram of all fast functions of one digital channel, but this one uses a real digital signal, IF_n, where n is the sample number. All functions in this figure must operate at the real ADC sampling clock rate, f_S. Recall that real ADC design guidelines require f_S to be faster than twice the spreading code rate. Thus, the sampling rate is always faster for the real IF signal than for the complex baseband signal.

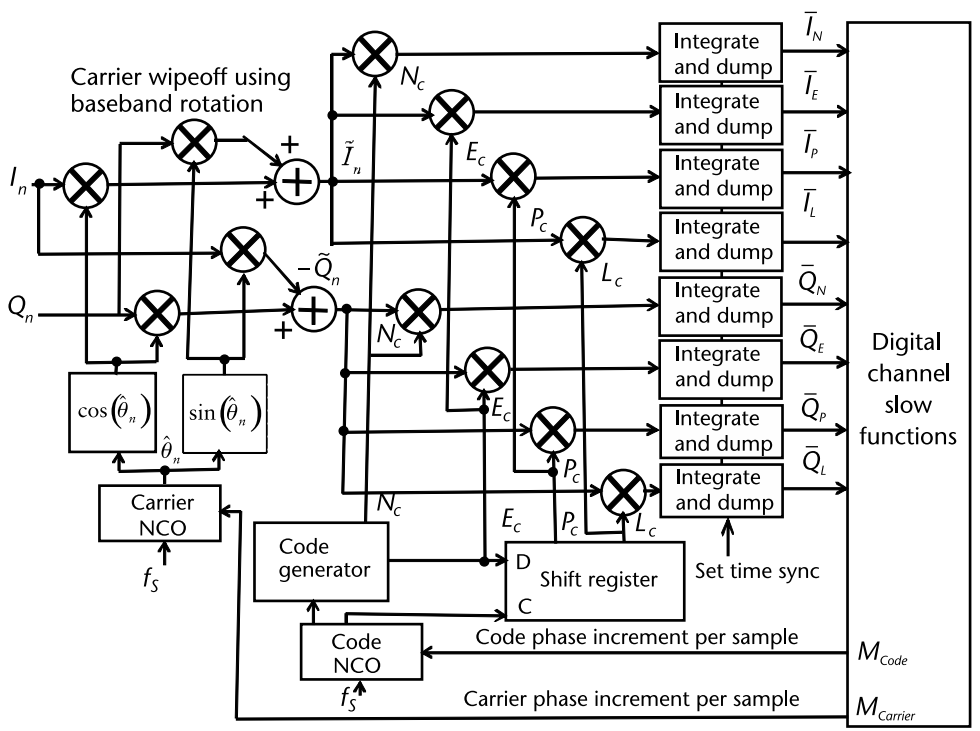


Figure 8.13 Digital channel closed loop fast functions using baseband input.

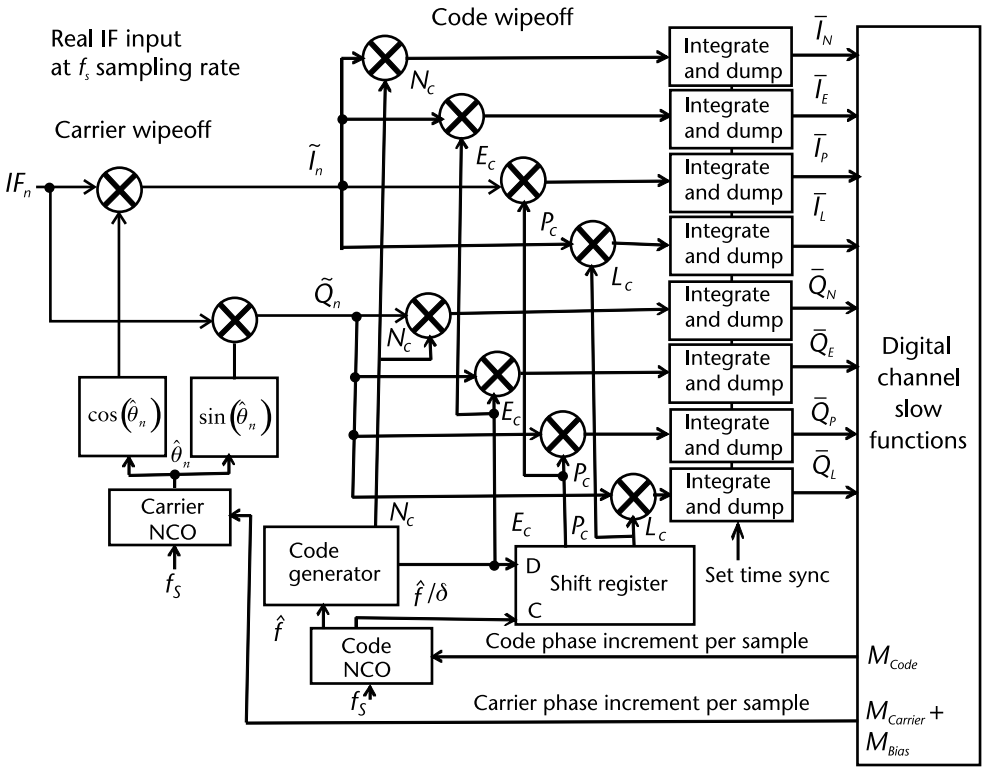


Figure 8.14 Digital channel closed loop fast functions using real IF input.

Note that the only difference between Figures 8.13 and 8.14 is the carrier wipeoff process. The complex baseband signal rotation in Figure 8.13 requires four multiplications and two additions while the real signal at the IF carrier frequency requires only two multiplications. All remaining functions are identical, so the net processing throughput burden for the higher sampling rate of the real IF signal implementation is greater than for the baseband case because the sampling clock rate is roughly two times higher. However, there is less analog processing for the real IF case and the digital processing that replaces the front-end analog process achieves superior results. Also, faster lower power digital technology makes it practical to reduce the analog processing.

8.4.1.1 Carrier Wipe-Off

Assume that the SV signal of interest is a single BPSK modulated signal like the GPS L1 C/A signal that has been downconverted by the front end to IF. The real signal equation is

$$r(t) = \sqrt{2P}d(t-\tau)c(t-\tau)\cos(2\pi f_{IF}t + \theta) + n(t) \tag{8.19}$$

where P is the received signal power (later defined as the carrier power, C), $d(t - \tau)$ is the delayed data modulation, $c(t - \tau)$ is the delayed spreading code modulation, the cosine function is the delayed carrier frequency downconverted to IF and n(t) is

the noise component. After the IF signal is downconverted to baseband, it becomes the complex signal

$$\begin{aligned} I(t) &= \sqrt{P}d(t-\tau)c(t-\tau)\cos(\theta) + x(t)/\sqrt{2} \\ Q(t) &= \sqrt{P}d(t-\tau)c(t-\tau)\sin(\theta) + y(t)/\sqrt{2} \end{aligned} \tag{8.20}$$

Note that the carrier signal has been removed in the complex signal, but θ can be time-varying such that the cos(θ) and sin(θ) terms represent the residual frequencies that remain for each SV tracked due to Doppler, reference oscillator frequency offset and other sources of frequency and phase error in the downconversion process. After the two baseband ADCs have converted these analog signals into complex digital signals, I_n and Q_n, they are rotated as shown in the carrier wipe-off stage of Figure 8.13. The rotation output equation is [27]

$$\begin{bmatrix} \tilde{I}_n \\ \tilde{Q}_n \end{bmatrix} = \begin{bmatrix} \cos\left(\hat{\theta}_n\right) & \sin\left(\hat{\theta}_n\right) \\ -\sin\left(\hat{\theta}_n\right) & \cos\left(\hat{\theta}_n\right) \end{bmatrix} \begin{bmatrix} I_n \\ Q_n \end{bmatrix} \tag{8.21}$$

where $\hat{\theta}_n$ is the nth sample phase estimate from the carrier numerically controlled oscillator (NCO) that closely matches the desired incoming signal phase in the closed loop operation depicted in Figure 8.13. Although the complex signals are buried in noise at this point, the signal processing gain that follows makes it possible for the slow functions (described later) to detect and track these signals with a positive signal-to-noise power ratio (in decibels) that enables the slow functions tracking process, in combination with the carrier NCO, to provide an accurate feedback estimate of $\hat{\theta}_{n+1}$ for the next underlying signal sample.

After $\hat{\theta}_n$ as been converted into cosine and sine signals (described later) the carrier wipe-off process takes place as shown in Figure 8.13. The equation for the complex outputs of the rotated signals is [27]

$$\begin{bmatrix} \tilde{I}_n \\ \tilde{Q}_n \end{bmatrix} = \begin{bmatrix} \sqrt{P}d_n c_n \cos\left(\theta_n - \hat{\theta}_n\right) + \tilde{x}_n/\sqrt{2} \\ \sqrt{P}d_n c_n \sin\left(\theta_n - \hat{\theta}_n\right) + \tilde{y}_n/\sqrt{2} \end{bmatrix} \tag{8.22}$$

where the last terms in both components account for the noise. Note that when the feedback estimate is equal to the incoming phase, the in-phase component becomes maximum because the cosine term is one and the quadra-phase term becomes minimum (noise only) because the sine term is zero. If this is the consistent case, it corresponds to the carrier-tracking loop being in phase lock with the incoming signal. Also note that data (unless a pilot channel is being tracked) and code samples remain in the signal at this point, that the code term is known and is removed by the following code wipe-off process, but the data term (if present) is usually unknown, so its bandwidth becomes the limiting factor in how much integration can take place in the stages following code wipe-off.

In Figure 8.14 the real digital signal, IF_n, is a digitized version of (8.19), so $IF_n = r_n$. In this case the carrier NCO input contains a constant number called the

carrier bias that ultimately removes the IF component in the incoming real signal while also performing the baseband rotation. As a result, the equation for the complex signals of the $\tilde{I}_n, \tilde{Q}_n$ outputs of the carrier wipe-off process contains the same signals shown in (8.22) plus some high frequency components that are removed by the lowpass filtering effect of the following integrate-and-dump functions. The net effect is that essentially identical signals appear at the integrate-and-dump functions outputs in both Figures 8.13 and 8.14.

Carrier Complex Signal Synthesis

As observed in both Figures 8.13 and 8.14, the carrier NCO produces a digital carrier phase estimate $\hat{\theta}_n$ that is used for carrier complex signal synthesis. This phase is converted into complex in-phase and quadra-phase components, $\cos(\hat{\theta}_n)$ and $\sin(\hat{\theta}_n)$, respectively, that have the same bit resolution as provided by the ADC input. Since the number of ADC bits and the number of discrete phases are typically small and because this is a fast function, the cosine and sine functions are typically performed by table look-up (mapping) with negligible loss in precision.

Carrier NCO

Figure 8.15 is a functional block diagram of the NCO design used by both carrier and code wipe-off functions. The design parameters and relevant equations are shown in the figure itself. The digital input and clocking functions are architecturally identical for both applications, but the outputs are different. Replica carrier generation uses the most significant bits of the NCO that are the same number of bits as the ADC bits, N_{ADC}, so the N_{ADC}-bit phase of the N-bit phase accumulator in Figure 8.15 produces $\hat{\theta}_n$ at the same resolution as the ADC.

Two NCO case examples are presented, the first from Section 8.3.7 for a pair of 3-bit baseband ADCs with sampling clock f_{SB} = 34 MHz and the second from Section 8.3.8 for one real IF ADC with f_{SIF} = 112 MHz and f_{IF} =140 MHz that originated in NZ(3), but was downsampled from NZ(3) to NZ(1) where the center frequency became $0.25f_{SIF}$ = 28 MHz. Both cases were designed for the GPS L5

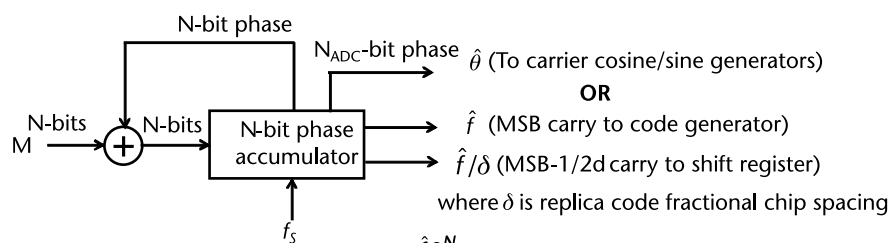


Figure 8.15 Numerically controlled oscillator (NCO) implementation for carrier or code generators and code shift register clock.

signal and $N_{ADC} = 3$ bits. The NCO phase accumulator in Figure 8.15 is assumed to be $N = 32$ bits and the N_{ADC}-bit phase is taken from the three most significant bits (MSBs) of the phase accumulator to synthesize the cosine and sine functions for both case examples. Inspection of Figures 8.13 or 8.14 and 8.15 shows that $M =$ phase increment per sample as the input of the NCO in general, but with an appropriate subscript. Letting $\hat{f}_{BDca}$ designate this specific baseband Doppler frequency, M_{BDca} designate the carrier Doppler NCO input, and f_{SB} designate the 34-MHz baseband sampling clock, then using the equation for phase accumulator output frequency shown in Figure 8.15:

$$\hat{f}_{BDca} = \frac{M_{BDca} f_{SB}}{2^N} = \frac{M_{BDca} 3.4E7}{2^{32}} = \frac{M_{BDca}}{126.3225675}$$

In general, the value of M is determined by the required $\hat{f}$, so for the baseband case, both M_{BDca} and $\hat{f}_{BDca}$ can be positive, negative, or zero (DC) depending on the SV signal Doppler (e.g., if the user is stationary, positive if the SV is rising, zero at highest elevation, and negative if setting).

For the IF ADC case, the IF_n input in Figure 8.14 contains a carrier IF frequency of 28 MHz requiring M in Figure 8.15 to contain a bias component to demodulate this to baseband. Designating this IF carrier bias as $M_{IFbiasca}$, $\hat{f}_{IFbiasca}$ as the 28-MHz IF bias frequency and f_{SIF} as the 112-MHz IF sampling clock, then

$$M_{IFbiasca} = \frac{2^N \hat{f}_{IFbiasca}}{f_{SIF}} = \frac{2^{32} 28}{112} = 1073741824$$

Designating M_{IFDca} as carrier IF Doppler NCO input and $\hat{f}_{IFDca}$ as its Doppler frequency, then

$$\hat{f}_{IFDca} = \frac{M_{IFDca} f_{SIF}}{2^N} = \frac{M_{IFDca} 1.12E8}{2^{32}} = \frac{M_{IFDca}}{38.34792229}$$

The composite IF frequency output is

$$\hat{f}_{IFca} = \hat{f}_{IFDca} + 2.8E7 + \frac{M_{IFDca} + M_{IFbiasca}}{38.34792229} = \frac{M_{IFDca} + 1073741824}{38.34792229}$$

Because of the 28-MHz bias, the composite output frequency of this carrier NCO, $\hat{f}_{IFca}$, never goes negative.

Both of these case examples demonstrate how M is computed for a desired carrier frequency synthesis by the carrier phase accumulator, but recall that the carrier phase accumulator output frequency is not directly used by the carrier wipe-off functions. As shown in Figure 8.15, the most significant bits (MSBs) of the carrier phase accumulator that match the number of bits of the ADC, N_{ADC}, are directly used to synthesize the replica carrier phase, described next.

It is relatively simple to create a table look-up for the cosine and sine functions for the 3-bit ADC as shown in Table 8.10. Note that only one quadrant of values is required in the table (plus application of the correct quadrant) to synthesize both replicas. In the 3-bit ADC case there are three values: 0V, 1V, and 0.7V. In Table 8.10, the columns for the 3-bit cosine and sine values contain three entries. The map entries are the actual 3-bit binary cosine or sine values mapped from the binary values of the NCO N_{ADC} bits. The decimal entries in the middle are the normalized decimal values of the cosine and sine values and the volt entries are the output signal amplitudes of the cosine and sine functions.

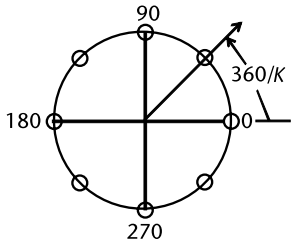
Referring to the top phasor diagram in Figure 8.16, the 3-bit N_{ADC} can synthesize one of $K = 2^{N_{ADC}} = 8$ discrete replica carrier phase states in each 360° cycle, that is, $\hat{\theta}_n = 0°, 45, 90°$, and so forth, so the carrier phase replica resolution is 45°, a very coarse replica phase resolution. The massive amount of integration (averaging) that takes place following the carrier wipe-off process plus the extremely high resolution of the 32-bit NCO carrier phase accumulator makes the effective (average) phase resolution orders of magnitude higher. As noted in Figure 8.16, the phasor rotation is counterclockwise if the output frequency is positive and vice versa. The generalized NCO parameters included in Figure 8.16 apply for any value of N_{ADC}. The bottom I and Q phasor diagram in Figure 8.16 depicts the map of (cosine, sine) output amplitudes (volts) shown in Table 8.10 for all 8 phase points of the 3-bit N_{ADC}.

Figure 8.17 (top) illustrates 17 discrete phase states of the N_{ADC}-bit phase accumulator shown in Figure 8.15 when the N_{ADC} input is $\frac{M}{2^{N-N_{ADC}}} = 1$. This input value corresponds to one 45° phase step per sample since it represents the LSB of the N_{ADC} portion of the carrier phase accumulator. For example, this input value would be 1 if $f_{SIF} = 112$ MHz and the carrier phase accumulator frequency, $\hat{f}_{IFca} = 14$ MHz. Coincidentally, this value is 2 when $\hat{f}_{IFca} = \hat{f}_{IFcabias} = 28$ MHz, the bias frequency for the IF input case example. This steady state bias value would synthesize 90° phase steps.

The cosine and sine samples in the middle and bottom, respectively, of Figure 8.17 are the outputs of replica carrier wipe-off functions, $\cos(\hat{\theta})$ and $\sin(\hat{\theta})$, respectively, shown in Figures 8.13 and 8.14. The 17 epochs produce one epoch ($1/f_S$) more than 2 cycles of f_S with 45° phase increments per epoch, where the first epoch

Table 8.10 Table Look-Up Design for 3-Bit ADC Cosine and Sine Functions

$\hat{\theta}_n$ degrees	$\hat{\theta}_n$ Radians	$cos(\hat{\theta}_n)$ volts	$sin(\hat{\theta}_n)$ volts	3-bit $\hat{\theta}_n$ N_{ADC} bits	3-bit $cos(\hat{\theta}_n)$ map	decimal	volts	3-bit $sin(\hat{\theta}_n)$ map	decimal	volts
315	5.4978	0.7071	−0.7071	111	110	+2/3	+0.7	001	−2/3	−0.7
270	4.7124	0	−1	110	100	+0	0	000	−3/3	−1
225	3.9270	−0.7071	−0.7071	101	001	−2/3	−0.7	001	−2/3	−0.7
180	3.1416	−1	0	100	000	−3/3	−1	100	+0	0
135	2.3562	−0.7071	0.7071	011	001	−2/3	−0.7	110	+2/3	+0.7
90	1.5708	0	1	010	100	+0	0	111	+3/3	+1
45	0.7854	0.7071	0.7071	001	110	+2/3	+0.7	110	+2/3	+0.7
0	0	1	0	000	111	+3/3	+1	100	+0	0



N_{ADC} = Number of ADC bits
= Number of NCO phase accumulator MSBs used

$K = 2^{N_{ADC}}$
= Number of ADC phase points

$\hat{f}_{car}$ = Carrier frequency
= Rate at which the phase points are traversed
= Positive when traverse is counterclockwise
= Negative when traverse is clockwise

$E_{MAX} = \frac{1}{2}\cos\left(\pi\left[\frac{K-4}{2K}\right]\right)$ (Maximum cosine error is at cosine zero crossing steps)

$\tilde{E}_{MAX} = \frac{\pi}{K}$ (Half the phase resolution in radians and numerically slighlty larger than E_{MAX})

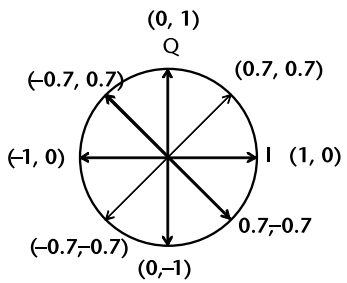


Figure 8.16 Carrier NCO (top) and cosine, sine (bottom) phasor diagrams for N_{ADC} = 3 bits plus generalized parameters.

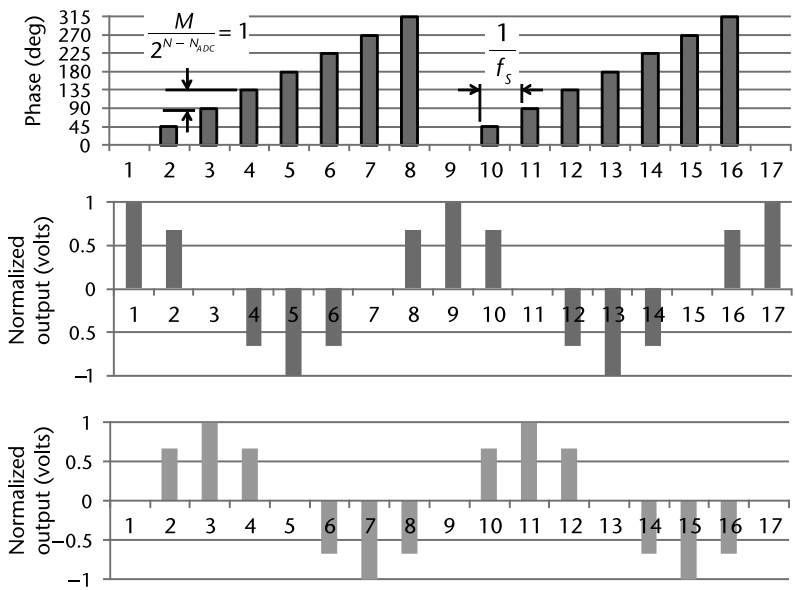


Figure 8.17 Sampled outputs of NCO (top), cosine function (middle) and sine function (bottom) for 3-bit N_{ADC} 45° phase increments.

corresponds $\hat{\theta}_n = 0°$. In general, the same phase appears multiple times for each epoch of f_S because the N_{ADC} resolution is coarse. From Figure 8.16, the upper bound amplitude error, E_{MAX}, occurs at the cosine zero crossing steps and is half the step value. For the 3-bit case example, this is 0.3333 based on the actual quantized step change at this point. It is 0.3535 based on the generalized cosine equation and 0.3927 based on the generalized radian approximation shown in Figure 8.16. Even though the 3-bit N_{ADC} makes coarse amplitude estimates as well as coarse phase

estimates, the massive averaging that takes place by the integration process that follows the carrier wipe-off process, averages these errors resulting in extremely precise carrier phase measurements.

GLONASS Carrier NCO

The carrier NCO is the function in the digital channel design where the GLONASS IF FDMA signal described in Section 8.3.11 is tuned using the architecture of Figure 8.14. As stated in Section 8.3.11, this digital channel FDMA tuning technique for each GLONASS frequency band selected lacks additional bandpass filtering beyond what is performed on the SV. Referring to Table 8.9 for this case example, the entire GLONASS L1 band has a bandwidth of 7.875 MHz, so the front-end bandwidth is wider (say, 10 MHz) and the GLONASS mid-band frequency of 1,601.71875 MHz (offset by 0.2813 MHz with respect to the carrier frequency of the SV on either side) is downconverted to the front-end 140-MHz IF. This IF is downsampled from NZ(9) into NZ(1) with the consequent sampling rate of $f_{SG} = 32.941$ MHz. Recall that the center frequency of NZ(1) is 0.25 f_S in general, so IF_n in Figure 8.14 for the GLONASS case example has the center frequency $f_{IFG} = 8.23525$ MHz. Within the bandwidth of IF_n there are 14 possible GLONASS SV center frequencies of $8.51655 + 0.5625N_G$ MHz, where $N_G = -7, -6, \ldots, 0, +1, +2, \ldots, +6$. This is because the GLONASS L1 FDMA band has been downconverted to 140 MHz, so $N_G(0) =$ 1,602 MHz at L-band has been downconverted to $f_G(0) = f_{IFG} + 0.2813$ MHz = 8.51655 MHz in the 140-MHz IF band. The carrier NCO bias equation to select the desired GLONASS center frequency using N_G is therefore

$$M_{IFGbiasca}(N_G) = \frac{2^N f_{IFGbiasca}(N_G)}{f_{SG}} = \frac{2^{32}(8.51655 + 0.5625N_G)}{32.941} = 73340794.27N_G$$

The carrier NCO bias for $N_G(0)$ is obviously 1,110,418,740. The 14 biases could be precomputed as a table look-up.

The primary concerns about this design are that the three lowest replica carrier frequencies will generate second harmonics that fall into three of the higher frequency bands within IF_n that may correlate enough with energy in those bands to cause interference. These combinations are as follows: $N_G(-7)$ interfered by $N_G(1)$, $N_G(-6)$ by $N_G(3)$ and $N_G(-5)$ by $N_G(6)$, but only if those SV combinations are in view simultaneously. The remedy would require a lowpass digital filter at the outputs of the replica cosine and sine functions to suppress the second harmonic when these lower frequencies are synthesized.

8.4.1.2 Code Wipe-Off

The code wipe-off combined with the massive integration that follows provides the signal processing gain that converts the spread spectrum signal buried in noise to a signal well above the noise level. This is because these fast rate processes collapse the spreading code down to very slow rate outputs (i.e., the wide transmission bandwidth is converted to a narrow signal detection and tracking bandwidth). Referring to Figures 8.13 and 8.14, observe that the inputs to the code wipe-off

correlators (multipliers) are the complex outputs of the carrier wipe-off functions and the replica code phases of the code generator shift register. The code wipe-off functions could theoretically be implemented before the carrier wipe-off function, but practically this increases the number of complex carrier wipe-off functions along with forced increase in the resolution of the code correlation functions. It has already been shown that the complex replica carrier signals should be quantized to the same number of bits as provided by the ADC, while the code wipe-off process can usually be performed with 1-bit precision, that is, using simple "exclusive-or" 1-bit multipliers with no carries (but there are exceptions with some of the modernized GNSS signals). There are diminishing returns on using higher replica code resolution to improve implementation loss. Also, increased code correlation resolution significantly complicates the code correlator design for very little performance payoff. The potential exceptions to this statement are described in Section 8.7. In any case, the wipe-off sequence depicted in Figures 8.13 and 8.14 is the preferred design.

Since the digital channel is assumed to be tracking the SV, then the replica carrier frequency, $\hat{\theta}_n$, is exactly the same as the downconverted incoming SV signal frequency and the prompt replica code phase (P_c) is well within one chip of the incoming SV signal code, so the same correlation properties occur that happen for the mathematical autocorrelation process for a given PRN code signal. However, the mechanics of the digital channel correlation process are very different from the autocorrelation process because only selected phases of the correlation envelope, early (E_c), prompt (P_c), and late (L_c) are replicated and multiplied with the complex carrier-stripped incoming signal samples, $\tilde{I}_n$ and $\tilde{Q}_n$, by the six digital correlators shown in Figures 8.13 and 8.14. Note that these replica code phases are provided by the code generator in combination with the code NCO and the code shift register. Since these fast functions are being described in the context of signal tracking, the replica code generator has determined the prompt replica code phase well within the 2-chip correlation region and continues to match the incoming signal code phase. The prompt code wipe-off function multiplies the prompt replica code, P_n, with the result of (8.22) to produce

$$\begin{bmatrix} \tilde{I}_n P_c \\ \tilde{Q}_n P_c \end{bmatrix} = \begin{bmatrix} \sqrt{P} d_n c_n \hat{c}_c \cos\left(\theta_n - \hat{\theta}_n\right) + \tilde{x}_n \hat{c}_c / \sqrt{2} \\ \sqrt{P} d_n c_n \hat{c}_c \sin\left(\theta_n - \hat{\theta}_n\right) + \tilde{y}_n \hat{c}_c / \sqrt{2} \end{bmatrix} \tag{8.23}$$

where $\hat{c}_c$ is the replica code phase state (typically +1 or –1) of the prompt replica code during the nth sample. This equation provides insight into the two-dimensional carrier and code wipe-off functions because the computation results of both can be seen simultaneously. First, the noise term in (8.22) becomes another noise term on average in (8.23) because the replica code sequence is pseudorandom and there is no correlation with noise. Second, the maximum value of $\tilde{I}_n P_c$ occurs when both the carrier and code replicas match because $\cos(0) = 1$ and $c_n \hat{c}_c$ both have matching signs on a continuous basis while $\tilde{Q}_n P_c$ is essentially noise because $\sin(0) = 0$. Finally, the output signal amplitude deteriorates proportionally with the two-dimensional mismatch of either or both replicas (code and carrier) until the mismatch point (tracking threshold) is reached (by either one or both), at which

point the signal is completely lost. Also note that the only variable remaining in the signal after code wipe-off is the data modulation, d_n. The known time duration between +1, –1 transitions in this usually unknown signal pattern limits how long the following coherent integration process may continue before dumping. Before the maximum coherent integration time can be used, the phase of the data transition boundary must be determined by a process called bit synchronization (or overlay code synchronization in the case of most modernized signals) described in Section 8.11, and the phase of the integrate-and-dump process must be adjusted accordingly. While this transition boundary is unknown, the integration and dump periods must be much smaller so that the corruption that results if a data bit transition occurs during the integration period is infrequent (e.g., 1 in 20 if the integration period is 1 ms and the data transition period is 20 ms). This data modulation is not present in a modernized GNSS signal pilot channel and the significant benefits of this feature are further described later. The prompt complex signal is used for carrier tracking, further described in Section 8.6. The complex early and late signal magnitudes are equal to but lower than and symmetrical around the prompt signal magnitude if the code phase is being tracked perfectly. The early minus late error is used for code tracking and further described in Section 8.7. The code generator noise meter signal, N_c, is described in the next section.

Code Generator

The specific design of every code generator is documented in an interface specification provided by the relevant space segment authority. Some of these designs are presented in Chapters 3 to 6, but the controlling document should be the final basis for the code generator design (as well as all other details of the space segment interface). The code generator in any BPSK design example requires only a 2-bit shift register to provide three replica code phases, namely, E_c (early), P_c (prompt), and L_c (late) that are typically spaced ½-chip apart. There are numerous variations of this basic design such as narrower early/late correlator spacing to improve multipath mitigation, additional very early and very late correlators to detect correlation ambiguity for BOC signals and extended correlators for deep integration with inertial measurement units to improve robustness in the presence of jamming, but this design provides the basis for understanding the variations.

Code Noise Meter

The code noise meter designated as N_c in Figures 8.13 and 8.14 is a 2-chip early code phase with respect to the replica code phases being synthesized by the code generator [28] or a noncorrelating spreading code if that can be formed from available components of the replica spreading code generator. The important uses of the 2-chip early noise meter design are further described in Sections 8.4.2 and 8.4.3, but the fact that it is early by 2 chips with respect to any other replica code output assures that only noise will be generated by its complex correlators if any of the other (later) replica codes are correlating with the incoming signal. The ideal noise meter is processed by the same functions as the signals of interest so as to provide a measure of the noise subject to the same processes (including the same distortions due to approximations) as the signals, but the complexity can be reduced to only the Q signal correlation path. During signal acquisition (further described in Section

8.4), the noise meter is used to set the detection threshold. During the transition into tracking and thereafter, it is used by the slow functions to measure the signal-to-noise ratio. The slow functions are further described in Section 8.4.3.

Code Setter

The code setter is a functional part of every code generator but not depicted in Figures 8.13 and 8.14 because it is a slow rate function interface to a fast rate function and its design must be tailored to each replica code generator design. Specifically, the code setter provides the interface between a slow rate function called the code accumulator and the replica code generator. The code accumulator controls and keeps track of the SV transmit time that is associated with the state of the replica code generator [29, 30]. However, the typical replica code generator transmit time range (period) is ambiguous with respect to the unambiguous time of week maintained on the SV. So the code accumulator ultimately acquires the same unambiguous SV transmit time of week, but partitioned in a manner such that the least significant bits match the replica code generator period. During the tracking state, the replica code generator NCO advances the replica code generator by 1 chip every time it overflows, so the state of this NCO at any time represents an extremely precise quantization measure of the fraction of one chip of the SV transmit time if the code tracking error is zero. The quantization error is virtually zero, so the SV transmit time error using this scheme is strictly due to signal noise. Many GNSS receiver designs do not use code NCOs, so quantization error becomes a factor in the pseudorange measurement error budget.

The code accumulator controls the state of the replica code generator when necessary using the code setter and it maintains a slow rate clock that can predict the SV transmit time any time the receiver control function requests a code range measurement, typically from all tracking channels at the same set time. With the help of other receiver processes, the code accumulator removes the time ambiguity in the code phase offset of the (typical) replica code generator so that the total SV transmit time content in the code accumulator corresponds unambiguously to the SV transmit time in 1 week. The exception to the previous statement is the GPS P(Y) code with a period of exactly one week so the SV transmit time of week is learned unambiguously by the receiver when the replica P(Y) code generator is tracking. Whatever time period is involved with the code generators of other restricted access GNSS signals with long periods, that time period must be learned and maintained by GNSS receiver and the code setter must be able to convert that time into the code state of its replica code generator. The code accumulator is maintained (updated) at the same slow rate input as the code NCO input update shown in the figures as the code phase increment per sample. The SV transmit time in the code accumulator is the natural measurement of a GNSS spreading code signal that is converted into pseudorange [29, 30]. The pseudorange measurement will be further described in Section 8.9.

The code setter is only one aspect of the code generator that is unique for every GNSS spreading code signal. The design must be derived from an interface specification that only describes how the PRN code is generated on the SV. The receiver version must be compatible with this design but is operated and controlled entirely different in the digital channel than in the SV environment. The digital

channel code generator design must support trial-and-error search patterns until successfully matching the incoming SV signal (see Section 8.4.3) and then track the Doppler, atmospheric effects, and other sources of error on the incoming signal in real time (see Section 8.4.2).

The code setter can be a slow function because all fast functions operate with constant NCO inputs until the slow functions change them. Thus, any future replica code generator state can be predicted from the time the known value M is used to update the code NCO until the next time that value is updated using the fact that f_S and $\hat{f}$ are known and $\hat{f}$ is constant until a new value of M arrives. The key to the code setter design is for it to have the ability to set the code NCO and code generator state to the replica code state that corresponds to the known measurement of SV transmit time and to keep track of the transmit time change as a slow function. For example, if the replica code generator is implemented by one linear feedback shift register and the only replica code state known is the reset state, the code setter becomes a linear counter capable of holding the total number of chips in one period of the replica code generator. Assuming that there is a slow function receiver channel process (sometimes called the code accumulator) that knows and can predict the SV transmit time at the code NCO update intervals, then it can store the known time offset to the beginning of the next replica code period into the counter at a specific code NCO update time, then enable the counter to synchronously count down the offset (at the code NCO rate), and then reset the replica code generator at the end of that delay. The replica code generator is then aligned with the incoming signal transmit time and will stay aligned if the code NCO is operating at the correct Doppler compensated chipping rate. Obtaining transmit time from the replica code generator when it has successfully searched and found the transmit time of the SV by the correlation process should not be necessary if the search process has been properly controlled by the slow function search process that is setting and controlling the replica code phase state. However, it could be obtained by the reverse process of determining when the replica code period ends at a specific code NCO epoch time. Although there appears to be only one replica code chip resolution in this transmit time process, there is also an N-bit fraction of a code chip resolution in the code NCO, so the transmit time measurement accuracy strictly depends on how much noise is in the code tracking loop. Ambiguous carrier tracking loop phase measurements in phase lock operation using interferometric techniques can further refine this.

Code Shift Register

In this BPSK design example only a 2-bit shift register is required to provide the replica code phases used during the tracking mode. As shown in Figures 8.13 and 8.14, the replica code spacing is accomplished synchronously by clocking the shift register with the $\hat{f}/\delta$ frequency rate obtained from the code NCO, where δ is the code spacing in chips. So, for $\delta = \frac{1}{2}$ chip, the code shift register is clocked at $2\hat{f}$. As shown in both figures, four code generator replica signals at different code spacing are fed to four complex (I and Q) code correlators. The eight correlator outputs are fed to eight integrate-and-dump functions whose outputs are fed to the slow channel functions for additional integration. The four in-phase integrated signals (in code phase order) are $\overline{I}_N, \overline{I}_E, \overline{I}_P, \overline{I}_L$. The four quadra-phase signals are $\overline{Q}_N, \overline{Q}_E, \overline{Q}_P, \overline{Q}_L$. The

earliest complex signal is used as a measure of the noise floor. The remaining three are the early, prompt, and late complex signals that are used for (prompt) carrier and (early minus late) code tracking (although there are code tracking techniques described in Section 8.6 that also include the prompt signal).

If synthesized by a stage lower than the MSB in the code NCO accumulator design shown in Figure 8.15 and inferred by Figures 8.13 and 8.14, δ can only be binary fractions (i.e., ½, ¼, and so forth chips) and this limitation is usually acceptable. This BPSK design example uses δ = ½-chip code spacing, so the shift register is clocked at twice the code generator rate resulting in E_c, P_c and L_c shift register outputs being ½-chip apart. Consequently, the slow function early minus late code discriminator that will be described in Section 8.7.1 computes the error in the code correlation with the spacing between the early and late correlators of Δ = 1 chip. Variations in code shift register designs are further discussed in Section 8.7.

Code NCO

The code NCO generates the $\hat{f}$ output that advances the code generator at the same spreading rate (plus code Doppler) of the incoming SV signal. By inspection of Figure 8.15, the value M at the input of the code NCO determines the value of $\hat{f}$ shown as the input to the code generator in Figure 8.13 (baseband input) or Figure 8.14 (real IF input). Using the baseband and the IF ADC case examples, the code NCO designs require that M synthesize the replica L5 spreading code rate (10.23 Mcps) plus code Doppler. Since the spreading code rate is the same for both types of signal input and only the variable code Doppler is tracked, a bias component that synthesizes the spreading code rate is added to the variable code Doppler to provide the composite input to the code NCO. For the baseband case example of L5, let $M_{Bbiasco}$ designate the baseband code bias used to replicate the constant $\hat{f}_{Bbiasco}$ = 10.23 MHz, with f_{SB} = 34 MHz, then

$$M_{Bbiasco} = \frac{2^N \hat{f}_{Bbiasco}}{f_{SB}} = \frac{2^{32}10.23}{34} = 1292279866$$

Designating M_{BDco} as the baseband code Doppler input into the code phase accumulator and $\hat{f}_{BDco}$ as its baseband Doppler frequency output, then

$$\hat{f}_{BDco} = \frac{M_{BDco} f_{SB}}{2^N} = \frac{M_{BDco} 3.4E7}{2^{32}} = \frac{M_{BDco}}{126.3225675}$$

The composite baseband code frequency output from its NCO is

$$\hat{f}_{Bco} = \hat{f}_{BDco} + 1.023E7 = \frac{M_{BDco} + M_{Bbiasco}}{126.3225675} = \frac{M_{BDco} + 1292279866}{126.3225675}$$

where $M_{BDco} = 126.3225675\ \hat{f}_{BDco}$.

Similarly, let $M_{IFbiasco}$ designate the baseband code bias used to replicate the constant $\hat{f}_{IFbiasco}$ = 10.23 MHz, with f_{SIF} = 112 MHz, and then the composite IF code frequency output from its NCO can be derived as

$$\hat{f}_{IFco} = \hat{f}_{IFDco} + 1.023E7 = \frac{M_{IFDco} + M_{IFbiasco}}{38.34792229} = \frac{M_{IFDco} + 392299245}{38.34792229}$$

where $M_{IFDco} = 38.34792229\,\hat{f}_{IFDco}$.

The difference between the values of these two examples is due only to the difference in the two ADC sample rates.

8.4.1.3 Integrate and Dump

The integrate-and-dump functions shown in Figures 8.13 and 8.14 provide low-pass filtering of the code wipe-off functions outputs prior to signal detection. Each complex pair of integrate-and-dump functions is often called the predetection filter because the following functions provide signal detection. The architecture of the integrate-and-dump function is similar to the NCO design shown in Figure 8.15 (i.e., it is an accumulate operation over a specified number of samples). That process is typically normalized by dividing the accumulated value by the number of samples before the result is passed on to the next stage and the accumulator zeroed. The integrate part of the function accumulates the outputs of the code wipe-off process. The value from each code wipe-off process is much less than the accumulator capacity, and that must be more than enough capacity such that overflow never occurs before its content has been normalized and transferred to the next stage at the same time the accumulator is reset to zero. The integrate-and-dump function provides the processing gain that changes the signal-to-noise ratio from negative at its input to positive at its output (in decibels) if the two-dimensional carrier and code wipe-off replica signals are closely matched with their respective parts of the incoming signal. The carrier and code wipe-off functions produce wideband error signals well below the noise level that the integration-and-dump process collapse into narrowband error signals well above the noise level (i.e., the despreading process). The ideal noise meter signal is processed exactly the same as the signal of interest in order to provide a matched measure of the prevailing noise level.

Set Time Sync

The set time sync function shown at the base of every integrate-and-dump function adds or subtracts a time increment (TINC) from one integration interval periodically to adjust the phase of the digital channel slow functions such that their processes align with the data bit (or symbol) transitions of the incoming signal. This alignment must be small with respect to the transition period of a data bit (or symbol). For example, if the data bit period is 20 ms (50-bps rate), the time increment (TINC) would typically be 0.25 ms or smaller. Each TINC operation is controlled by the receiver control and processing function after the receiver has learned where these transitions are with respect to the set time and when they have changed in phase enough to require a TINC adjustment. It is very important that the slow

functions be advised of every TINC so that the correct predetection integration time, T, is used in the computations during the next slow cycle.

This phase mismatch in input signal transition boundaries occurs because the SVs in view are typically at different ranges from the user and their ranges are changing. As a result, their data transitions arrive at different phases from each other and specifically different from the set time of the GNSS receiver. The typical set time increment for a GNSS receiver is 10 ms or 20 ms, that is, the typical symbol transition period (100-sps rate) or data bit transition period (50-bps rate), respectively.

Note that the set time sync function is just as important for modern GNSS signals providing data and pilot channels. Typically, the pilot channel is used exclusively for tracking because there are no data transitions and therefore no squaring loss in the carrier-tracking loop, but both signals are processed in the same digital channel. Both code generators must be implemented, but they can share a common carrier wipe-off function and a common code NCO, both maintained by the pilot replica code tracking loop, thereby slaving the data code generator in proper phase relationship to the pilot code generator. Only the prompt replica code of the data replica code generator is needed to perform code wipe-off of the data modulated signal. This architecture is described in Section 8.4.2.2. Following integrate and dump, the prompt signal is passed to the data demodulation function. There are trade-off issues relating to the increased robustness of using the added power in the data channel for tracking. This will be described in more detail in Section 8.11.

8.4.1.4 Fast Function Design Trends

The digital channel fast functions have been presented in the context of hardware-defined functions because this was their design origin. The reason that fast functions of the digital channel are separated from slow functions (described in Section 8.4.2) is that there is a trend in GNSS receiver design toward software-defined fast functions, that is, software-defined receivers (SDRs) using modern digital signal processors (DSPs). The slow functions have always been software-defined. These design trends are presented in the foregoing sections.

Hardware-Defined Fast Functions

Traditionally, the fast functions of the digital channel for high production GNSS receivers are hardware-defined and implemented in one or more application specific integrated circuits (ASICs). This remains as the preferred embodiment for high volume GNSS receivers, although provisions for software definition is often provided where practical in functions predicted to require redefinition during that production cycle and DSP-based ultrafast search engines are often included (introduced in Section 8.4.3.3 and further described in Section 8.5.5).

Nonreal-Time Software-Defined Fast Functions

Many published papers on intended real-time DSP-based GPS SDRs became nonreal-time postmission processing receivers because the authors underestimated the fast functions throughput requirement, but post-mission processing is a valuable asset to GNSS technology development. The earliest sophisticated GPS SDRs that

included the ability to select and experiment with numerous sophisticated versions of every GPS receiver function were developed as nonreal-time receivers used for prototyping and design trade-off analyses [31]. Most GNSS receiver companies have developed extensive proprietary computer-aided design (CAD) resources that include nonreal-time SDRs to debug their present generation receivers and speed-up the development of next-generation receivers with fewer implementation flaws. This is also the trend in GNSS educational training and research. For example, there is a MATLAB compatible version of a highly sophisticated CAD GNSS SDR toolbox available at no cost for educational and noncommercial research use [32]. Reference [32] also provides an extensive list of manufacturers and additional resources that support GNSS receiver research and development.

Software-Defined Fast Functions Using Programmable Hardware

The first successful high performance real-time GPS SDRs used programmable hardware, called field programmable gate arrays (FPGAs), to implement the fast functions instead of ASICs. FPGAs achieve almost the same high levels of digital hardware integration and low power of ASICs at much lower nonrecurring development cost, but higher production cost. The unique replica code generator for each type of spreading code used in modern GNSS receivers represents the most challenging fast function design change (and they also increase the complexity and throughput design impact in the slow function area as well). Linear feedback shift register code generators and NCOs do not run efficiently when software is defined using current-generation DSPs.

Software-Defined Trends in Spreading Codes

The newest GPS L1C spreading code is one of several GNSS ranging codes that cannot be generated by linear feedback shift register techniques. This spreading code is definitely a trend away from a traditionally hardware-defined spreading code generator and toward a software-defined synthesis. Even though the underlying theory is extremely complex, the replica code generation is straightforward. The detailed specifications required to implement this design are provided in [33], but the following high-level design description verifies the relatively simple logic involved. The same 10,223-chip length Legendre sequence and the same 7-chip expansion sequence (0110100) are used to generate every unique 10,230-chip length Weil-based L1C spreading code. Two uniquely specified parameters are used to define the L1C PRN number: the Weil Index (w = 1 to 5,111) and the Insertion Index (p = 1 to 10,223). One (w, p) pair for the pilot channel and another different pair for the data channel, both pairs unique for each L1C PRN number. The Legendre sequence is provided in [33, Table 6.2-1] or it can be generated as follows:

$L(0) = 0$; and for $t = 1$ to 10,222:

$L(t) = 1$, if there exists an integer x such that t is congruent to x^2 modulo 10,223;

$L(t) = 0$, if there exists no integer x such that t is congruent to x^2 modulo 10,223.

To generate the L1C replica spreading code using its specified parameter values, the Weil sequence is first constructed by the exclusive-or of the Legendre sequence $L(t)$ with itself shifted by the Weil Index and the result stored into the Weil

sequence as: $W_i(t; w) = L(t) \oplus L(t+ w)$ for $t = 0$ to 10,222. For example, $W_i(0; w) = L(0) \oplus L(0+ w)$, $W_i(1; w) = L(1) \oplus L(1 + w)$ … $W_i(10{,}222; w) = L(10{,}222) \oplus L(10{,}222+ w)$. Note that the circular shift of the Legendre sequence leaves the Weil sequence exactly 7 chips short of the 10,230 chips required for the Weil sequence (i.e., the L1C pilot or data channel spreading code). The insert point is specified by Insertion Index p, where $p = 1$ to 10,223. The expansion sequence is inserted before the pth value of the Weil code. For example, the Insertion Index for L1Cp PRN 1 is $p = 412$, so the insertion would be: … W_i (411; w), 0110100, W_i (412; w), W_i (413; w) ... $W_i(10{,}222; w)$.

L1C is a prime example of a spreading code that must be software-defined before operational use. The typical software-defined implementation would be to compute the constant Legendre sequence and store it along with the constant 7-chip sequence 0110100 as part of the code generation firmware, then pregenerate the particular PRN spreading code using its specific w to compute the Weil sequence and then its specified p to properly insert the remaining 7 chips. A 10,230-chip circular shift register (hardware or software-defined) could be used to generate this replica spreading code. A DSP SDR would precompute and store the L1C spreading codes in the same manner described in the next section. Other GNSS spreading codes use memory codes that cannot be generated by a linear feedback shift register or software-defined, but rather must simply be stored in nonvolatile memory.

Software-Defined Fast Functions

Software-defined fast functions using DSPs originally replaced the parallel hardware-defined schemes with table look-up (TLU) schemes, thereby trading increased memory and lower resolution for improved sequential computational efficiency. However, DSPs are not only improving in sophistication of their instruction sets and support software but also achieve speed improvements by means of parallel processing utilizing single instruction multiple data (SIMD) operations commonly utilized in modern microprocessors. This architectural advantage will eventually be utilized in future generations of software-defined GNSS receivers, and there will likely be more similarity between software-defined and hardware-defined fast functions.

Since the first SDR DSPs could not perform the parallel hardware-defined fast functions shown in Figures 8.13 and 8.14 efficiently, the earliest successful, real-time, software-defined fast functions replaced them with maps of precomputed PRN codes onto a set of sample times stored in bit-wise parallel formats. This technique enabled efficient DSP block processing using bit-wise parallel software correlations with the incoming signals [34]. The bit-wise scheme speeds up software correlation by operating in parallel on multiple samples using TLU techniques. The downside of the bit-wise parallel correlation technique includes suboptimal quantization (2-bit sign/magnitude ADC) and coarse replica carrier/replica code alignment during closed-loop tracking with subsequent loss of measurement precision because TLU techniques have much lower measurement resolution than their typical 32-bit NCO counterparts.

An impressive single DSP was used to implement an equivalent 43 channel L1 C/A GPS SDR developed for university scientific research in 2006 [35]. This SDR used efficient DSP techniques including a 2-bit sign/magnitude ADC, fixed-point

processing, bitwise parallel correlators and a fast Fourier transform (FFT)-based acquisition scheme. The FFT scheme performed rapid acquisitions at C/N_0 (carrier to noise power ratio in a 1-Hz noise bandwidth) of 33 dB-Hz or higher. The SDR references cited in [35] provide insight into the historical progress of achieving real-time DSP operation that successfully overcame the computational challenges of the digital channel fast functions. In 2011, [36] described a production L1 C/A and L2C SDR using the same bitwise parallel correlators plus modifiable open-source software targeted for scientific space applications, one version packaged for laboratory experimentation and another compact configuration for space applications. The L2C acquisition scheme used the reduced uncertainty (aiding) following initial signal acquisition by the fast FFT C/A code scheme.

A major benefit of software-defined functions is that each programmed function can be re-entrant (i.e., only one software function or combination of functions is programmed). This program is multiplexed (reentered) multiple times to effectively provide multiple channel uses of the same function but with different operating variables and even different constants. It can be multiplexed as many times as the host processor can permit while still meeting its allocated real-time loading budget with some remaining time margin. A re-entrant program requires a unique indexed memory space (usually provided by the calling program) for the variables or unique constants associated with the program. So the DSP must be fast enough to support as many reentrant fast functions as required for multiple channels. One paradox is that a fast DSP can be the most suitable engine for fast functions, but a fast microprocessor with built-in fast floating point capability and extensive memory capacity is a more suitable engine for the remaining (slow) functions. So the use of one or more fast DSPs and a fast central microprocessor should be considered for a multiconstellation general-purpose SDR.

It should now be apparent that the fast functions perform relatively simple and repetitive operations in steady state that are updated by comparatively very slow (and usually more complex) functions. So instead of performing the operations synchronously (in real time) using parallel hardware-defined functions (as is the usual case for hardware-defined implementations, including FPGAs), the purely software-defined scheme uses a DSP to take in blocks of data in real time intervals. These blocks are processed serially and asynchronously (but much faster than real time) until all of the processes equivalent to their hardware-defined counterparts have computed all required outputs for that real time interval. These outputs are passed on to their respective slow functions (that have always been software-defined). These slow functions process their inputs and provide corrective feedback to the fast functions. As will be seen in Section 8.8.5, computation delay in the slow functions plays a key role in maintaining closed-loop stability.

Design Comparisons

Reference [37] discusses and compares the attributes of ASIC, FPGA and DSP technologies, then predicts their roles in future generation GNSS receivers. The article recognizes that modernized GNSS signals are creating a new transition period that could last as long as 6 to 10 years where GNSS receiver developers will employ technology that can be changed in the field. During this transition period field programmability is advantageous because it allows design flexibility during continuous

evolution of improvements that take place. After this transition period, established chip sets and original equipment manufacturer (OEM) modules/receivers will once again dominate the market. Table 8.11, adapted from [37], ranks the three technologies with respect to five categories normally considered during product trade studies. Based on ranking, it is likely that mass-market commercial applications will continue to use ASIC technology in GNSS receivers because it provides the smallest production single-unit costs. FPGA and DSP technology may be used as development tools and in GNSS receivers for limited-market applications.

8.4.2 Slow Functions

The slow functions are the ones that operate with the GNSS signals after they have been despread into narrowband signals. Figure 8.18 illustrates the slow functions involved with baseband code and carrier tracking in closed-loop operation. It also depicts the carrier power estimate, $I_P^2 + Q_P^2$, formed with the prompt signals and the noise power estimate, $I_N^2 + Q_N^2$, formed with the noise signals that are passed on to the C/N_0 meter described in Section 8.12 along with other special slow baseband functions that use the unmodified I_P, Q_P prompt signals. The data demodulation function described in Section 8.11 uses I_P if the signal contains navigation data such as the GPS C/A code and P(Y) code signals. There would not be data modulation in the carrier tracking loop prompt signals if a modernized signal (such as GPS L5) with separate data and pilot channels is being tracked because the pilot channel signal is used here. In this case, the data demodulation function will receive an open loop prompt signal from the data channel function that will be described in Section 8.4.2.2. The combination of these slow carrier and code tracking functions and the fast carrier and code wipe-off and predetection integration functions form the complete closed loops of one digital channel. These slow functions are typically implemented in the receiver and control processor shown in Figure 8.1.

8.4.2.1 Integrate and Dump

The four I and four Q signals from the integrate-and-dump outputs of the fast functions shown in Figures 8.13 and 8.14 (as separate I and Q groups) have been rearranged in Figure 8.18 (in matching I and Q pairs) as four complex inputs into the slow rate integrate and dump functions. The total combined duration of the fast and slow rate integrate-and-dump functions establishes the predetection integration

Table 8.11 Predicted Technology Preferences for Modernized GNSS Receivers

Technology	*Development Costs*	*Performance*	*Power Consumption*	*Single-Unit Costs*	*Flexibility*
ASIC	Major disadvantage	Major advantage	Major advantage	Major advantage	Major disadvantage
FPGA	Minor disadvantage	Major advantage	Minor advantage	Minor disadvantage	Minor advantage
DSP	Major advantage	Major advantage to minor advantage	Minor advantage to major disadvantage	Minor advantage to minor disadvantage	Major advantage

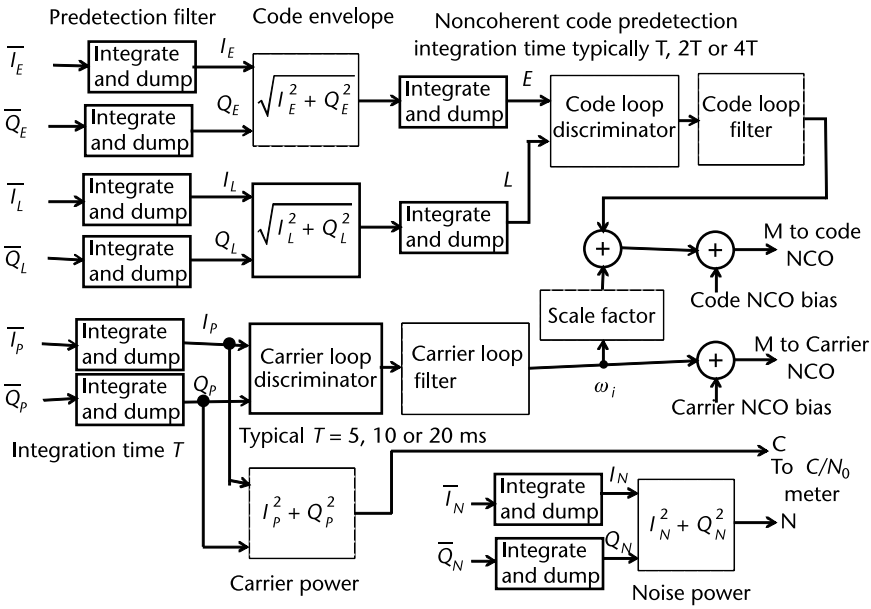


Figure 8.18 Digital channel closed loop slow functions.

time (T) for the digital channel. For example, one sample of the normalized slow function in-phase prompt signal can be expressed in equation form as

$$I_P = \frac{1}{f_S T} \sum_{n=k}^{n=k+f_S T} \tilde{I}_n P_n \tag{8.24}$$

where the fast prompt samples and sampling clock rate terms are taken from Figures 8.13 or 8.14 and the slow prompt sample term from Figure 8.18. The other slow rate input equations are similar. As shown in Figure 8.18, typical values of T for the carrier-tracking loop are 5, 10, and 20 ms (when in closed-loop operation), but there are temporary search slow functions where $T = 1$ ms and in some cases closed-loop operation with $T = 2$ ms such as tracking the L5 GPS/WAAS signal at 500 sps [38]. Assuming $T = 20$ ms, then for the baseband ADC case example, where $f_s = 34$ MHz, the number of integrated fast function samples is 680,000. So the fast to slow function processing speed ratio is 680,000 to 1. Table 8.12 summarizes these very significant processing speed differences for both the baseband and IF case examples for typical values of T.

For the GPS L1 C/A signal, there is 50-bps data modulation present on the signal and the steady state value of T is set to 20 ms to permit the carrier-tracking loop to operate in phase lock with the maximum time between data transitions. The start and stop time (phase) of these integrate and dump functions should not straddle the data bit transition boundaries because each time the SV data bits change sign, the subsequent I and Q values are reversed in sign. Consequently, if the boundary is straddled and there is a data transition, the integration and dump result for that interval is degraded. In the worst case, if the data transition occurs at the mid-point (T/2), the signal is effectively canceled for that interval. During the initial (cold

Table 8.12 Ratio of Fast Functions to Slow Functions Processing Speed

Predetection Integration Time	*Ratio of Fast Function to Slow Function Processing Speed*	
T (ms)	Baseband ADC f_s = 34 MHz	IF ADC f_s = 112 MHz
20	680,000:1	2,240,000:1
10	340,000:1	1,120,000:1
5	170,000:1	560,000:1
1	34,000:1	112,000:1

start) C/A code signal search process, the receiver does not know where the SV data bit transition boundaries are located. The performance degradation has to be tolerated until the bit synchronization process, described further in Section 8.4.3, locates these boundaries. During these times, shorter values of T are used so that the majority of integrate and dump operations are assured to not contain a data transition. For example, a typical value of T in the carrier tracking loop prior to bit sync is 5 ms. If a modernized signal is being tracked, the pilot channel is used and data transitions are of no concern to the tracking loops, but there are other pilot channel concerns while getting into the steady state carrier tracking mode that are introduced in the following section.

8.4.2.2 Carrier Tracking Loop

In Figure 8.18 the I_P, Q_P slow function samples are fed to the carrier loop discriminator that determines the carrier phase error for each sample. Each phase error sample is fed to the carrier loop filter that removes noise and predicts a Doppler frequency correction sample, ω_i. The design details of these two functions are provided in Section 8.6. This output is fed to the carrier NCO after a bias (if any) is added. Recall that no bias is required for the baseband ADC case while the IF ADC case does require a bias term. That output becomes the M input of the carrier NCO illustrated in Figure 8.15 and the carrier phase increment per sample shown in Figures 8.13 and 8.14. Each ω_i sample is also multiplied by a scale factor and added to the code loop filter output to provide carrier aided code. Since the code and carrier outputs are both range terms, the carrier aiding effectively removes the dynamic stress from the code loop. There is a smaller amount of Doppler in the code tracking loop than in the carrier tracking loop, so the carrier aiding to the code loop must be scaled down. The scale factor is determined by

$$\text{Scale factor} = \frac{R_c}{f_L} \ (\text{cycles/chips}) \tag{8.25}$$

where R_c = spreading code rate (Mcps) and f_L = L-band carrier frequency (MHz) of the signal.

Table 8.13 shows the scale factors for the GNSS open signals. This table does not include the future GLONASS open signals L2OF (L2 FDMA signals) and

Table 8.13 GNSS Open Signal Scale Factors for Carrier-Aided Code

Constellation	*Signal*	*Carrier Frequency (MHz)*	*Spreading Code Rate (Mcps)*	*Carrier-Aided Code Scale Factor*
GPS	L1 C/A	1,575.420	1.023	0.000649351
	L1 C_D, L1 C_P	1,575.420	1.023	0.000649351
	L2 CM, L2 CL	1,227.60	0.5115	0.000416667
	L5 I5, Q5	1,176.450	10.23	0.008695652
GLONASS	L1OF	$1{,}602 + 0.5625N_G$	0.5115	0.000320075 $N_G(-7)$
		$N_G = -7, -6, -5, -4, -3, -2 -1, 0,1, 2, 3, 4, 5, 6$		0.000319288 $N_G(0)$ 0.000318617 N_G (6)
	L3OC	1,202.025	10.230	0.008510638
Galileo	E1 B, E1 C	1,575.42	1.023	0.000324675
	E5a-I, E5a-Q	1,176.450	10.23	0.000434783
	E5b-I, E5b-Q	1,207.140	10.23	0.000423729
BeiDou	B1I	1,561.098	2.046	0.000327654
	B2I	1,207.140	2.046	0.000423729

L1OC and L2OC (L1 and L2 CDMA signals). Note that there are 14 carrier frequencies and scale factors for the GLONASS L1 signals, but the same replica code generator is used for all 14. The recent GLONASS L3 signal is a CDMA signal. Also, the future BeiDou open signals, the B1-C and the B2 AltBOC (15, 10) signal consisting of QPSK-R(10) signals B2b and B2a, are not included for lack of a final interface specification.

Pilot Channel Carrier Tracking

Up to this point, the digital channel architecture has been described in the context of legacy GPS and GLONASS signals that are modulated by data. Many modernized GNSS signals provide separate data and pilot (dataless) channels. The pilot signal is the clear choice for carrier tracking because it is significantly more robust than the data signal. This is further described in Section 8.6.1. Although the pilot channel removes concerns about data transitions in the carrier tracking loop, data transitions remain in the data channel demodulation process. There are also secondary codes in the modernized signals that cause degradation in the signal-to-noise ratio until the replica secondary codes of both data and pilot channels are synchronized with their counterparts in the SV signal. The secondary code generation function is included with each code generator design and they do provide mitigation to multipath and narrowband interference when synchronized with their respective codes. Secondary codes are further described in Section 8.5.

There are relatively few architectural changes needed to make a legacy digital channel design compatible with a modernized GNSS signal with data and pilot channels. The architectural addition is shown in Figure 8.19. There are now two code generators, data channel and pilot channel, with their respective shift registers. The data channel is synchronized to its incoming signal by the pilot channel. The prompt code of the data channel is mixed with $\tilde{I}_n$ followed by the data channel fast function integrate and dump to produce $\overline{I}_{Pd}$ that is sent to its slow function

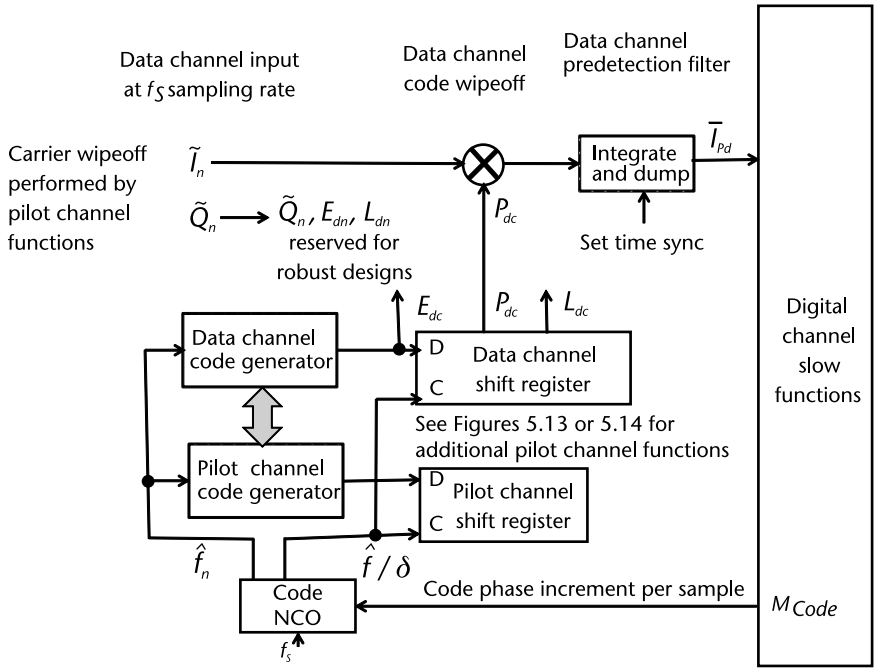


Figure 8.19 Minimum data channel fast functions using pilot channel code generator for synchronization.

counterpart. After the slow function integrate and dump produces I_{Pd} (not shown), this signal is fed to the data demodulation function. The modernized data demodulation function, which may include a forward error correction (FEC) decoder, is more complex than its legacy counterpart and far more effective at reducing the bit error rate. As stated earlier for the legacy design of Figure 8.18, I_P is sent to the data demodulation function only when a data channel is controlling the carrier-tracking loop. Data demodulation is described further in Section 8.11.

Since the data and pilot channels are synchronous at their origin in the SV and modulated onto the same carrier, then digital channel closed loop operation with the pilot channel also provides sufficient information to synchronously demodulate the data channel in the open loop manner depicted in Figure 8.19. The data channel performance would be worse if operated independently in closed loop with its own carrier and code tracking loops. As noted in Figure 8.19, $\tilde{Q}_n$, E_{dn}, L_{dn} are reserved for robust designs (i.e., when there is a need to enhance the pilot tracking robustness using the power in the data channel for other than data demodulation).

Phase Alignment With Data/Symbol Transitions

Figure 8.20 illustrates the phase alignment of integrate and dump intervals with SV data or symbol transition boundaries. At the top of figure, the SV data or symbol transition boundaries gradually change phase (owing to SV range changes) and are usually not aligned with the receiver's set time boundaries shown second from the top of the figure. The SV transition boundaries movement relative to the set time can be in either direction and typically very slow (e.g., for a stationary user, slowly

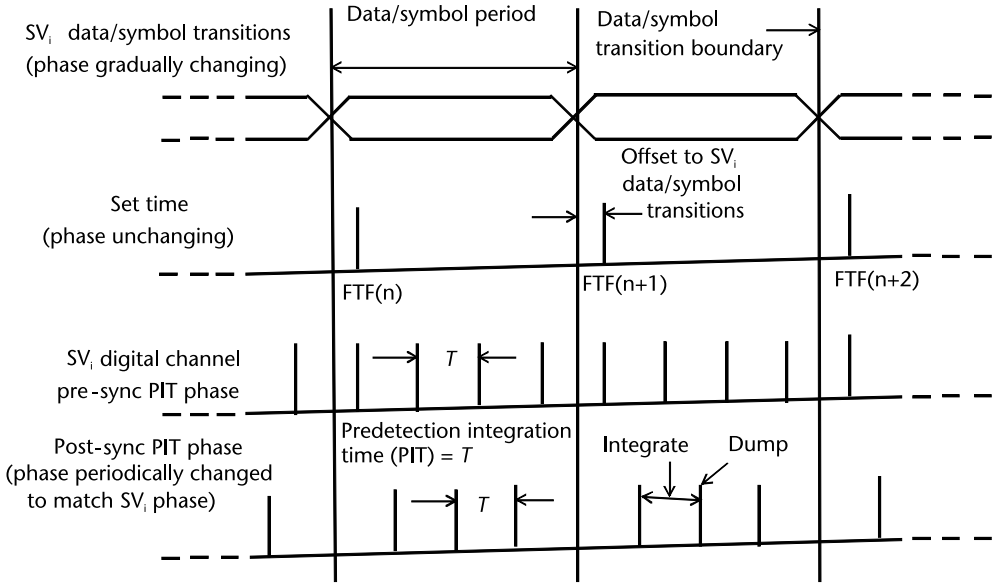


Figure 8.20 Phase alignment of predetection integrate-and-dump intervals with SV data/symbol transition boundaries.

to the right if the SV is rising, stationary at highest elevation and slowly to the left if setting, with about a 20-ms total change in phase between zenith and either horizon). The SV data or symbol transition boundaries are generally different for each SV and therefore different in each digital channel.

As stated in Figure 8.20, the set time phase is unchanging and its period, named fundamental time frame (FTF) in the figure, is the same phase in every digital channel. The FTF count is shown under each epoch in the second line. The FTF count is monotonic and is typically maintained by a 32-bit accumulator that periodically rolls over. It keeps the same time for all of the digital channels and is maintained by the receiver control and processing function. The digital channel design must adjust its predetection integration time (PIT) from the misaligned (presync) phase shown in the third line of Figure 8.20 to the aligned (post-sync) phase shown on the bottom line. When the receiver does not know its position and precise time, it cannot compute the SV range to predict the phase of the SV transitions. However, the receiver control and processing function, in cooperation with each digital channel, learns this initial phase offset by a process called bit synchronization (or secondary code synchronization). Bit sync occurs shortly after the signal is acquired and uses the set time sync feature shown in Figures 8.13 and 8.14 to remove the offset between the FTF and SV transitions that are shown in the second line from the top of Figure 8.20. After this alignment to the SV data or symbol transition, the (suboptimum) integrate-and-dump boundary (i.e., PIT shown on the bottom line) is widened to match to the data or symbol period of the SV. This is the desired (optimum) PIT for data demodulation. It is usually the desired PIT for phase lock loop operation with the legacy signals that have data transitions in the tracking signal. For pilot signals there are no data boundary restrictions on the PIT, but since there

is no squaring loss in these signals the PIT is also less important. The PIT defines the value of T that is used prolifically in most of the sections to follow.

8.4.2.3 Code Tracking Loop

Referring back to Figure 8.18, the slow functions of the code-tracking loop are shown at the top using the early and late fast function signals. The slow function integrate-and-dump process completes the predetection integration time. This is followed by the envelope computation shown in the figure. The early and late envelopes are sent to the code loop discriminator after (optional) noncoherent filtering by integrate-and-dump functions shown in the figure. This additional integration time is made possible by the carrier aiding to the code loop described earlier because dynamic stress has been removed from the code tracking loop (i.e., it is only tracking atmospheric delay variations). If the noncoherent integration option is used, the code loop filter is updated at a slower rate than for the carrier loop filter thereby removing more noise and reducing computation burden. Inspection of the carrier-aided code design in the figure reveals that code loop filter output is constant for one or more periods of carrier loop iterations, but the scaled carrier aiding is added to the code loop output at the carrier loop update rate. The nominal code spreading rate (called code NCO bias) is added to the carrier-aided code Doppler output and fed to the code NCO, resulting in the same code loop update rate as the carrier loop.

Even though these are slow functions in terms of how often they are performed, when the time comes (the dump sample time) for all of their computations to be performed, the computation delay should ideally be completed before the next sample time occurs. This corresponds to zero computation delay in the real-time digital process. However, this is typically not the case. This design issue is further described as part of the loop filter stability topic in Section 8.8.5.

There are numerous design variations in code tracking loop designs. For example, a common design practice is to use the power computation of the complex inputs (avoiding the square root to provide the envelope), but this is a carryover from analog technology that produces a nonlinear code discrimination function even at high signal to noise ratios. On the upside, if the carrier-tracking loop is in phase lock, then coherent code tracking using only I_E and I_L can be used (unmodified) since the Q signals contain only noise under this condition. Coherent code tracking is ideal because it produces the most accurate range measurements, but requires quick fallback to the noncoherent code discriminator process if phase lock in the carrier loop is lost. So the prudent use of coherent code tracking depends on a very reliable phase lock status indicator design that is further described in Section 8.13.2. The various code loop discriminator designs are described in Section 8.7.1.

8.4.3 Search Functions

Now that the digital channel architecture and its basic functions are better understood from a high-level steady-state signal tracking perspective, the search functions can be described from the perspective of getting the digital channel into the tracking state. Three search modes are typically used in a commercial GNSS receiver:

1. Sky search: This uses the search engine under cold-start and warm-start uncertainty conditions where the receiver is highly uncertain about numerous parameters including the uncertainty that the SV being searched is in view. Sky search is a sequential search process (i.e., the search engine searches for one SV at a time, dismissing any SV after the full uncertainty has been explored for that SV without success). The worst-case PVT and SV location (visibility) uncertainty condition is cold-start acquisition where the GNSS receiver typically has no PVT information and little or no knowledge of the reference oscillator frequency offset, but usually has at least crude almanac data for the SV constellations used. The almanac data cannot provide SV location without at least a coarse estimate of position and time, but is available for initial measurement incorporation when the first four SVs have been acquired (without the added delay waiting for SV broadcast ephemeris data to be demodulated). The search condition from standby mode when the previous mode was tracking is called warm start. It has lower uncertainty than for cold start, such as knowledge of the reference oscillator frequency offset from its specified value (that reduces Doppler uncertainty) and a more accurate estimate of time and almanac data, but not necessarily position (needed to locate the SVs in view), but the last known position and current time is a good starting point for warm start. Sky search would be an extremely time-consuming process without a search engine. The GPS L1 C/A code is a very efficient acquisition signal in space and was originally designed to serve this purpose. It has a short code length of 1,023 chips, but its signal is not as robust as are its modernized counterparts.
2. Aided search: This uses reduced search resources of the digital channel when the receiver is tracking in at least four other digital channels, thereby providing very low uncertainty conditions. Aided search is a parallel search process whereby every digital channel is acquiring different SVs at the same time. The aiding provided to each digital channel under this search condition results in almost instant acquisition unless the signal is blocked or the interference level is high enough to prevent acquisition. In the latter case, the situational awareness feature of the front-end provides information to best determine the acquisition strategy. Under this search condition, the aiding includes the certainty that the SV is visible unless obstructed.
3. Reacquisition: This initially uses aided search and the reduced search resources of the digital channel with aiding provided by dead reckoning. Reacquisition occurs when some or all signals are lost due to dynamic stress, signal blockage, or interference and is a parallel search process. The initial uncertainty under this condition is usually low but grows exponentially with outage time if fewer than two SVs are being tracked (assuming that altitude hold and time bias rate hold modes are used during a 2 SV tracking condition). When the predicted uncertainty has increased to the point where the digital channel search resources are inadequate, the search engine is used. When a subset of digital channels lose track for any reason, but four or more remain tracking, that subset is technically reacquiring, but functionally in aided search.

8.4.3.1 Search Engine

The search engine is a massive two-dimensional search function. It can be implemented by traditional time-domain search techniques or by modernized frequency domain techniques. Frequency-domain techniques are computationally intensive and are practical only if available computational resources can support it. The time-domain architecture is described first, followed by the frequency-domain architecture. These acquisition designs are further described in Section 8.5. The search engine is used only in the worst-case condition where the receiver can provide little or no information to assist in the search process (i.e., it must explore the maximum uncertainty in both the carrier Doppler and code range dimensions).

Carrier Doppler Range Uncertainty

The determination of the maximum Doppler uncertainty must take into account contributions due to the changing range to the SV caused by the maximum SV velocity toward the user plus the maximum user velocity toward the SV plus the frequency offset error in the reference oscillator. The SV contribution to Doppler uncertainty is described first.

Figure 8.21 illustrates the orbit geometry of the GNSS SV with respect to a stationary user on the surface of the Earth, neglecting Earth rotation and relativity effects due to the high velocity of the SV. The geometry assumes a constant Earth radius, R_E = 6,378,137m (WGS-84 equatorial radius), constant SV orbit radius, R_{ES}, around the Earth and that the SV orbit is in the plane of the user (i.e., the worst-case Doppler range condition where SV closest approach to the user is directly overhead). The user height above the Earth surface, h_U, is assumed to be zero because it is usually a small order effect, but it could simply be added to the Earth radius, R_E.

By inspection of Figure 8.21, the SV velocity is

$$v_S = \frac{2\pi R_{ES}}{T_{orbit}} \ \text{(m/s)} \tag{8.26}$$

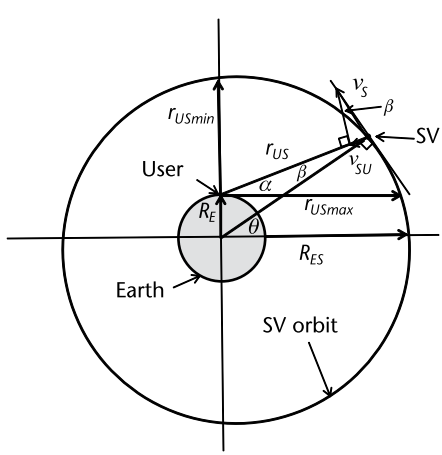


Figure 8.21 GNSS space vehicle (SV) orbit geometry with respect to stationary user on Earth's surface.

where T_{orbit} is the SV orbit period in seconds and R_{ES} is the SV orbit radius from the Earth's center in meters (usually specified as the standard orbit semimajor axis).

The SV velocity component toward the user is

$$v_{SU} = v_S \sin\beta \ \text{(m/s)} \tag{8.27}$$

where β is the angle at the SV between the line from the center of the Earth, that is, the SV orbit radius, R_{ES}, and the line from the user to the SV (i.e., the user range to the SV, r_{US}). Note in Figure 8.21 that the projection from the end of vector v_S onto r_{US} is perpendicular to vector v_{SU} that is aligned with r_{US}, so the angle opposite to r_{US} is also equal to β.

Using the law of sines:

$$\begin{aligned} &\frac{\sin\beta}{R_E} = \frac{\sin(\alpha+90)}{R_{ES}} = \frac{\cos\alpha}{R_{ES}} \\ &\sin\beta = \frac{R_E}{R_{ES}}\cos\alpha \qquad \text{(radians)} \\ &\beta = \sin^{-1}\left(\frac{R_E}{R_{ES}}\cos\alpha\right) \end{aligned} \tag{8.28}$$

The satellite velocity component toward the user can now be defined in terms of the user elevation angle, α, in radians by substituting (8.28) into (8.27)

$$v_{SU} = v_S \sin(\beta) = v_S \sin\left(\sin^{-1}\left[\frac{R_E}{R_{ES}}\cos\alpha\right]\right) = v_S\left(\frac{R_E}{R_{ES}}\cos\alpha\right) \ \text{(m/s)} \tag{8.29}$$

Using (8.29) and then (8.26), the worst-case Doppler effect at the user position caused by the SV carrier frequency is

$$f_D = \frac{f_L v_{SU}}{c} = \frac{v_{SU}}{\lambda_L} = \frac{v_S R_E}{\lambda_L R_{ES}}\cos\alpha = \frac{2\pi R_E}{T_{orbit}\lambda_L}\cos\alpha \ \text{(Hz)} \tag{8.30}$$

where $\lambda_L = \dfrac{c}{f_L}$ is the wavelength (m) of the transmitted L-band frequency, f_L (Hz), using the velocity of light, $c = 299{,}792{,}458$ m/s.

The range from the user to the SV is determined by

$$r_{US} = \sqrt{R_E^2 + R_{ES}^2 - 2R_E R_{ES}\sin\left(\alpha + \sin^{-1}\frac{R_E}{R_{ES}}\cos\alpha\right)} \ \text{(m)} \tag{8.31}$$

Using (8.31) and referring to Figure 8.21, $\alpha = 0$ radians corresponds to the maximum range to the SV, $r_{US\max} = \sqrt{R_E^2 + R_{ES}^2 - 2R_E^2} = \sqrt{R_{ES}^2 - R_E^2}$ (m) , and $\alpha = \frac{\pi}{2}$ radians corresponds to the minimum range, $r_{US\min} = \sqrt{R_E^2 + R_{ES}^2 - 2R_E R_{ES}} = R_{ES} - R_E$(m).

Note that the Doppler is maximum when the SV is at the user horizon (α= 0 and π radians), zero at user zenith, positive when the SV is rising ($0 \leq \alpha < \frac{\pi}{2}$ radians) and negative when the SV is setting $\frac{\pi}{2} < \alpha \leq \pi$ radians.

The equation for the added Doppler due to user motion is

$$f_{DU} = \frac{v_{US}}{\lambda_L} = \frac{v_U \cos\alpha}{\lambda_L} \text{ (Hz)} \tag{8.32}$$

where v_{US} is the user velocity toward the SV and v_U is the user velocity in the user plane in the direction of the SV orbit plane. In practice, v_U is assumed to be the maximum user velocity (m/s).

The reference oscillator frequency offset is specified as a dimensionless ratio, $\frac{\Delta f}{f}$, with a value in parts per million (ppm). Specific value examples are ±0.5 ppm (very high quality), ±1 ppm (high quality), ±2 ppm (medium quality), and ±3 ppm and higher (low quality). A frequency offset at any lower frequency in the receiver front end is always referenced to the original L-band frequency of the SV. In other words, frequency offsets are the same at baseband as they are at L-band. Therefore, the equation for the Doppler effect caused by the reference oscillator frequency offset (regardless of its specified frequency) is

$$f_{DO} = f_L \frac{\Delta f}{f 10^6} = f_{LMHz} \frac{\Delta f}{f_1} \text{ (Hz)} \tag{8.33}$$

where f_{LMHz} is the L-band carrier frequency expressed in units of megahertz and $\frac{\Delta f}{f_1}$ is the ppm value ±0.5, ±1.0, and so forth without the "million" in the denominator.

Code Range Uncertainty

The optimum code search increment is ½-chip, so the search engine code range uncertainty for a civilian GNSS signal with short code length, L_C (chips), is 2 L_C code bins, unless there is an overlay code. Overlay codes are shorter in length, $L_O < L_C$, but with longer periods than the spreading code generator ($T_O > T_C$) and considerably slower clock rates ($f_O << f_C$), resulting in uncertainty of the overlay code phasing with the replica code periods because the overlay period spans multiple spreading code generator periods. The ambiguity is typically the overlay code length.

There are two approaches that can be used by a search engine to accommodate GNSS signals with overlay codes. First, the combination of the spreading code and secondary code can be treated as a long code (with length = $L_O L_C$). This approach has the drawback that it greatly increases the search engine code range uncertainty (to $2L_O L_C$ assuming ½-chip increments) and the advantage that it permits for very long PITs. A second approach is for the search engine to treat the known secondary code as unknown data symbols, and simply use a PIT that minimizes the impact of data transitions. With this approach, or for configuring the search engine for GNSS signals modulated with navigation data, the PIT, T, should be less than or equal to the unknown data symbol period.

In the code generators of Figure 8.19, there is a code overlay sequence generator associated with its respective replica code generator. This design would be appropriate to implement the first approach described above for the search engine. The code NCO (Doppler compensated) frequency, $\hat{f}$, is divided appropriately to advance the overlay code generator at $\hat{f}_o$ and also counted so as to reset the overlay code generator at the end of its period. Both the replica code generator and the overlay code generator are reset at the beginning of their joint sequences, although this is arbitrarily at the beginning of one replica code generator period until the correct one is known after the phase ambiguity has been learned. These shifting and reset processes are Doppler compensated if the search is conducted in real time. The overlay and code generator sequences are multiplied at the replica code generator output rate, $\hat{f}$, as a 1-bit multiply using exclusive-or logic. The overlay transition boundary is synchronous with the code generator period, so the number of search engine code bins will be $2L_C L_O$. Alternatively, the search engine could require the code generators to suppress the overlay code sequence until the ambiguity is determined (as per the second approach discussed above). When the overlay code ambiguity is resolved, the replica overlay code is activated and phased accordingly. One of the many benefits of the overlay code is that when its ambiguity is resolved by the search process, this also resolves the ambiguity of the data or symbol transition boundary of the companion data channel thereby avoiding the need for a bit synchronization process (further described in Section 8.11).

Search Engine Carrier Doppler and Code Uncertainty Ranges

Table 8.14 shows the maximum carrier Doppler and code uncertainty ranges required by the search engine. This is provided for all GNSS open signals with a defined interface specification at the time of this writing. In many practical receiver designs, the code uncertainty range would be reduced significantly for those signals with very long overlay codes (e.g., GPS L1C) by initially treating the overlay code bits as unknown data symbols during the search (see Code Range Uncertainty in Section 8.4.3.1). The Doppler uncertainly computation uses (8.30) assuming a minimum user elevation angle of 15° (0.2618 radian).

The Doppler uncertainty from Table 8.14 must be increased to include the maximum user velocity using (8.32) and reference oscillator frequency offset error using (8.33) when these are unknown by the receiver during cold start.

Two case examples are used to compare the cold-start search uncertainty for the same maximum user velocity and reference oscillator offset using the GPS L1 C/A signal (λ_{L1} = 0.1903 m, f_{L1MHz} = 1,575.42 MHz) and the GPS L5 Q5 pilot

signal (λ_{L5} = 0.2548 m, f_{L5MHz} = 1,176.45 MHz). Both case examples assume a minimum user elevation of 15° (0.2618 radian), a maximum user velocity of 100 mph (44.7 m/s), and a reference oscillator maximum frequency offset of ±1 ppm. From (8.32) and (8.33):

$$f_{DUL1} = \frac{v_U \cos\alpha}{\lambda_L} = 44.7\cos(0.2618)/0.1903 = 227 \text{ Hz}$$
$$f_{DOL1} = (ppm) f_{L1MHz} = 1575 \text{ Hz}$$
$$f_{DUL5} = 44.7\cos(0.2618)/0.2548 = 169 \text{ Hz}$$
$$f_{DOL5} = (ppm) f_{L5MHz} = 1176 \text{ Hz}$$

From Table 8.14, the L1 C/A Doppler range is ±4,722 Hz, so the total Doppler search range is ±6,524 Hz. The total code search range is 2,046 half-chips. The L5 Q5 Doppler range is ±3,526 Hz, so the total Doppler search range is ±4,871 Hz and the total code search range is 409,200 half-chips.

A rule-of-thumb estimate of the useful bandwidth of one Doppler bin is $2/(3T)$ where T is the dwell time per cell consisting of one Doppler bin and one code bin. For the C/A-code, a typical search dwell time under good signal to noise ratio conditions is T = 1 ms, so each Doppler bin width is 667 Hz. For L5 Q5, if the search engine is correlating over a full repetition of the Neuman-Hofman (NH) overlay code that is 20 bits long with each bit period equal to the 1-ms spreading code period, then T = 20 ms and the Doppler bin width is 33 Hz.

With the above assumptions, the L1 C/A signal requires 21 Doppler bins (an odd number is always required to search symmetrically above and below the first, 0 Hz, Doppler bin) and the L5 Q5 signal requires 293 Doppler bins. The L1 C/A signal requires a total of 21 × 2,046 = 42,966 search cells while the L5 Q5 signal requires 293 × 409,200 = 119,895,600 search cells. The number of search cells for Q5 can be reduced to 15 × 20,460 = 306,900 if the NH code is initially treated as unknown data and using T = 1 ms, but sensitivity is diminished and the overlay code ambiguity must be resolved after the code ambiguity is resolved.

Importantly, there are false peaks when the L5 Q5 replica code is aligned with the incoming code but the NH replica overlay is misaligned with the incoming overlay code [41]. Such peaks occur for any first GNSS signals with an overlay code. These peaks are not encountered when the first approach outlined in Section 8.4.3.1 (Code Range Uncertainty) is used within the search engine to initially treat the overlay code bits as unknown data symbols. This approach is followed in [41], which also outlines a simple method to resolve the overlay code timing ambiguity. Once the spreading code replica is aligned with the incoming spreading code, there are only L_O possible starting times for a length-L_O overlay code. Successive correlation sums using a PIT equal to the spreading code repetition period can be correlated against the overlay code under all L_O start time possibilities, and the largest result indicates which possibility was correct.

Using the second approach outlined in Section 8.4.3.1 (Code Range Uncertainty) in which the combination of spreading code and overlay code are correlated against the incoming signal, false peaks can be avoided in at least two ways. First, if all of the cells in the two-dimensional code/Doppler uncertainty space are searched

Table 8.14 GNSS Open Signal Code and Carrier Doppler Search Engine Ranges

Constellation	R_{ES}[1] T_{orbit}[1]	*Doppler Range α ≥ 15° (Hz)*	*Open Signals (Pilot Only)*	*Carrier Frequency (MHz)*	*Code: L_C (T)[1] Overlay: L_O (T)[1] L = chip length T = period*	*Code Range (1/2-chips)*
GPS	*26,578 km* 11 hours/58 minutes/2 seconds = 43,082 seconds	4,722	L1 C/A	1,575.420	1,023 (1 ms) no overlay	2,046
		4,722	$L1C_P$	1,575.420	10,230 (10 ms) 1,800 (1,800 ms)	36,828,000
		3,679	L2 CL	1,227.60	767,250 (1.5 s) no overlay	1,534,500
		3,526	L5 Q5	1,176.450	10,230 (1 ms) 20 (20 ms)	409,200
GLONASS	25,478 km 11 hours/14 minutes/30 seconds = 40,472 seconds	5,098 $N_G(-7)$ 5,100 $N_G(-6)$ 5,119 $N_G(0)$ 5,120 $N_G(5)$ 5,122 N (6)	L1OF	1,602 + $0.5625N_G$ $N_G = -7, -6, -5, -4, -3, -2 -1, 0, 1, 2, 3, 4, 5, 6$	511 (1 ms), no overlay	1,022
		3,965 $N_G(-7)$ 3,967 $N_G(-6)$ 3,975 $N_G(0)$ 3,982 $N_G(5)$ 3,984 N (6)	L2OF	1,248 + $0.4375N_G$ $N_G = -7, -6, -5, -4, -3, -2 -1, 0, 1, 2, 3, 4, 5, 6$	511 (1 ms), no overlay	1,022
		3,835	L3OCp	1,202.025	10,230 (1 ms), 10 (10 ms)	204,600
Galileo	29,600 km 14 hours/4 minutes/41 second = 50,581 seconds	4,022	E1 CBOC-	1,575.42	4,092 (4 ms), 25 (100 ms)	204,600
		3,003	E5a-p	1,176.450	10,230 (1 ms), 100 (100 ms)	2,046,000
		3,082	E5b-p	1,207.140	10,230 (1 ms), 100 (100 ms)	2,046,000
BeiDou	27,778 km 12 hours/52 minutes/4 seconds = 46,324 seconds	4,351	B1I	1,561.098	2,046 (1 ms), 20 (20 ms)	81,840
		3,365	B2I	1,207.140	2,046 (1 ms), 20 (20 ms)	81,840

Note 1: Each constellation orbit radius and period and each SV code length and overlay length and corresponding periods are from [39]

simultaneously (e.g., using FFT techniques), the noise on each false peak is highly correlated with the noise in the correct code/Doppler cell and an incorrect detection is highly improbable. Secondly, if a sequential search is used, once the signal is acquired, the other $L_O - 1$ timing possibilities for the overlay code can be monitored to see if any of these possibilities yields correlation sums with greater power (which would both indicate that the overlay code was mistimed when acquisition first occurred, and also how to resolve the overlay timing ambiguity correctly).

There are obviously numerous other acquisition strategies for GNSS signals with overlay codes, but from the above discussion it should be clear that GNSS

signals with short spreading codes and no overlay code (e.g., GPS C/A-code) are the better search engine choice under favorable signal acquisition circumstances.

8.4.3.2 Basic Time-Domain Search Functions

Figure 8.22 depicts the digital channel time-domain search mode functions for one carrier Doppler and one code bin. During the search mode, the carrier and code functions are operated open loop (i.e., under step-by-step two-dimensional search control of the receiver control and processing function). The carrier Doppler and code NCO inputs used during the search pattern are precomputed and stored in a look-up table (one table for every type of search). One digital channel can synthesize only one Doppler bin per search dwell time, T, but it can synthesize multiple (M + 1) code bins per dwell time using the code generator depicted in Figure 8.23 configured for the search engine mode. Note that the only difference from the tracking

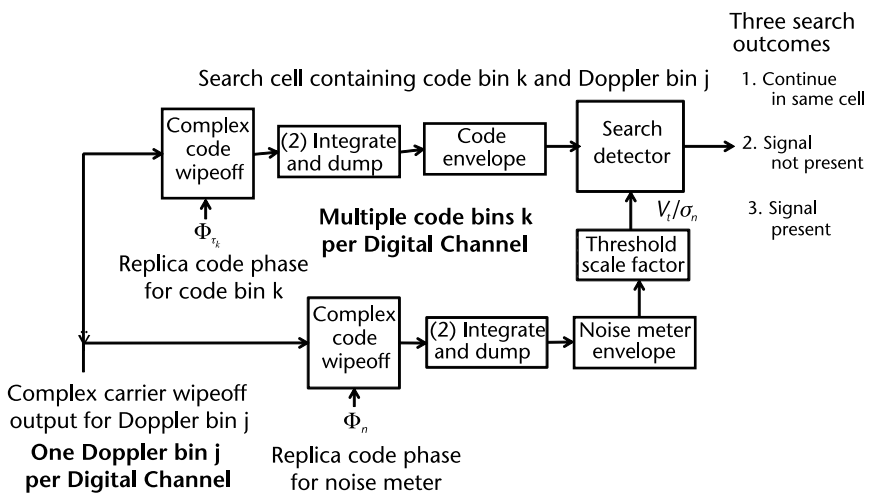


Figure 8.22 Digital channel open loop search functions for one cell.

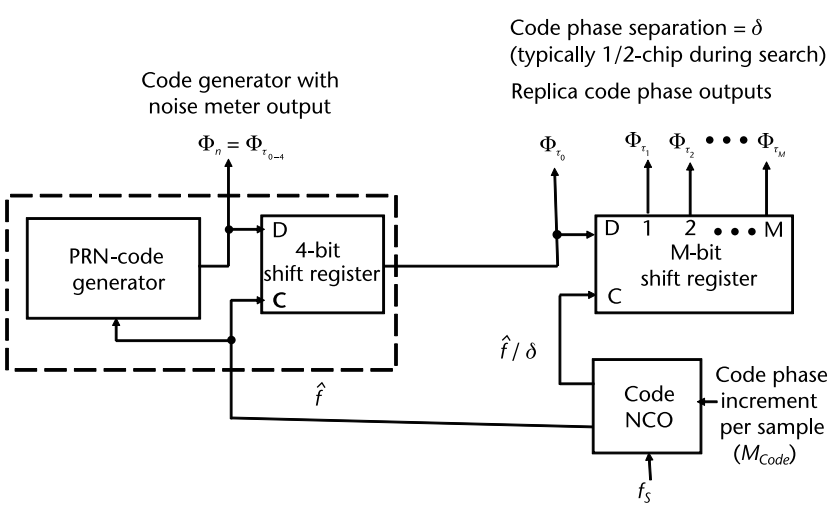


Figure 8.23 Search engine replica code generator with noise meter output.

mode is the extended (M stage) shift register. Also note that this figure illustrates how the 2-chip early noise meter code phase is synthesized. An alternate method of generating the noise meter is to use an uncorrelated signal from the replica code generator such as the G1 register of the C/A replica code generator. This method does not require the 2-chip early delay since the synthesized replica code never correlates with the incoming signal.

Beginning at the bottom left corner of Figure 8.22, the complex carrier wipe-off output for Doppler bin j is applied to two complex code wipe-off functions (one for signal and the other for the noise meter). This input corresponds to the $\tilde{I}_n, \tilde{Q}_n$ complex samples produced by the carrier wipe-off functions shown in Figures 8.13 or 8.14 where the carrier phase increment per sample ($M_{Carrier}$ or $M_{Carrier} + M_{Bias}$, respectively) of the carrier NCO has been set to synthesize Doppler bin j for this part of the search process using table look-up values. Back to Figure 8.22, at the top left corner, the replica code phase for code bin k, Φ_{τ_k} is the corresponding shift register phase output (one of the $\Phi_{\tau 0}$ to $\Phi_{\tau M}$ code phases) of the replica code generator configured for search mode shown in Figure 8.23. The code phase increment per sample (M_{Code}) in Figure 8.23 is set to the nominal spreading code rate of the incoming signal (plus the appropriately scaled code Doppler bin j value). The noise meter output, Φ_N, from Figure 8.23 is fed to the complex code wipe-off function in the lower signal path shown in Figure 8.22. These complex code and noise outputs are integrated for dwell time, T, typically 1 ms for good signal-to-noise ratio conditions, then dumped into their respective envelope functions (square root of the sum of the squares of I and Q). The noise term is multiplied by a scale factor that determines the search threshold denoted V_t/σ_n, where the numerator is the voltage threshold and the denominator is the 1-sigma noise level. Because the search threshold is critical, this scale factor is typically optimized for low, medium and high signal to noise ratios for each SV signal and T using computer simulations. Using this threshold, the search detector makes a binary decision (often called a one-shot decision) based on the signal being above or below the noise threshold. That decision (1 if true, 0 if false) is passed to a more sophisticated search detector that may require many one-shot decisions in the same Doppler and code bin to make its final decision. Three search detector decision outcomes are possible as shown in the figure (top right corner): (1) continue in the same cell (in which case nothing is changed); (2) signal not present (in which case the next unsearched cell is selected, but only after all other parallel search detectors have terminated); or (3) signal present (in which case a carrier Vernier search and peak code search are performed followed by carrier and code loop closure). Time-domain search detector designs, including search threshold, are derived mathematically in Section 8.5. The search process ends when the loop closure is successful as determined by the carrier-to-noise power ratio meter and ultimately by the phase lock indicator described in Section 8.12. The loop closure process must pull in any remaining frequency and code uncertainty. The weak link is the carrier loop typically using a frequency lock loop (FLL) carrier discriminator that has a much wider frequency error pull-in range than does a phase lock loop (PLL). The FLL will automatically transition into PLL operation if the FLL-assisted PLL loop described in Section 8.8.3 [40] is used.

Figure 8.24 illustrates a typical two-dimensional search pattern using the legacy C/A code with 1023 possible code phases as an example. In the code range

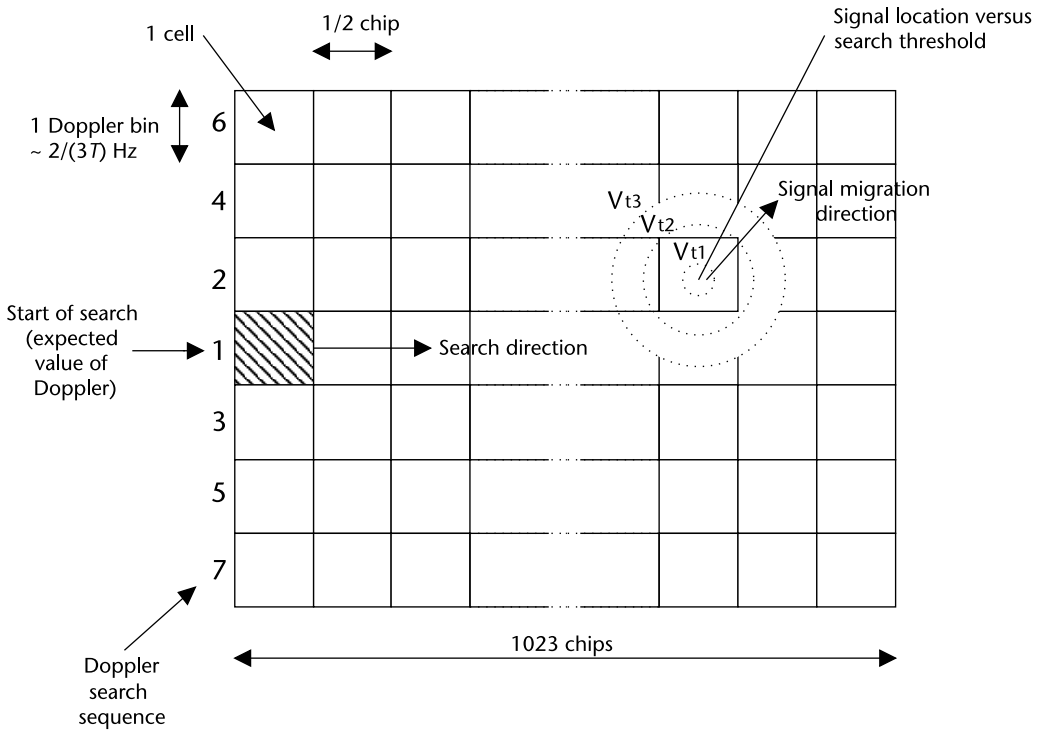


Figure 8.24 Two-dimensional C/A code search pattern.

dimension, 1,023 chips are searched in ½-chip increments. This is the optimum code bin size for the search process, but subtle code generation and correlation features are required for BOC signal ambiguity removal (described further in Section 8.5.6). The code cell search pattern is always from early to late (left to right as indicated in Figure 8.24) in the code dimension because false multipath signals always arrive late with respect to the true GNSS signal.

Figure 8.25(a) is a plot of the ideal (infinite bandwidth) BPSK code correlation envelope equation, $A^2\left(1-\frac{|\tau|}{T_{chip}}\right)$ for $|\tau| \le T_{chip}$ and 0 elsewhere, where A is the signal amplitude (assumed normalized to 1), τ is the offset (in chips) between the replica and incoming spreading symbols and T_{chip} is the period of each symbol (1 chip). The actual (finite bandwidth) correlation envelope is rounded at the peak and the transition regions are not straight lines. The figure depicts the maximum signal loss factor (0.25), amplitude roll-off (–2.5 dB) and power loss (–1.25 dB) corresponding to the worst-case alignment between the replica and the incoming spreading symbols where the replica is exactly ½-chip offset from the incoming signal as marked on the code correlation envelope. The plot is marked at all of the ½-chip intercepts but only the two equal and highest intercepts are analyzed.

There are only 7 Doppler bins shown in Figure 8.24. Note that there are always an odd number of Doppler bins because the carrier Doppler search pattern should begin centered on the mid-Doppler bin of the uncertainty region (the zero Doppler bin for cold start). As implied by the numbers to the left of the Doppler bins, a sequential search pattern is symmetrically spaced around the mid-Doppler bin. For

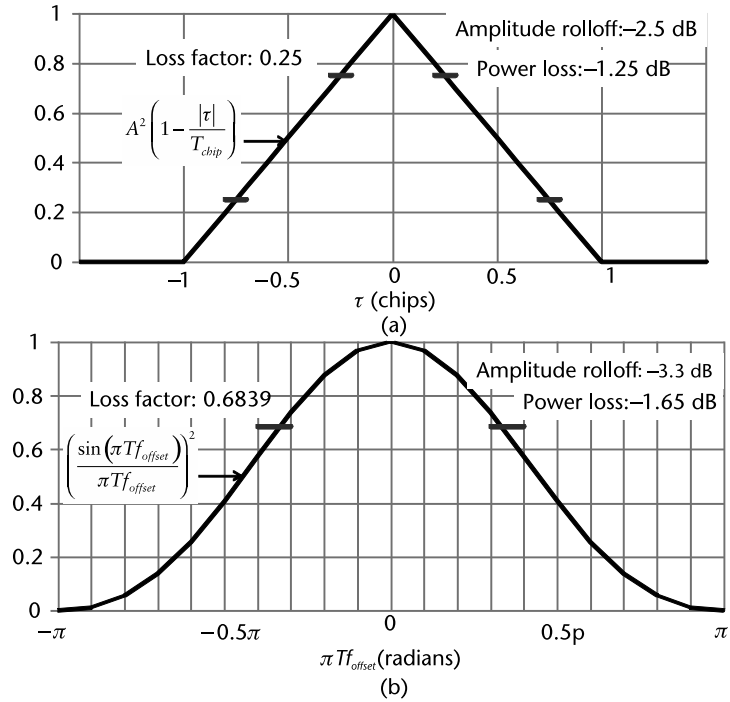


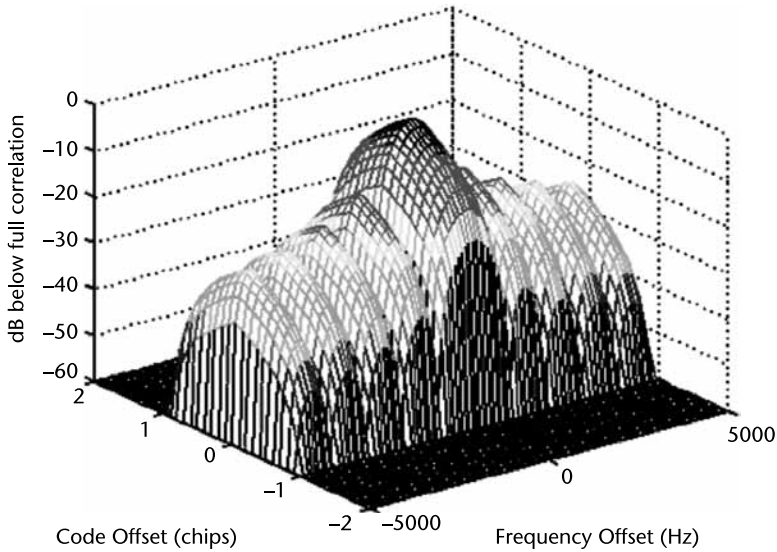
Figure 8.25 Maximum signal loss using (a) ½-chip code bins and (b) 2/(3T) Doppler bins.

cold start, the Doppler search pattern is from the highest user elevation angle to lower in a symmetrical pattern, alternating between the possibility that the SV may be rising or setting and the user velocity toward or away from the SV line of sight.

Figure 8.24 shows a useful rule-of-thumb value for the Doppler bin frequency width as $2/(3T)$ Hz, where T is the search dwell time in seconds. Figure 8.25(b) shows the maximum signal loss factor (0.6839), amplitude roll-off (–3.3 dB) and power loss (–1.65 dB), corresponding to this approximation. The figure is a plot of the loss factor equation $\left(\frac{\sin(\pi T f_{offset})}{\pi T f_{offset}}\right)^2$, where f_{offset} is the frequency mismatch between the replica and the incoming signal, T is the dwell time. The plot is marked at 2 points where $f_{offset} = \pm 1/3\text{T}$ Hz. The exact Doppler bin frequency spacing for a maximum amplitude rolloff of exactly –3 dB is $2/\pi T$ Hz. The calculations for $f_{offset} = \pm -1/\pi T$ maximum loss factor is $\sin^2(1) = 0.70807$ and for maximum power loss is $10\log_{10}(0.70807) = -1.5$ dB.

As shown in the Figure 8.24 search pattern snapshot, the peak signal is located in Doppler bin 2 and at chip 1022, specifically in the second half-chip of the replica C/A code phase state, but it is obvious that the signal could be detected in 2 nearby Doppler bins and code bins depending on the threshold voltage, V_t, setting (and the noise level, σ_n, not illustrated in the figure).

Figure 8.26 from [27] graphically illustrates the C/A code signal detection region in three dimensions. Note that the width of the code correlation response is characterized by the 2-chip wide correlation envelope that is maximum when the replica matches the incoming signal and essentially zero when mismatched by 1-chip or more on either side of the maximum. The peak amplitude of the maximum



T = 1 ms, perfect code assumed (i.e., no autocorrelation sidelobes)

Figure 8.26 Code phase and carrier Doppler frequency two-dimensional search in signal capture region.

code correlation is dependent on how well the Doppler is matched. Figure 8.26 also reveals that the carrier wipe-off output signal is characterized by a $\left(\frac{\sin(\pi f_{offset} T)}{\pi f_{offset} T}\right)^2$ response where f_{offset} is the frequency offset in hertz between the replica and the incoming signal carrier Doppler, and T is the cell dwell time in seconds. In this example, T = 1 ms, so the main lobe has nulls at ±1 kHz with offset-diminished sidelobe amplitudes spaced 1 kHz apart on either side of the main lobe.

Figure 8.27 shows how the 2-chip early noise meter code phase will never correlate with the incoming signal if any of the following (later) code phases are correlating with the incoming signal in the code search dimension (assuming the replica Doppler is closely matching the incoming signal in the carrier Doppler dimension). Obviously, the noise meter will not correlate with the incoming signal until the earliest replica signal ceases to correlate, so any signal detection dwell for

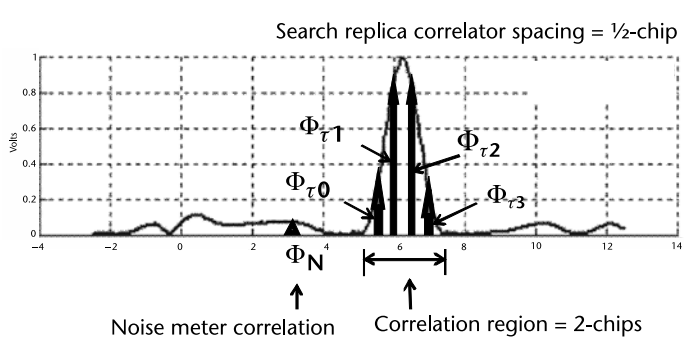


Figure 8.27 Illustration of 2-chip early noise meter correlator with respect to later ½-chip code correlators within correlation region.

the situation that the noise meter is correlating with the signal will result in immediate dismissal of that cell by the search detector. During the search mode the noise meter correlator provides the denominator of $\frac{V_t}{\sigma_n}$ used by the search detector shown in Figure 8.22. During the tracking mode the same noise meter provides the noise power estimate for the signal to noise power ratio measurement function implied by Figure 8.18.

Searching Uncertainty in Code and Doppler Dimensions

Recognizing that Figure 8.22 represents the search functions of one cell of the search pattern depicted in Figure 8.24, the time-domain search engine consists of a massive number of these cell functions operating in parallel. As a simple search engine example for the C/A code sequential search pattern described by Figure 8.24, consider that if there are 7 digital channels available for this search process, each channel having 2,046 code correlators and search detectors, then $7 \times 2{,}046$ cells can be searched simultaneously, thereby resolving all of the search uncertainty in the code dimension for 7 bins of Doppler uncertainty after all search detectors have terminated, assuming that the signal to noise ratio is equal to or higher than the expected value for the search parameters involved.

Figure 8.23 shows how the replica code generation function must be expanded in every digital channel during high uncertainty search modes (such as cold start) to support the search engine. Referring to Figure 8.23, the code NCO output designated as $\hat{f}/\delta$ uses $\delta = ½$ since the replica code spacing is ½-chip and the M-bit shift register provides M+1 replica codes. This supports a simultaneous search of M+1 half-chip code cells using M+1 search detectors. If $M+1 = 2L_C$ where L_C = replica code length (chips), then all possible combinations of ½-chip code phases can be searched in parallel (i.e., the total range uncertainty in one Doppler bin that is provided by one digital channel). If there are J digital channels available to this search engine, then $J(M+1)$ cells can be searched simultaneously. If J Doppler bins covers the worst-case Doppler uncertainty, then the search engine will find the SV after the last search detector has reported an outcome of signal found or signal not found, but only if the SV is in view and unobstructed and the signal to noise ratio is adequate for the dwell time T. The situational awareness feature in the precise gain control of the front end provides information with respect to the level of interference present that is used to select T and V_t/σ_n, but only the search process can provide signal power level awareness (i.e., the signal may or may not be obstructed even if the receiver knows that it should be visible). The only way to determine the signal power level is by the search process outcome. The antenna noise temperature rises when the receiver is operated indoors, but this small increase in thermal noise is not a reliable indicator of signal obstruction because there are so many other sources of increased noise.

Searching with real-time processes can take a considerable amount of time when the carrier Doppler uncertainty and the code range uncertainty are high. Recall that the real-time search process involves correlation between the replica code and the incoming signal code following the carrier wipe-off process until the replica carrier Doppler phase and the replica code phase closely match simultaneously with the incoming signal. As seen in Figure 8.24 this uncertainty is mapped in

discrete Doppler bins and code bins whose intersections are called cells. From the perspective of using the same type of real-time processes to build a search engine that is used during steady state tracking, the fastest real-time signal acquisition time is achieved when there are enough real-time processing resources to examine all of the cells simultaneously. Lacking those resources, the real-time search process must repeat the search pattern systematically either until the total uncertainty has been searched or the signal is found earlier in the search pattern. Described next are much faster and more computationally efficient search engine signal acquisition techniques using frequency domain processing made practical by modern ultrahigh-speed DSP technology.

8.4.3.3 Frequency Domain Search Engine

Instead of the ADC sampled data stream being processed sequentially sample by sample in real time using parallel hardware, frequency-domain techniques process a block of samples (e.g., N samples at a time corresponding to the real-time dwell period, T). For this reason, GNSS frequency-domain processing is often called block processing. All of the associated frequency domain processes must be completed faster than T, corresponding to the time interval of that block of data (i.e., the entire block of data must be fully processed by the time the next block of data arrives in real time, but interim results can be accumulated noncoherently until the predicted signal-to-noise ratio is sufficient for signal detection or the signal is not found after exploring all signal uncertainty). Block processing is made practical by the FFT version of the discrete Fourier transform (DFT) in combination with modern DSP technology designed to support discrete frequency-domain processing, specifically using the FFT algorithm, its complex conjugate, and its complex inverse, available as turnkey programs that run significantly faster than the underlying real time that marches on while this frequency-domain processing is taking place.

FFT Versus DFT Computational Efficiency

The DFT and FFT both produce the same results, but for a large number of complex samples, N, the computational efficiency of the FFT is orders of magnitude faster. The DFT execution time is $k_{DFT}N^2$ and the FFT execution time is $k_{FFT}N \log_2 N$, where the k factors are constants of proportionality [42].

A case example for $N = 4{,}096$ complex samples (hereafter called "point" in this example) using the Texas Instruments TMS320VC5505 DSP dramatically demonstrates the processing speed advantage of the FFT over the DFT for large N. Start with the specified 1,024-point FFT in 7,315 cycles in [43] that takes 7,315/150 = 48.8 μs at its maximum specified clock speed of 150 MHz, so $k_{FFT}N \log_2 N = 4.88\text{E-}05$ seconds and $N \log_2 N = 10{,}240$, then $k_{FFT} = 4.88\text{E-}05/10{,}240 = 4.77\text{E-}09$. The value for k_{DFT} is typically about a factor of 2.5 longer than for k_{FFT} [42], so assume $k_{DFT} = 1.19\text{E-}08$. For a 4,096-point DSP, the execution time for the DFT is $k_{DFT}N^2$ = 200 ms and for the FFT is $k_{FFT}N \log_2 N = 234$ μs. Therefore, the 4,096-point FFT is 853 times (nearly 3 orders of magnitude) faster than the 4,096-point DFT, plus the FFT is more accurate because there are fewer computations resulting in lower round-off error. Refer to Chapter 12 of [42] for more details on DFT and

FFT implementation, including programs written in BASIC for the complex DFT, FFT, inverse FFT (IFFT) as well as FFT and IFFT for real signals.

As shown in Figures 8.13 and 8.14, the input signal can be complex or real, respectively [i.e., each input stage contains either a complex value with a real and imaginary part or a real value (with only a real part)]. In the case of GNSS signal samples, the real and imaginary parts are the in-phase (I) and quadra-phase (Q) samples, respectively. Figure 8.28 adapted from [44] illustrates how a block of N samples is transformed by the DSP for both the (a) real and (b) complex case using $N = 8$ and $n = 3$ in both cases. (Note that N is intentionally short for simple graphic illustration purposes only and is not a typical number of points.) The typical FFT transforms only complex inputs and is based on the lower left part of Figure 8.28 (complex DFT and FFT in the time domain) with the restriction that $N = 2^n$, n = any positive integer, also called radix 2 (i.e., $\log_2 N = n$). The crosshatched time-domain samples suggest how the FFT can transform a real signal by showing the common values between the real DFT and the complex DFT. Within the complex DFT (and FFT), each sample corresponds to an ordered memory location that contains two values, one real and one imaginary (e.g., I and Q samples). So any real signal can be converted into a complex signal and transformed by the conventional FFT by simply attaching an imaginary value of length N containing all zeros. Also, since the FFT operates on complex signals of length N based on some power of 2, then any input signal sample length that contains fewer samples than this is made equal to length N by adding zeros into the remaining samples, called zero padding. As illustrated in [42], a real FFT program can be written that accepts the complex FFT format with the imaginary part filled with all zeros, but does not waste computation time processing them. Also, there are now FFT programs that will convert signals that are not radix 2.

FFT Simplicity and Efficiency

This section begins with a familiar time-domain computation process that is simplified by frequency domain processing. The real-time discrete convolution process

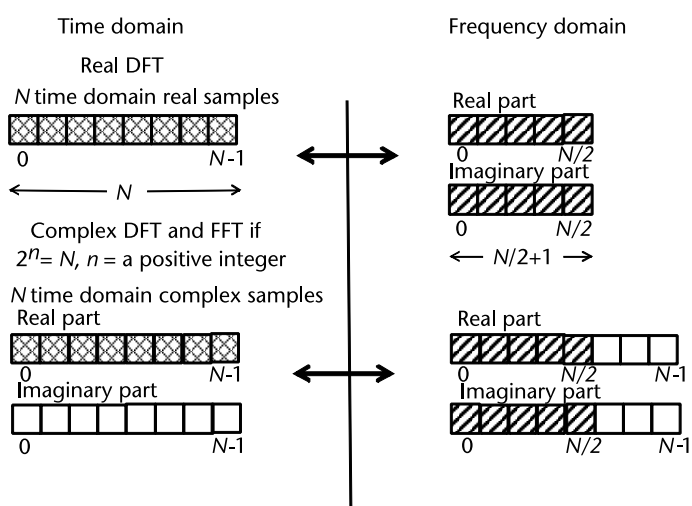


Figure 8.28 Time- and frequency-domain comparisons between real and complex DFT. Crosshatched time-domain samples show common values between the two DFTs.

between a digital filter with finite impulse response $h(n)$ and a digital signal input $x(n)$ is usually abbreviated as $x(n)*h(n) = y(n)$, where the star symbol means convolved with. The real-time convolution process is commutative (i.e., the order of the input signal and impulse response does not matter if the convolution process is computed correctly for each case). Assume that $x(n)$ is an N point digital input signal running from 0 to $N - 1$ and $h(n)$ has an M point finite impulse response running from 0 to $M - 1$, then the convolution of the two will be an $N + M - 1$ point output signal running from 0 to $N + M - 2$ using the formal discrete convolution equation

$$y(i) = \sum_{j=0}^{M-1} h(j)x(i-j) \quad i = 0,1,2,\ldots,M+N-1 \tag{8.34}$$

The convolution algorithm is not as simple as it looks. The index $i = 0$ to $M + N - 1$ always defines the length of the output, $y(i)$. Refer to Chapter 6 of [42] for two different BASIC programs that perform the convolution process of (8.34), one from the viewpoint of the input signal and the other from the viewpoint of the output signal. The index in $x(i - j)$, is $i = 0$ to $N - 1$ for the outer loop of the input signal viewpoint algorithm and i = 0 to $M + N - 1$ in the outer loop of the output signal viewpoint algorithm except no computation is performed for either of the inner loop conditions: $i - j < 0$ or $i - j > N - 1$.

The convolution operation is simplified in the frequency domain because convolution in the time domain becomes multiplication in the frequency domain. After converting both real-time functions into the frequency domain as $X(k) = \text{FFT}\ [x(n)]$ and $H(k) = \text{FFT}\ [h(n)]$, the following frequency-domain equation is used

$$Y(k) = X(k)H(k) \quad k = 0,1,2,\ldots,M+N-1 \tag{8.35}$$

This equation is derived from (8.34) in [42]. Clearly, it is easier to implement the time domain convolution process using multiplication in the frequency domain and it also turns out to be computationally efficient for a large number of samples. The convolution process in the time domain is simply the inverse FFT of the frequency domain product using the equation

$$y(i) = IFFT\left[Y(k)\right] = IFFT\left[X(k)H(k)\right] \quad i = 0,1,2,\ldots,M+N+1 \tag{8.36}$$

Note that sampled data that had a continuous real-time source can have poor resolution using discrete signal processing. The question naturally arises as to how much resolution can be obtained in the FFT discrete frequency response. The answer is it can be infinitely high if the impulse response is padded with an infinite number of zeros (i.e., there is nothing limiting the frequency resolution except the length of the FFT). As noted earlier, in order to comply with the radix 2 FFT requirement, there is usually some zero padding and that increases the frequency resolution. A related issue is that the impulse response is usually a real discrete sampled signal but it represents a continuous frequency response. As observed in Figure 8.28, an N-point real DFT of this impulse response provides $N/2 + 1$ samples of this continuous signal. If the DFT is made longer, then the resolution improves and

comes closer to becoming the original continuous signal. For example, suppose it were possible to add an infinite number of zeros to the time-domain signal. This would produce a time-domain signal that has an infinitely long period (i.e., an aperiodic signal). Since the frequency domain would achieve an infinitesimally small spacing between samples, it would become a continuous signal. However, the DFT considers the time-domain signal to be infinitely long and periodic (i.e., it assumes the N points are repeated over and over from negative to positive infinity in the time domain in order for the frequency-domain result to be aperiodic). The same analogy applies to the FFT operation on complex time-domain signals.

With this insight into the benefits as well as the limitations of the FFT for simplifying the convolution process, the next step is to use the same computational efficiency and simplicity to perform the (less familiar) code correlation process that takes place in a GNSS receiver search engine (i.e., the process of despreading an incoming PRN signal using a replica of this signal). That real-time computation is very similar to the convolution equation, but instead of the finite length impulse response $h(n)$, the replica code signal will be represented by $y(n)$ with a periodicity of M points running from 0 to $M - 1$ and $x(n)$ is the digital input signal (after the carrier wipe-off process) with the same periodicity of M points running from 0 to $M - 1$ and the correlation of the two will be periodic with an M point output signal running from 0 to $M - 1$ using the following equation

$$z(i) = \sum_{j=0}^{M-1} x(j)y(i+j) \quad i = 0,1,2,\ldots M-1 \tag{8.37}$$

Note the sign and the output length difference between (8.37) and (8.34), but like (8.34) it is commutative. If the impulse response were the same length as the input signal in (8.34), the output length would be $2M$ for the convolution process but is only M for the (8.37) correlation process. The correlation process is less complicated since it is simply a circular shift in increments of 1 point of all M points of the replica $y(n)$ with the fixed phase of one period of the input signal $x(n)$ and, at each phase shift, all the points are multiplied and added to produce one output point. The index i where the two signals most closely align is where maximum correlation occurs. This correlation equation has the following frequency domain counterpart, using $X(k)$ = FFT $[x(n)]$, $Y(k)$ = FFT $[(y(n)]$, and $Y^*(k)$ is the complex conjugate of $Y(k)$

$$Z(k) = X(k)Y^*(k) \quad k = 1,2,\ldots M-1 \tag{8.38}$$

This equation is derived from (8.37) in [41, 44]. Note that there is an additional step of performing the complex conjugate of $Y(k)$ involved when using the FFT to perform the correlation process. Performing the complex conjugate of $Y(k)$ is simply changing the sign of the imaginary part of $Y(k)$. The last step is to compute the real-time correlation output as follows

$$z(i) = IFFT\left[Z(k)\right] = IFFT\left[X(k)Y^*(k)\right] \quad i = 0,1,2,\ldots M-1 \tag{8.39}$$

GPS C/A Code FFT Acquisition Schemes

The earliest frequency-domain acquisition techniques were developed for use with the GPS C/A code that is only 1,023 chips long and has no overlay codes. As observed in Table 8.14, the modernized GNSS spreading codes are much longer and most have overlay codes, so this significantly increases the acquisition time, thereby making frequency-domain acquisition techniques even more appealing, but also increasing the DSP computational burden. Figure 8.29 is a high-level block diagram that illustrates two GPS L1 C/A code block processing acquisition schemes called: (a) parallel frequency [44] and (b) parallel code (also called circular correlation) [44, 45].

Referring to Figure 8.29(a), [44] uses a front-end signal with a low IF of 1,250 kHz and an ADC sample rate of 5 MHz with code wipe-off performed as the first step of the acquisition process. However, the signal detection process in [44] is not the frequency-domain detection technique as shown in scheme (a). Instead, the maximum time-domain amplitude resulting when the Doppler compensated replica C/A code matches the phase of the incoming C/A code within less than ±1 chip. The Doppler compensation is only approximated in discrete steps, but when it closely approximates the IF plus Doppler (plus reference oscillator frequency offset) the resulting signal output becomes a pure CW signal. That coincidence ideally produces a single line spectrum in the frequency domain, but practically several lines of different amplitudes when the correlation region is entered.

As indicated in Figure 8.29(a), the detection process takes place in the frequency domain by selecting the line with the maximum amplitude that also exceeds a predetermined threshold (above the expected noise floor). However, this signal coincidence is also occurring and can also be detected in the first-stage time-domain process using similar search threshold techniques. There are multiple strategies for refining this signal acquisition scheme, but the fundamental scheme used in [44] is described as follows.

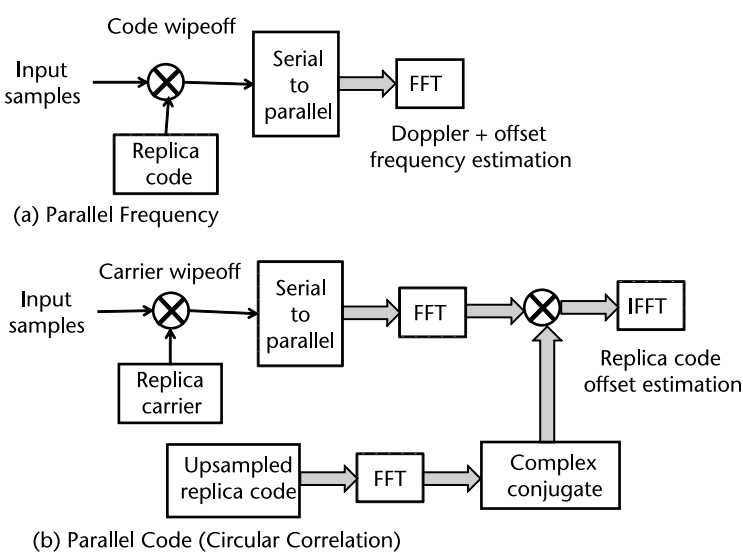


Figure 8.29 Two block processing FFT acquisition schemes.

The replica code shown in Figure 8.29(a) consists of 21 precomputed code bins, each code bin containing a precalculated replica C/A code, C_S, for satellite PRN number S of length 1,023 chips, sampled at the same 5-MHz rate of the incoming signal, producing $C_S = 5{,}000$ samples. There are $C_S/1{,}023 = 4.888$ samples per C/A code chip. The assumed total Doppler plus offset uncertainty is ±10 kHz, so each precalculated Doppler compensated code bin is obtained using $C_S e^{j2\pi f_i t}$ where f_i = 1,250-10, 1,250-9, …, 1,250-1, 1,250 +0, 1,250+1, …, 1,250+10 (kHz) and $i = 21$ frequency bins, each bin separated by 1 kHz [recall that the rule-of-thumb Doppler bin spacing is $2/(3T) = 0.667$ kHz]. The first 5,000 incoming real samples are read and the initial code phase index is set $k = 1$. These real samples are complex multiplied by the 5,000 complex samples of all 21 frequency bins, one bin at a time and point by point, resulting in each frequency bin containing 5,000 complex samples that represent $T = 1$ ms of real time. For each bin, each of the 5,000 real and imaginary output values are squared, all of these values are added and the square root of this sum becomes one output frequency bin. This process is the basis for the name of this scheme being called parallel frequency. At this point, all 21 parallel frequency bins contain an amplitude corresponding to $k = 1$ of a possible 5,000 code offsets. Each code offset represents 1/4.888 = 0.205 chip (recall that ½-chip code spacing is optimum). Then k is incremented by 1 (circular shift of the incoming signal by 0.205 chip) and the same process is repeated for the remaining 20 parallel frequency bins. This process continues until all possible incoming code phases have been circular shifted as indicated by $k = 5{,}000$. At the end of these iterations, a matrix results with $21 \times 5{,}000$ (105,000) amplitudes organized in 21 frequency bin rows and 5,000 code bin columns similar to Figure 8.24. These amplitudes have been generated in less than 1 ms, which is as required faster than the next 1-ms block of real-time 5,000 input samples. All 105,000 amplitudes are examined and all that exceed a threshold are selected, then the highest amplitude among these is selected. That selection corresponds to the kth input code phase shift and the ith Doppler bin. Then a fine frequency process follows prior to entering the tracking process. For weaker signals, this process can be repeated several times using a new 1-ms input each time and the amplitudes noncoherently integrated before the detection process is performed.

The signal detection process could have been in the frequency domain as depicted in Figure 8.29(a) using a constant frequency search scheme similar to that described for Figure 8.24 that would produce 5,000 complex samples in each code bin obtained by summing all 5,000 I components and all 5,000 Q components into one complex point per bin. The serial-to-parallel process prior to the FFT shown in Figure 8.29(a) could be accomplished as k is indexed at constant frequency. The zero padding (if needed) takes place in the serial-to-parallel operation where the target memory that is based on radix 2 is completely zeroed, but the serial transfer of samples is shorter. The end result is zero padding in the remaining least significant bit memory locations. The detection scheme is carried out in the frequency domain and no inverse FFT is required.

The parallel frequency scheme provides good insight into the use of a DSP for fast signal acquisition, but is not computationally efficient nor is it a frequency domain scheme in [44], that is, the calculation sequence would have to be sequential frequency as described above in order to use the frequency-domain scheme

efficiently. The parallel code, also known as circular correlation scheme in Figure 8.29(b), is far superior in computational efficiency and speed.

Referring to Figure 8.29(b), the carrier wipe-off process is performed first followed by serial to parallel conversion for block processing (depicted by block arrows) with the block size equal to one period of the replica code with no greater than ½-chip spacing. The time-domain carrier wipe-off process can be by either of the methods shown in Figures 8.13 or 8.14. Then the FFT is performed on the complex signal and multiplied by the complex conjugate of the FFT of upsampled replica code. The replica code sampling must be the same as the incoming signal. Each cycle of the circular correlation process produces all correlation combinations for one Doppler bin.

In [45], a 10-kHz Doppler uncertainty range is searched assuming negligible contributions by user velocity and reference oscillator frequency offset. This is searched in 1-kHz frequency bins [recall that the rule-of-thumb Doppler bin spacing is $2/(3T) = 0.667$ kHz]. The block size used for the C/A code signal corresponds to a dwell time of $T = 1$ ms (i.e., one C/A code period, $T_c = 1$ ms). The circular correlation cycle is repeated K times in the same Doppler bin, each cycle using a new 1-ms input block. The detection process is in the time domain where the absolute value of each code offset cell is taken and summed from $i = 0$ to $K - 1$. The value $K = 20$ is used in order to increase the signal to noise ratio enough to detect the offset code under weaker signal conditions than for the (a) example where $K = 1$. No details are provided about the search pattern, signal detection process or the uncertainty refinement processes that must follow before the tracking loops are closed. Radix 2 FFT processing is used, so both the incoming signal length $2L =$ 2,046 ½-chip samples and the upsampled signal length $2L = 2{,}046$ ½-chip samples are padded with two zeros, so that the length of both is 2,048 samples.

Reference [45] reported a 500-ms acquisition time using this scheme with their existing DSP that could support an 8 times faster acquisition than the real-time signal acquisition scheme that took 4 seconds. Obviously, the parallel code acquisition scheme is much faster than its real-time sequential search counterpart. How much faster depends on the speed of the DSP. To achieve the maximum speed, all 10 frequency bins would have to be processed in less than $T_c = 1$ ms for each K block (i.e., using the same input data block). This requires ten 2,048-point FFTs, ten 2,048-point complex multiplications, and ten 2,048 IFFTs in less than 1 ms to allow time for the remaining overhead processes. The upsampled replica code FFT and its complex conjugate are only computed once, so it can be precomputed. Although the input signal code contains code Doppler, this is neglected in this scheme. The detection process must wait until K blocks of 2,046 samples have been padded and processed. Assuming that $K = 20$, all of the above processing including the detection process must be computed within $K\,T_c = 20$ ms. This requires a DSP that is 25 times faster than in [45], resulting in signal acquisition 200 times faster than its real-time counterpart.

Although the ADC sample rate is not mentioned, it is a commensurate sampling frequency of 2.046 MHz designed to produce a C/A code signal block size of length $L = 2{,}046$ ½-chips for a 1-ms dwell time (i.e., the optimum search spacing). Commensurate sampling violates the first rule for the sampling frequency described in Section 8.3.7. Figures 8.30 and 8.31 illustrate the adverse consequences of using commensurate sampling.

Figure 8.30 depicts the correlation distortion that occurs in a low IF front-end design when the ADC sampling rate, $f_S = 8\ f_0$, where $f_0 = \text{P(Y)}$ spreading symbol frequency and the front-end IF, $f_{IF} = 2\ f_0$. Figure 8.31 depicts a similar but slightly smaller distortion when the ADC sampling rate remains commensurate, $f_S = 8\ f_0$, but $f_{IF} = 2.01\ f_0$ is asynchronous. The reduced effect is because the spreading code is oversampled. The P(Y) code signal correlation uses an exaggerated high signal-to-noise ratio to magnify the commensurate distortion visually. The same effect is

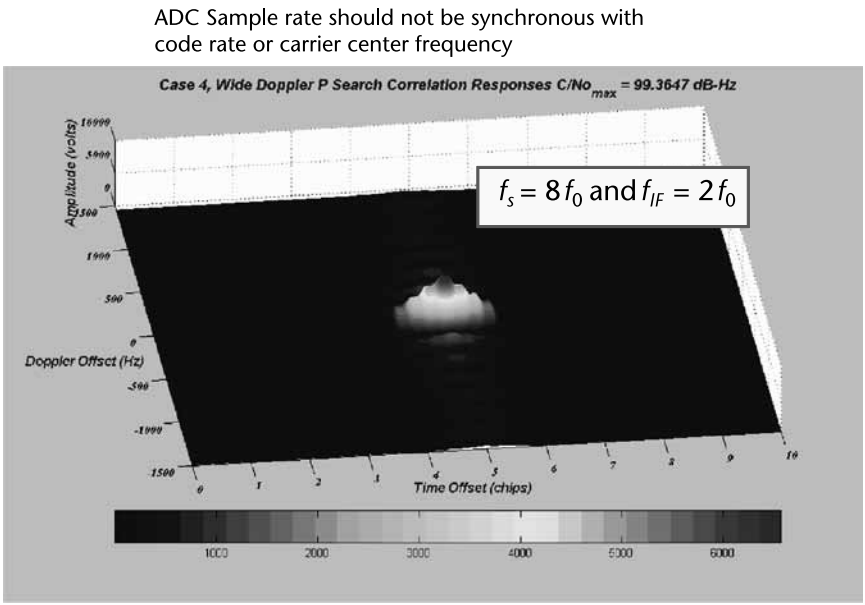


Figure 8.30 Adverse effect of commensurate sampling (ADC sampling rate synchronous with spreading symbol frequency and IF). (Graph provided courtesy of Logan Scott, L. S. Consulting, Inc.)

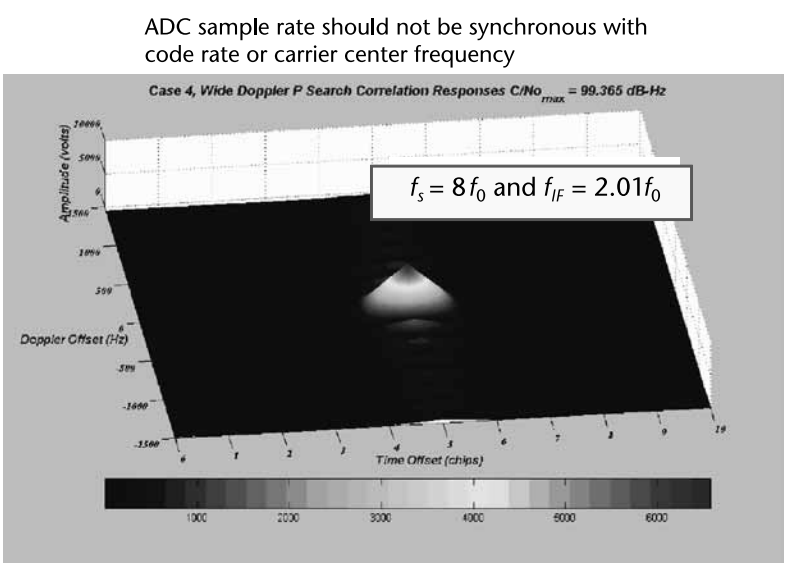


Figure 8.31 Adverse effect of commensurate sampling (ADC sampling rate synchronous with spreading symbol frequency but asynchronous with IF). (Graph provided courtesy of Logan Scott, L. S. Consulting, Inc.)

experienced with any spreading symbol that has been commensurately sampled. The correlation distortion is caused because of the lack of assured migration of the samples that prevents consistent sampling in the spreading symbol transition boundary regions [i.e., where the value is midway between +1 and –1 (zero)]. This distortion is avoided by obeying this rule in the front-end design and by synthesizing the replica code at the same asynchronous sampling rate. This may penalize the efficiency of the DSP signal acquisition scheme because there will be not be exactly 2 samples per code chip, as was the oversampled case in [44] described using Figure 8.29(a). This rule presents no problem to the real-time signal acquisition and tracking process because the code NCO and shift register combination readily accommodate ½-chip spacing (or other powers of 2 in the denominator) to a very high resolution regardless of the sampling frequency.

8.5 Acquisition

The basic concepts of modern GNSS signal acquisition from a high-level design perspective were introduced in the previous section. There is a large amount of literature on legacy GPS signal acquisition in direct sequence receivers. Reference [46] described legacy GPS receiver search techniques when time-domain acquisition was the only viable signal processing approach using custom ASIC components in the baseband hardware plus the limited microprocessor power available at the time. Although the baseband architectures are now evolving toward software-defined implementation, the real-time acquisition techniques are still the most viable when the search uncertainties are small. Reference [47] described rapid signal acquisition techniques for the legacy GPS signals using DSP-based FFT processing that introduced advanced frequency-domain processing concepts and also provides an extensive list of references relating to GPS acquisition schemes using frequency-domain signal processing techniques. Reference [48] provided a comprehensive review of current FFT-based acquisition architectures being considered for a next-generation GNSS receiver with a stated objective of minimizing FPGA resources in that architecture. This section focuses not only on refinement of the acquisition architecture for the search engine but also on the underlying theory of search detection.

8.5.1 Single Trial Detector

Detection processes typically begin with a single trial detector that is a statistical process because each cell either contains noise only or signal plus noise. Each of these two cases has a unique probability density function (pdf). Figure 8.32 illustrates the four possible outcomes using these same two pdfs for a single trial detector based on a complex signal envelope that will be analyzed later. As was shown in Figure 8.22, the single trial threshold is V_t/σ_n and this is chosen to provide an acceptable probability of false alarm, P_{fa}. As is shown in Figure 8.32, σ_n has been normalized to unity, so any cell envelope that is at or above the threshold, V_t, is detected as the presence of the signal. Any cell envelope that is below the threshold, V_t, is detected as noise. There are four outcomes of the single trial (binary) detection processes illustrated in Figure 8.32: two are wrong and two are correct. The single

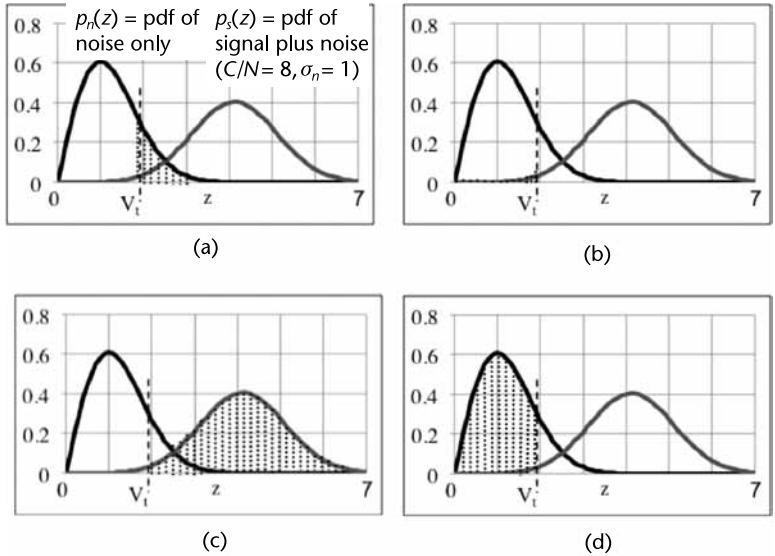


Figure 8.32 Probability density functions for single trial detector of a complex envelope: shaded areas represent: (a) probability of false alarm (used), (b) probability of false dismissal (not used), (c) probability of detection (used), and (d) probability of correct dismissal (not used).

trial probability can be computed by an appropriate integration with the threshold, V_t, as one limit and infinity or zero as the other. These integrations are shown as the shaded areas in Figure 8.32. The two statistics that are actually used for signal detection are the two on the left side of Figure 8.32: (c) single trial probability of detection, P_d, and (a) single trial probability of false alarm, P_{fa}. If σ_n has not been normalized, these are defined as

$$P_d = \int_{V_t/\sigma_n}^{\infty} p_s(z)\,dz \tag{8.40}$$

$$P_{fa} = \int_{V_t/\sigma_n}^{\infty} p_n(z)\,dz \tag{8.41}$$

where:

$p_s(z)$ = pdf of the envelope containing signal plus noise;
$p_n(z)$ = pdf of the envelope containing noise only.

If the root mean square of the signal amplitude (signal only) is formed by the envelope $A_{Env} = \sqrt{I_S^2 + Q_S^2}$, where the in-phase (I) and quadra-phase (Q) components have statistically independent Gaussian pdfs, then $p_s(z)$ is a Ricean distribution (similar to [49, pp. 693–694]) defined as

$$p_s(z) = \frac{z}{\sigma_n^2} e^{-\left(\frac{z^2 + A_{Env}^2}{2\sigma_n^2}\right)} I_0\left(\frac{A_{Env} z}{\sigma_n^2}\right) \text{ for } z \geq 0; p_s(z) = 0 \text{ otherwise} \tag{8.42}$$

where

z = random variable;

σ_n^2 = noise variance;

$I_0\left(\frac{A_{Env} z}{\sigma_n^2}\right)$ = modified Bessel function of the first kind with order zero;

$$I_0(x) = 1 + \frac{(x/2)^2}{(1!)^2} + \frac{(x/2)^4}{(2!)^2} + \frac{(x/2)^6}{(3!)^2} + \ldots$$

Defining the dimensionless predetection carrier-to-noise power ratio as $C/N = A_{ENV}^2 / 2\sigma_n^2$, as presented to the single trial detector, then (8.42) for $z \cdot 0$ can be expressed as

$$p_s(z) = \frac{z}{\sigma_n^2} e^{-\left(\frac{z^2}{2\sigma_n^2} + C/N\right)} I_0\left(\frac{z\sqrt{2\,C/N}}{\sigma_n}\right) \tag{8.43}$$

where

$C/N = (C/N_0)T$

N_0 = noise power in a 1 Hz bandwidth (W)

T = search dwell time (s)

The pdf $p_s(z)$ for the case where both signal and noise are present in (8.43) is plotted in all examples of Figure 8.32 for $C/N = 8$ and $\sigma_n = 1$ (normalized).

For the case where there is no signal present, then evaluating (8.42) for $A_{Env} = 0$ yields a Rayleigh distribution for $p_n(z)$ defined as

$$p_n(z) = \frac{z}{\sigma_n^2} e^{-\left(\frac{z^2}{2\sigma_n^2}\right)} \tag{8.44}$$

The pdf $p_n(z)$ for the case where only noise is present in (8.44) is plotted in all examples of Figure 8.32 with $\sigma_n = 1$ (normalized). The mean of the Rayleigh distribution is $\mu_{Ray} = \sigma_n\sqrt{\pi / 2}$ and the variance is $\text{var}_{Ray} = \sigma_n^2(2 - \pi/2)$. The result of integrating (8.41) using the pdf of (8.44) is

$$p_{fa} = e^{-\left(\frac{V_t^2}{2\sigma_n^2}\right)} \tag{8.45}$$

Taking the natural log of both sides of (8.45) and solving for the threshold value in terms of a desired single trial probability of false alarm

$$V_t/\sigma_n = \sqrt{-2\ln P_{fa}} \tag{8.46}$$

For example, if it is desired that $P_{fa} = 16\%$, then $V_t/\sigma_n = 1.91446152$ and the single trial probability of detection, P_d, can be computed for the expected C/N using (8.40) and (8.43) with $\sigma_n = 1$ (normalized) as follows

$$\begin{aligned} P_d &= \int_{V_t}^{\infty} ze^{-\left(\frac{z^2}{2}+C/N\right)} I_0\left(z\sqrt{2\,C/N}\right) dz \\ P_d &= 1 - \int_{0}^{V_t} ze^{-\left(\frac{z^2}{2}+C/N\right)} I_0\left(z\sqrt{2\,C/N}\right) dz \end{aligned} \tag{8.47}$$

Some examples of single trial probability of detection, P_d, using (8.47) are shown in Table 8.15 for input C/N ratios from 1 to 9, then for each C/N computing $(C/N)_{\text{dB}} = 10\log_{10} C/N$ and tabulating the corresponding $(C/N_0)_{dB} = (C/N)_{dB} - 10\log_{10} T$ for $T = 1$, 2.5, 5, and 10 ms, with $P_{fa} = 16\%$.

By inspection of Table 8.15, the low probability of detection at C/N below 4 and especially the poor false alarm rate from a single trial detector are usually unsatisfactory for GNSS applications. Single trial search detector schemes are seldom used alone, but are combined with either variable or fixed dwell time detectors. A variable dwell time detector makes a yes or no decision in a variable interval of time if the first single trial decision is yes. If no, then a typical design will proceed immediately to the next cell (a more conservative design will require two no answers in a row to proceed to the next cell). If a maybe condition is present, it will remain in that cell (using new binary decisions for each trial) until the algorithm makes a decision (i.e., the dwell time is variable). A fixed dwell time detector makes a yes or no decision in a fixed interval of time using a vote on the outcome of a fixed number of single trials in the same cell, using new binary decisions for each trial. If the signal-to-noise ratio is good when there is a signal present, a properly tuned variable dwell time (sequential) multiple trial detector will search faster than

Table 8.15 Single Trial Probability of Detection with $P_{fa} = 16\%$

C/N	P_d	$(C/N_0)_{dB} = (C/N)_{dB} - 10\ log_{10}\ T$ (dB-Hz)			
(ratio)	(dimensionless)	$T = 1$ ms	$T = 2.5$ ms	$T = 5$ ms	$T = 10$ ms
1.0	0.431051970	30.00	26.02	23.01	20.00
2.0	0.638525844	33.01	29.03	26.02	23.01
3.0	0.780846119	34.77	30.79	27.78	24.77
4.0	0.871855378	36.02	32.04	29.03	26.02
5.0	0.927218854	36.99	33.01	30.00	26.99
6.0	0.959645510	37.78	33.80	30.79	27.78
7.0	0.978075147	38.45	34.47	31.46	28.45
8.0	0.988294542	39.03	35.05	32.04	29.03
9.0	0.993845105	39.54	35.56	32.55	29.54

a fixed dwell time multiple trial detector because it quickly dismisses the noise only conditions. As will be seen, the recommended search detector is a combination of both types of detectors with P_{fa} and V_t/σ_n adjusted to make the overall probability of false alarm of this combination of detectors suitable.

8.5.1.1 Envelope Approximations

To reduce the computational burden of forming the actual complex (signal plus noise) envelope, $A = \sqrt{I^2 + Q^2}$, there are two commonly used approximations. The most accurate but higher computational burden version of the two is the Jet Propulsion Laboratory (JPL) approximation defined by

$$\begin{aligned} &A_{JPL} = X + Y/8 \qquad \text{if } X \geq 3Y \\ &A_{JPL} = 7X/8 + Y/2 \qquad \text{if } X < 3Y \\ &\text{where} \\ &X = \text{MAX}\left(|I|, |Q|\right) \\ &Y = \text{MIN}\left(|I|, |Q|\right) \end{aligned} \tag{8.48}$$

The JPL approximation can also be expressed logically as
If $|I| \leq |Q|$, then $X = |Q|$, $Y = |I|$
else $X = |I|$, $Y = |Q|$
If $X \geq 3Y$, then $A_{JPL} = \text{X} + \text{Y}/8$
else $A_{JPL} = 7X/8 + Y/2$

The least accurate but lowest computational burden version is the Robertson approximation defined by

$$A_{Rob} = \text{MAX}\left(|I| + |Q|/2,\ |Q| + |I|/2\right) \tag{8.49}$$

The Robertson approximation can also be expressed logically as
If $|I| \leq |Q|$, then $A_{Rob} = |Q| + |I|/2$
else $A_{Rob} = |I| + |Q|/2$

Table 8.16 compares the accuracy performance of both the JPL and Robertson approximations for $A = \sqrt{I^2 + Q^2}$ assuming $A = 1$ (normalized) for one quadrant in 15° increments.

The more accurate JPL approximation (2.8% error worst case at 45°) is typically used during tracking modes while the computationally efficient Robertson approximation (11.6% worst-case errors at 30° and 60°) is typically used during acquisition. Since the Robertson approximation adds quantization noise to A, then the single trial detector threshold, V_t/σ_n, computed from (8.46) must be increased slightly to compensate. For example, [46] taken from [50], uses a correction factor of $(V_t/\sigma_n)_R = 1.08677793\ V_t/\sigma_n$, so $(V_t/\sigma_n)_R = \sqrt{-2.3621724 \ln P_{fa}}$. The determination of the most suitable single trial probability of false alarm, the overall probability of false alarm, and the overall probability of detection is a tuning process.

Table 8.16 Accuracy Comparisons of JPL and Robertson Envelope Approximations

θ (degrees)	$I = A\cos\theta$	$Q = A\sin\theta$	A_{JPL}	Error JPL	Error %	A_{Rob}	Error Robertson	Error %
0	1	0	1	0	0	1	0	0
15	0.965925826	0.258819045	0.9983	0.002	0.2	1.0953	−0.095	−9.5
30	0.866025404	0.5	1.0078	−0.008	−0.8	1.116	−0.116	−11.6
45	0.707106781	0.707106781	0.9723	0.028	2.8	1.0607	−0.061	−6.1
60	0.5	0.866025404	1.0078	−0.008	-0.8	1.116	−0.116	−11.6
75	0.258819045	0.965925826	0.9983	0.002	0.2	1.0953	−0.095	−9.5
90	0	1	1	0	0	1	0	0

8.5.2 Tong Search Detector

The Tong detector is a suboptimal search algorithm that requires an average of only 1.58 longer to make a decision than a maximum likelihood (optimum) search algorithm [51]. A maximum-likelihood search algorithm must search all possible uncertainties, which has already been shown to be practical with current DSP technology for high uncertainty conditions for GNSS PRN signals with relatively short code lengths (but not for extremely long code lengths of military signals such as GPS P(Y) and M code signals that are described in Section 8.5.6). However, these high uncertainty conditions rapidly disappear after the initial GNSS receiver search is successful, so for succeeding acquisitions or reacquisitions, the Tong detector has a reasonable computational burden and is excellent for detecting signals with an expected $(C/N_0)_{dB}$ of 25 dB-Hz or higher with a very low probability of false alarm. Since the Tong detector tends to mush (experience extremely long periods of indecision) under very low $(C/N_0)_{dB}$ conditions, a mush counter must be used to terminate this condition if it is encountered. A scheme that substitutes the mush counter with a fixed dwell time detector is described in Section 8.5.4 along with the recommended search algorithm.

The overall probability of false alarm and the overall probability of detection, respectively, for the Tong detector in [46] (the numerator in [46, equation (16)] for P_D is incorrect) taken from [52] is

$$P_{FA} = \frac{\left(\frac{1-P_{fa}}{P_{fa}}\right)^{B} - 1}{\left(\frac{1-P_{fa}}{P_{fa}}\right)^{A+B-1} - 1} \tag{8.50}$$

$$P_{D} = \frac{\left(\frac{1-P_{d}}{P_{d}}\right)^{B} - 1}{\left(\frac{1-P_{d}}{P_{d}}\right)^{A+B-1} - 1} \tag{8.51}$$

Figure 8.33 in [46] taken from [53] is a plot of (8.51) as a function of input signal-to-noise ratio expressed as $(C/N)_{dB}$ into the Tong detector for $B = 1$ and with A as a running parameter ranging from 2 to 12 with $P_{FA} = 1.0\text{E-}06$.

Referring to Figure 8.33, from left to right, each curve is based on a different single trial detector threshold (i.e., the smaller values of A requiring a larger threshold setting to keep the overall false alarm rate constant). In fact, the main attribute of the Tong detector is that it significantly improves the false alarm rate of the single trial detector for $A > 4$, but somewhat underperforms the probability of detection of the single trial detector at the threshold setting selected for Tong parameter A. Note in Figure 8.33 that increasing A increases the detection sensitivity, but this also decreases search speed. The detection sensitivity is also increased if $B = 2$, but also at the cost of search speed. Selecting these parameters is a tuning process because there are trade-offs between search speed and probability of detection for a desired probability of false alarm. Typical values are $B = 1$ and $A = 12$ for expected low $(C/N_0)_{dB}$ (25 dB-Hz or higher) to $A = 8$ for expected high $(C/N_0)_{dB}$ (39 dB-Hz or higher).

The Tong parameter B is the number of initial false decisions in a row required to dismiss a cell (e.g., if B =1, then if the first decision is false, that cell is immediately dismissed). The Tong parameter A is the number of initial true decisions in a row required to declare the signal present (e.g., if $A = 8$, then if the first eight input decisions are true, the signal is declared present). Mixed input decisions make the Tong decision time variable.

Table 8.17 shows the single trial detector threshold (assuming the Robertson approximation) and the resulting P_{fa} required to keep the Tong overall probability of false alarm constant at $P_{FA} = 1.0\text{E-}06$ using typical values of 1 and 2 for parameter B with Tong parameter A ranging from 2 to 12. Note that the false alarm rate of the single trial detector for values of Tong parameter $A > 4$ is significantly improved and that there is no improvement if $A = 2$.

Figure 8.33 did not use these thresholds because it assumed no computational error in the envelope. The curves in this figure move to the right about 1 dB (i.e., have reduced sensitivity) if the Robertson envelope approximation thresholds of Table 8.17 are used.

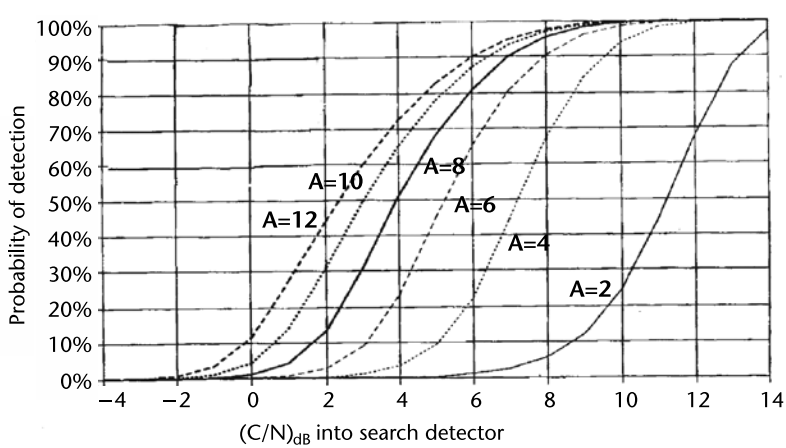


Figure 8.33 Probability of detection for Tong search detector with $P_{FA} = 1.0\text{E-}06$ and B = 1.

Table 8.17 Threshold and Single Trial P_{fa} to Keep Tong P_{FA} = 1.0E-06 Using Robertson Envelope Approximation

	Tong parameter B = 1		*Tong parameter B = 2*	
Tong parameter A	$(V_t/\sigma_n)_R$	P_{fa}	$(V_t/\sigma_n)_R$	P_{fa}
2	5.712671848	1.00E-06	5.712671848	1.00E-06
4	3.300511027	9.94E-03	3.301665179	9.90E-03
6	2.577174394	6.01E-02	2.58253793	5.94E-02
8	2.218994177	1.24E-01	2.227909726	1.22E-01
10	2.008200517	1.81E-01	2.019037457	1.78E-01
12	1.87100656	2.27E-01	1.882636288	2.23E-01

The Tong search speed varies from very fast with no signal present (noise only) to very slow with the signal present under high noise conditions. The mean number of dwell times to dismiss a cell (mean number of dwells per cell) containing noise only is

$$N_n = \frac{1}{1-2P_{fa}} \ \text{(dwells/cell)} \tag{8.52}$$

So for the noise only condition the Tong detector search speed can be estimated using:

$$R_{Tong(noise)} = \frac{C_c}{N_n T} \ \text{(chips/s)} \tag{8.53}$$

where C_c = chips per cell and T = dwell time (s).

For example, using (8.52) with P_{fa} = 16%, the mean number of dwells with no signal present is N_n = 1.47, so using (8.53) for a dwell time of 5 ms, and ½-chip per cell, the Tong search speed with no signal present is $R_{Tong(noise)}$ = 68 chips/s, but this is not representative of the overall average search speed, especially under poor signal to noise ratio conditions. However, it demonstrates that the Tong detector searches very fast when there is no signal present.

8.5.3 M of N Search Detector

The second example of a search algorithm is a fixed interval detector called the M of N search detector. The M of N search detector takes N envelopes and compares them to the threshold for each cell. If M or more of them exceeds the threshold, then the signal is declared present. If not, the signal is declared absent and the process is repeated for the next cell in the search pattern. These are treated as Bernoulli trials and the number of envelopes, n, that exceed the threshold has a Binomial distribution.

The overall probability of false alarm in *N* trials in [46] taken from [54] is

$$P_{FA} = \sum_{n=M}^{N} \binom{N}{n} P_{fa}^{n} \left(1 - P_{fa}\right)^{N-n} = 1 - \sum_{n=0}^{M-1} \binom{N}{n} P_{fa}^{n} \left(1 - P_{fa}\right)^{N-n}$$
$$= 1 - B\left(M-1; N, P_{fa}\right) \tag{8.54}$$

where $B(k;N,p)$ is the cumulative probability density function. The overall probability of detection in N trials from [46] taken from [54] is

$$P_{D} = \sum_{n=M}^{N} \binom{N}{n} P_{d}^{n} \left(1 - P_{fa}\right)^{N-n} = 1 - B\left(M-1; N, P_{d}\right) \tag{8.55}$$

Figure 8.34 from [46] taken from [54] illustrates the M of N probability of detection versus $(C/N)_{dB}$ into the detector for $N = 8$ and $M = 3$, 4, 5, and 6 when $P_{FA} = 1 \times 10^{-6}$.

By inspection in Figure 8.34, it is clear that $M = 5$ is the optimum value (i.e., a simple majority criteria of $M = N/2 + 1$). The data were generated by computing P_{fa} given M, N and P_{FA} = 1.0E-06 using the following equation from [46] taken from [54]

$$P_{fa} = B^{-1}\left(M-1; N, 1 - P_{FA}\right) \tag{8.56}$$

This value for P_{fa} is used in (8.46) to determine the single trial detector threshold, V_t/σ_n, that is normalized and used as the upper limit of the integral at the bottom equation in (8.47). This can be integrated in discrete increments of the random variable, z, by forming Δz between each increment out to the threshold value (z =

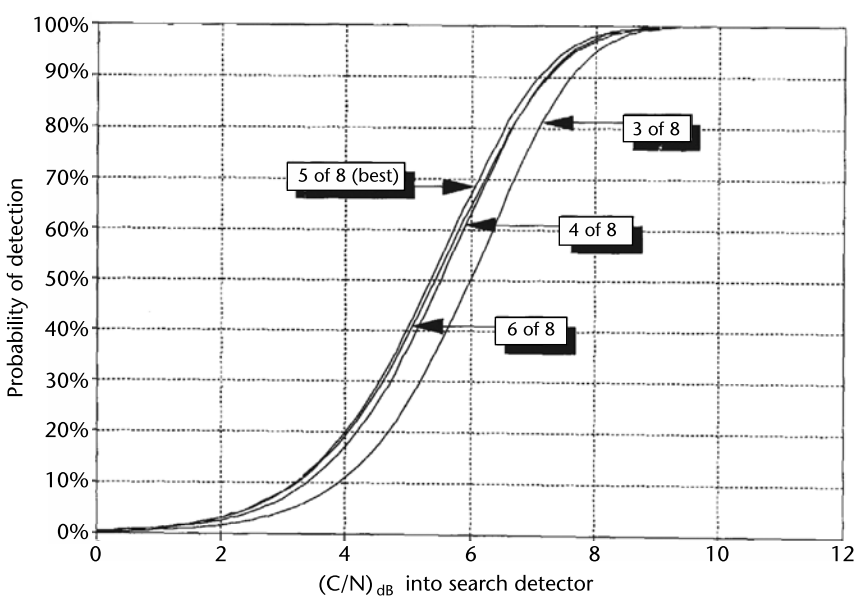


Figure 8.34 Probability of detection for M of N search detector with P_{FA} = 1E-06 and N = 8.

V_t) so that $p_D \approx 1 - \sum_{z=0}^{z=V_t} p_S(z)\Delta z$. This is performed for each value of C/N as determined from $(C/N)_{dB}$ used in the abscissa and each P_D is determined using (8.55) for every point of the abscissa. This computation sequence must be repeated for each combination of M and N using the new threshold that holds the overall probability of false alarm constant. The accuracy of the plots is extremely sensitive to the resolution of P_{fa} and Δz even though the actual thresholds cannot be maintained to several decimal places.

Table 8.18 tabulates the single trial threshold (assuming the Robertson approximation is used for the envelope computation) and its corresponding probability of false alarm that keeps the M of N detector overall probability of false alarm at 1.0E-06 for a multiplicity of values of N and assuming a simple majority vote ($M = N/2 + 1$). For the simple majority criteria note that these thresholds are higher than for the Tong detector when the fixed number of trials N compares in size with the Tong parameter A.

Figure 8.34 did not use these thresholds because it assumed no computational error in the envelope and there is a small loss in sensitivity when the Robertson approximation is used.

Under all signal conditions of the M of N detector there is a fixed number of dwells per cell, N, so the search speed is

$$R_{MofN} = \frac{C_c}{NT} \text{ (chips/s)} \tag{8.57}$$

For example, assuming $N = 8$ dwells per cell, $C_c = ½$-chip per cell and $T = 5$ ms per dwell, $R_S = 12.5$ chips/s. For the noise only condition, this is more than 5 times slower than the Tong search speed for the Tong $P_{fa} = 16\%$ example. When there is a high C/N signal present, the Tong detector slows down to about the same speed as the M of N detector if the Tong parameter A is the same as N, but if there is a low C/N signal present, the Tong detector takes much longer to make a decision than the M of N detector and, as stated earlier, can actually enter a mush condition where it becomes indecisive. For this reason, a combination of the two detectors is presented in the following section.

Table 8.18 Threshold and Single Trial P_{fa} to Keep M of N Detector P_{FA} = 1E-06 Using Robertson Envelope Approximation

N	M	$(V_t/\sigma_n)_R$	P_{fa}
8	5	2.89731174	0.028619
10	6	2.73614506	0.042032
12	7	2.613541817	0.055484
14	8	2.516464178	0.068506
16	9	2.437256335	0.080885
18	10	2.37111023	0.092543
20	11	2.314834153	0.103473

8.5.4 Combined Tong and M of N Search Detectors

Figure 8.35 illustrates the combined Tong and M of N search detector algorithm [28]. There is an ideal synergism using the combination of the two search detectors: the Tong detector searches faster in the presence of noise only and improves the probability of false alarm more efficiently than the M of N detector, but the M of N detector never takes longer than the prescribed number of dwells in the same cell (i.e., it does not mush).

The threshold design is always based on the superior Tong detector, so the *M* of *N* detector design must be based on the threshold chosen for the Tong detector. For example, if the Robertson approximation is used and the Tong parameters are B =1 and A = 12, then Table 8.17 says that V_t/σ_n= 1.87100656 to keep the Tong detector P_{FA} at 1.0E-06, but Table 8.18 shows no example of an *M* of *N* detector with a threshold that can maintain this false alarm probability. A mush counter of 20 is typical for this Tong detector design, so choosing $N = 20$ and $M = 15$ for the *M* of *N* design closely matches the Tong false alarm probability and also gives the Tong detector ample opportunity to be the primary decision maker. (Note that using the typical majority vote criteria for this *M* of *N* detector would not provide the desired false alarm performance.)

Using the above values as a case example, the operation of Figure 8.35 is described as follows. Three variables are initialized before the algorithm begins: The Tong counter K_t is set equal to the Tong value for B (assume the typical value $K_t = B = 1$ for this example); the *M* of *N* counter, K_m, is set to $N = 20$, and the *M* of *N* index, I, is set to 0. The stored constants for this example are: Tong parameter $A = 12$, *M* of *N* decision parameter $M = 15$. The operation begins with the single trial detector decision outcome for one dwell time in one search cell. If the envelope amplitude, ENV_k, is greater than threshold, V_t/σ_n, then K_t is incremented by 1 and I is incremented by 1; if not, then K_t is decremented by 1. (Note that K_m is decremented by 1 for either decision.) Then up to 4 decisions are made: if K_t is zero, the signal is declared not present; if not, then if K_t has reached $A = 12$, the signal is

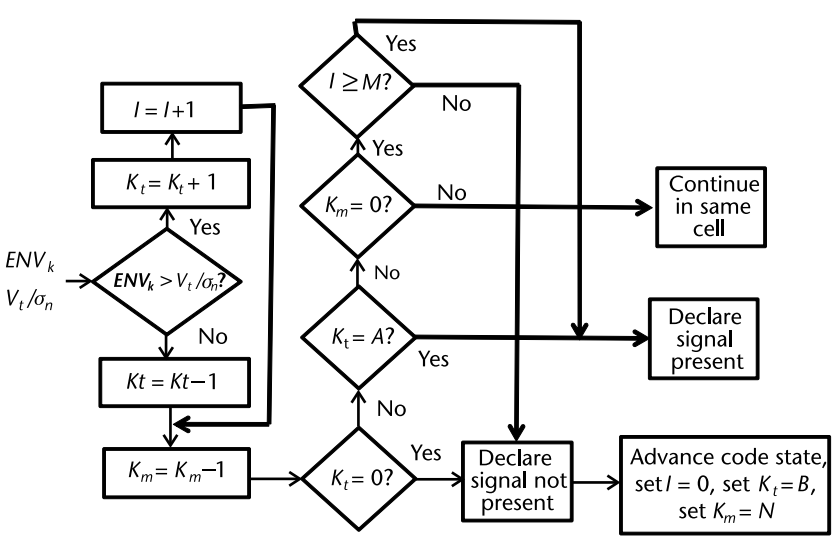


Figure 8.35 Hybrid Tong search detector with M of N detector as mush counter and N > A.

declared present; if not, then if K_m has not been decremented to 0, another single trial detector decision is requested in the same cell; if $K_m = 0$, then the M of N detector has taken over and the signal is declared not present if I is not greater than or equal to M; otherwise, the signal is declared present. When the signal is declared not present, the code state is advanced to a new cell and the detector is initialized again. When the signal is declared present, a peak search is performed (described in Section 8.5.7) before loop closure is performed.

8.5.5 FFT-Based Techniques

Early FFT-based search techniques based on the GPS L1 C/A code (and their limitations) were introduced in Section 8.4.3.3, but a computationally efficient technique is described next that embodies features that should be considered for FFT-based search engines designed for all GNSS signals.

8.5.4.1 Computationally Efficient FFT Acquisition Scheme

A block processing acquisition scheme using a parallel code technique modified for minimum computation is shown in Figure 8.36. Note that the complex baseband IF input scheme that performs the carrier wipe-off function in Figure 8.36 could be replaced by the real IF input scheme shown in Figure 8.14 because this also produces the required baseband I and Q signals after carrier wipe-off, albeit with some high-frequency components that are filtered out prior to the final detection process. This scheme is adapted from [41] and was specifically designed as an FFT-based acquisition algorithm for the GPS L5 signal, but the technique can be applied to

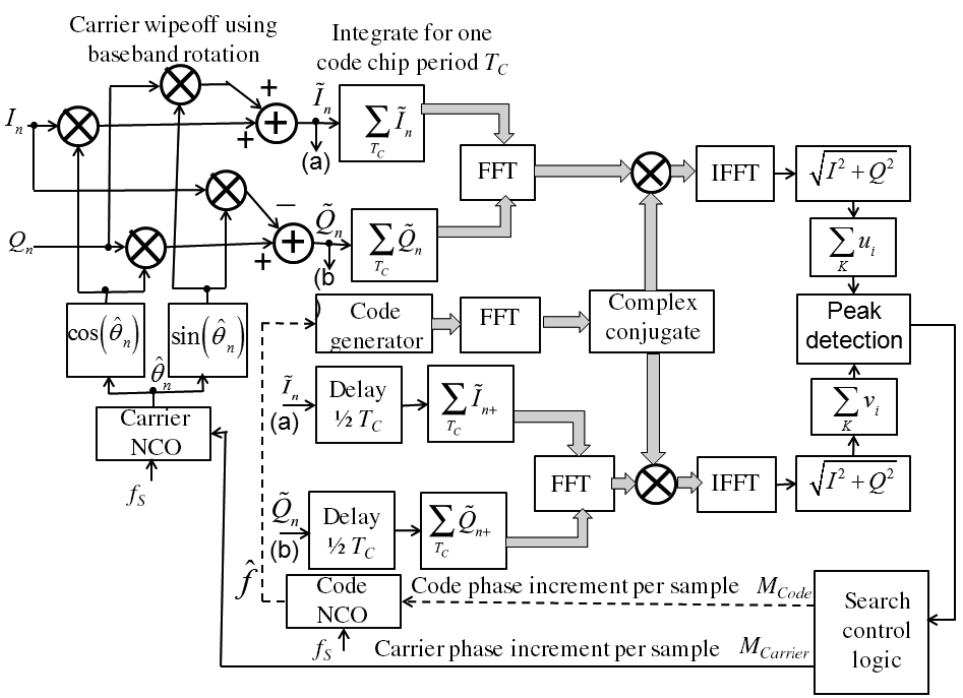


Figure 8.36 Block processing acquisition using parallel code technique modified for minimum computations.

any FFT-based acquisition scheme with care taken to avoid correlation ambiguity in GNSS signals with split spectrum.

Referring to Figure 8.36, the block processing begins after the carrier wipe-off process signified by the block arrows for the parallel signal flow (i.e., real-time processing is taking place up to the block processing point). Here, blocks of real-time samples are processed in the frequency domain much faster than real time and must be fast enough that block processing of one block of data is completed before the next block of real time data is ready (i.e., has been stored in memory). The serial to block operation typically uses a double-buffered memory so that one block can be stored in real time while the second block is being processed faster than real time.

The key to the computational efficiency of this scheme is the parallel operation of integrating $\tilde{I}_n$ and $\tilde{Q}_n$ over one chip period, T_C (shown in the upper left of Figure 8.36), and the same integration over T_C of $\tilde{I}_n$ and $\tilde{Q}_n$ [labeled (a) and (b) in lower left of Figure 8.36] after being delayed by ½-chip. This supports the optimum ½-chip search pattern in the code dimension. The FFT of these two complex signals are taken and separately multiplied by the complex conjugate of the FFT of the replica code generator to account for all possible phase states with respect to the incoming signal. The length of the FFT (number of integrated code periods) depends on the number of chips chosen for the coherent integration time of the correlated outcome. In the case of the Q5 pilot signal of [41] the coherent integration time was 20 ms because it has a 10,230 code chip period in 1 ms and the Neumann-Hofman Q5 overlay period is 20 ms (and there are no data transition boundaries), so the FFT length is a minimum of $10{,}230 \times 20 = 204{,}600$ if the FFT algorithm does not require a power of 2 length or 262,144 (with zero padding of 57,544) if a radix of 2 is required.

The key to achieving maximum computation efficiency is the use of precomputed complex conjugate functions of all FFTs of one period of the replica PRN replica code including the effect of any overlay code (see Table 8.14 for open service GNSS overlay code lengths). The replica PRN code is selected from the precomputation of all possible PRN codes for that signal. However, this consumes a massive amount of memory and these are typically not Doppler compensated replica codes that require an even greater increase the memory storage requirement. If precomputed codes are used then the dotted line functions including code NCO, code generator, FFT, and complex conjugate functions shown in Figure 8.36 would not be activated during FFT acquisition. Instead, the complex conjugate function would perform table look-ups of precomputed complex conjugates for each block required by the current real-time Doppler bin search being conducted by the search control logic. If there is sufficient processing power to compute those blocks using the Doppler compensated code NCO at the same real-time operation as the Doppler compensated carrier NCO (and the subsequent real time carrier wipe-off), then the dotted line functions would be activated to provide nearly perfect complex conjugate functions and the least amount of memory storage, but at the cost of decreased frequency-domain computation efficiency.

Resuming the functional description, the Doppler-shifted samples, $\tilde{I}_n$ and $\tilde{Q}_n$, are accumulated for each chip time, T_C, to form a sequence of incoming code phases over the full code period and the complex FFT of this sequence is complex-multiplied with the complex-conjugate FFT of the local replica code. The same process is also performed for the lower ½-chip delayed counterpart. The correlation sequence

for the predefined FFT length is obtained by taking the inverse FFT (IFFT) in both the upper and lower output correlation signal paths that result from the right circular shift of the code generator replica PRN sequence. The envelopes (e.g., Robertson amplitudes) of the upper and lower complex correlation sequences, $\sqrt{I^2 + Q^2}$, are stored as vectors $\mathbf{u}_i$ and $\mathbf{v}_i$, respectively, along with their corresponding replica code phase index. (In [41] the upper and lower square magnitudes of the 20-ms coherent correlation sequences are stored as $\mathbf{u}_i$ and $\mathbf{v}_i$.) As shown in Figure 8.36, these vectors may be noncoherently integrated K times to improve the signal to noise ratio. These vectors are block searched for the maximum correlation value and its corresponding code phase index. If the maximum correlation value exceeds the predetermined acquisition threshold value, then the corresponding code phase index is used to conduct a peak search that aligns the replica code phase with the incoming code phase.

The ADC sampling frequency, f_S, plays an important role in this scheme because of the ½-chip delay design requirement that an integer number of samples be equal to ½-chip. However, specifying an exact value results in commensurate sampling that disobeys the rule for ADC sampling described in Section 8.3.7 and criticized in the Figure 8.29(b) parallel code circular correlation FFT design presented in Section 8.4.3.3 with Figures 8.30 and 8.31 depicting the adverse correlation effect of commensurate sampling. Consider the three case examples of ADC designs for the L5 signal provided in Section 8.3.7 (baseband f_S = 34 MHz) and Section 8.3.8 (140-MHz IF f_S = 112 MHz and an antialias SAW filter version of the 140-MHz IF design with f_S = 62.22 MHz). Only the 62.22-MHz sample rate design is satisfactory as is because it provides 6.082 samples per chip and an almost ideal ½-chip delay using a 3-sample delay (the integer value of 3.041 samples). The baseband design at 34-MHz sample rate would have to be refined upward to provide a 44-MHz sample rate that would provide 4.008 samples per chip with the integer value of 2.004 samples providing the ½-chip delay. The 140-MHz IF design at 112-MHz sample rate is not satisfactory for this FFT design, so the 62.22-MHz sample rate design is the best solution if the 140-MHz IF remains the same because the sampling frequency rules are based on the chosen IF.

Reference [41] provided considerably more insight into the acquisition performance of this FFT technique as well as of conventional acquisition techniques in the context of the GPS L5 signal. For example, this FFT technique successfully acquired the L5 Q5 (pilot) signal received with a $(C/N_0)_{dB}$ = 25 dB-Hz using FFT length T = 20 ms and noncoherent integrations K = 25, requiring 0.5-ms dwell time for each Doppler bin search increment of 25 Hz.

8.5.6 Direct Acquisition of GPS Military Signals

In all cases of direct-P(Y) or direct-M military signal acquisition, it is impossible to search all the uncertainty in the code length (as is the case for typical commercial GNSS receivers) because the encrypted codes are infinitely long. The acquisition uncertainty in the Doppler dimension is the same as for the commercial GPS receiver that is designed to operate with the same expected maximum user velocity. Therefore, given that the modern military receiver has been designed to retain (or obtain) GPS time of week to typically much less than 1 second, the code range dimension uncertainty is defined by the receiver time uncertainty (converted into code range)

plus the maximum range change between the user antenna and the SV for any GPS SV in view. Review of the equations associated with Figure 8.21 for the GPS orbit translates that range uncertainty to about 20 ms (converted into code range).

For a sky search (cold start) acquisition example, first assume ideally that the user GPS time estimate is perfect and there is valid almanac data for the GPS SVs. The receiver cannot use the almanac to locate visible GPS SVs initially because it is missing an essential parameter: a rough estimate of its own location. So a typical initial assumption is that the receiver is located at the center of the Earth and chooses the first SV to search essentially at random, but chooses the following SVs in a sequence based on the distribution of SVs at the acquisition time. It performs a sky search over the total Doppler uncertainty plus the code range uncertainty based on its estimate of true GPS time. Since the SV is transmitting a PRN code that for this analysis can be assumed to be perfectly aligned with true GPS time, the receiver always searches early to late, so it starts with the replica code set to the SV transmit time as if was at zenith (zero SV Doppler and closest approach) based on its perfect estimate of GPS time as the transmit time. Then it searches out to the replica code corresponding to the SV transmit time as if it was at a lower elevation angle of interest (farthest approach). That would correspond to a range uncertainty of less than 20 ms assuming that very low-elevation SVs are undesirable initially and statistically most SVs are in the mid-elevation angle regions.

The next example assumes the reality that there is estimated GPS time uncertainly, so half of the total GPS time uncertainty (converted to code range) is subtracted from the first assumption of initial replica code setting and the search range starts there and continues through the first example code search range (less than 20 ms) plus the positive half of the total GPS time uncertainly (converted to code range). After the first SV has been acquired, the GPS time uncertainty immediately reduces to less than 20 ms that, in turn, reduces the total code range uncertainty to less than 40 ms. So finding the first SV is the most time-consuming portion of the sky search. At this point, the user location is assumed to be on the surface of the Earth (unless there is an independent source of user altitude available) and directly under the first SV successfully acquired. This enables a very coarse use of almanac data to select the (apparent) highest-elevation SV for the second search. When the second SV is found, the user location is assumed to be at the midpoint of the two points on (or above) the Earth defined by the two SV to Earth-center vectors. With 3 SVs, an altitude-hold three-dimensional solution is now possible that provides a much lower user position and time bias uncertainty, resulting in much better selection of visible SVs. In this manner, the receiver eventually bootstraps itself into a significantly reduced PVT uncertainty along with very small code range and carrier Doppler uncertainties. After the first 4 SVs have been acquired and measurements have been incorporated by the navigation function, the FFT search engine should not be needed for subsequent acquisitions because the uncertainties are so small. Modern military FFT search engine technology readily supports 1 second or more of time uncertainty with fairly rapid direct acquisition times even under interference conditions. The cold start search pattern also takes into consideration the most likely GPS SVs to be in view as the receiver reduces uncertainty with SV acquisitions.

The Military GPS User Equipment (MGUE) program has developed M-code-capable GPS receivers that are mandated by Congress after fiscal year 2017. These

MGUE receivers are not only remarkably lower in size, weight, and power consumption but also significantly more robust in acquisition and tracking of the more secure and powerful military signals with significantly increased operational reliability and accuracy. A specific requirement of every class of these modernized MGUE receivers is the ability to perform direct acquisition of the M code (direct-M acquisition). Direct-M acquisition means the ability to acquire the M code of visible GPS SVs directly in a sky-search mode without assistance from any other signal in space and with only a coarse knowledge of GPS time. The major direct-M acquisition requirement difference between classes of MGUE receivers is the minimum signal-to-noise ratio of the received M code signal during acquisition, with maximum dynamic stress a close second. The level of GNSS receiver sophistication to achieve higher levels of robustness in the presence of interference and spoofing are described in Chapter 9. These sophisticated but first principal techniques apply to MGUE receivers but are typically implemented using the highest state-of-the-art techniques because the potential operating environments are much harsher than for the majority of commercial GNSS applications.

Since there is a transition period in the GPS control segment, space segment, and (military) user segment between the original L1 and L2 P(Y) code military signals and the full capability of the modernized L1 and L2 M code signals, the current generation of MGUE receivers have dual P(Y) and M code capability, including the requirement for direct-P(Y) signal acquisition. While P(Y) code was originally designed for acquisition through C/A code, the concept of direct-P(Y) acquisition predates the M code concept using massive parallel correlators. Modernized FFT search engines in military GPS receivers have only recently evolved in MGUE technology after they were first demonstrated commercially with the very short code length of the GPS C/A code. Prior to FFT capability with the military codes, the use of massively parallel correlators was feasible for search uncertainty conditions where the prediction of satellite transmit time required less search time to acquire the P(Y) code by direct sequence than to perform a C/A code search and handover. However, the primary motivation for direct-P(Y) acquisition is that it can acquire the GPS signals under higher interference conditions than for C/A code. Reference [55] described a typical massive correlator architecture that supports rapid direct P(Y) code acquisition in the presence of jamming.

In contrast to the P(Y)-code signal, the M-code signal was designed so that direct acquisition would be the primary means of acquisition, drawing on advances in acquisition algorithms and integrated circuit technology. The BOC(10,5) modulation allows separate acquisition processing on upper and lower sidebands, with processing at the 5.115-MHz spreading code chip rate, and noncoherent integration of the results from the two sidebands, as illustrated in Figure 8.37 [56]. Note that the two sidebands may be selected and processed at the digital baseband part of the receiver rather than by two L-band downconverters and two ADCs. For example, if the 140-MHz IF undersampled L5 ADC design described in Section 8.3.8 was adapted for M code using a 30-MHz SAW bandwidth, but still undersampled in NZ(5) with $f_S = 62.22$ MHz, this would place the M-code center frequency in NZ(1) at $0.25f_S = 15.555$ MHz. The lower sideband would be selected with a carrier wipe-off signal of $15.555 - 10.23 = 5.325$ MHz ± Doppler and the upper sideband would be selected with a carrier wipe-off signal of $15.555 + 10.23 = 25.785$ MHz ± Doppler. Also note that the replica M code does not contain the biphase

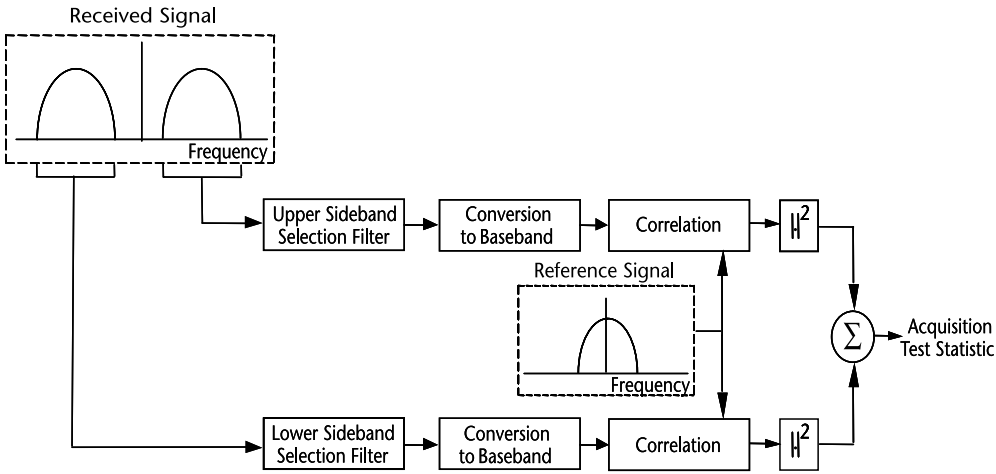


Figure 8.37 Sideband acquisition processing of the M-code signal.

modulation signal, so there is no ambiguity in the correlation pattern [i.e., the two sidebands can be searched in the same manner as the P(Y)-code BPSK signal]. This approach suffers only a fraction of a decibel in performance compared to coherent processing of both sidebands.

Interestingly, when the sideband acquisition processing approach is used, M-code signal direct acquisition processing uses approximately half the arithmetic operations and half the storage of Y-code signal direct acquisition processing [57]. An integrated circuit based on this processing approach demonstrated direct-M acquisition even with relatively large initial time uncertainties in significant levels of jamming [58]. The processing architecture is based on computation of short-time correlations, followed by FFT backend processing for parallel search of multiple frequency values.

Acquisition in jamming requires long integration times. Coherent integration times are limited by data bit boundaries, oscillator stability, and dynamics. They also lead to narrow Doppler bins. Consequently, a large number of noncoherent integrations are employed in the presence of jamming. Detection performance is readily predicted using standard theory. The output signal-to-noise-plus interference ratio (SNIR) after a cross-correlation is given by

$$\rho_o = \frac{T}{L}\frac{0.25C}{N_0 + J_0} \tag{8.58}$$

where T is the coherent integration time used in the correlations, L is the implementation loss expressed as a number greater than or equal to unity, C is the received signal power, the factor of 0.25 accounts for splitting the received signal power into four distinct segments (upper and lower sidebands, even and odd spreading symbols) in each coherent integration time, N_0 is the power spectral density of the thermal noise at the receiver front end, and J_0 is the effective power spectral density of the received jamming signal.

The detection probability is found using the generalized Marcum Q function. Using the notation $P_N(X, Y)$ [58] as the probability that the random variable with $2N$ degrees of freedom and SNIR of X exceeds threshold value of Y allows the detection probability to be expressed as

$$P_d = P_{4N_n}\left(\rho_o, V_t\right) \tag{8.59}$$

where N_n is the number of coherent integrations times used and V_t is the detection threshold calculated to provide the needed false alarm probability for the given number of noncoherent integrations. The 4-subscript notation multiplied by the N_n subscript notation in (8.59) accounts for the fact that the number of complex quantities being noncoherently combined is four times the number of coherent integration times used, reflecting the combination of upper and lower sidebands and even and odd spreading symbols.

The expressions (8.58) and (8.59) can be used to determine the number of coherent integration times needed to achieve a specific detection probability at a given false alarm probability.

The time (in seconds) to search the initial time uncertainty of $\pm\Delta$ seconds and an initial frequency uncertainty of $\pm\Phi$ Hz is then

$$T_{\text{search}} = N_n \left\lceil \frac{\Delta}{T} \right\rceil \left\lceil \frac{\Phi T}{N_{\text{STC}}} \right\rceil \tag{8.60}$$

where T is the coherent integration time, N_{STC}, is the number of short-time correlations within the coherent integration time, and the notation $\lceil x \rceil$ means the smallest integer greater than x.

Alternatively, the M code can be digitized conventionally at baseband or at IF and have the ambiguity removed by two different baseband techniques. The first technique converts the incoming baseband signal to a BPSK signal by multiplying it with the replica Doppler compensated spreading code square wave (that is coherent with the prompt replica code) before carrier wipe-off and then performing code wipe-off using the unspread replica M code in a conventional BPSK manner [59]. This coherent technique is satisfactory during signal acquisition, but it loses the code tracking precision and measurement accuracy of conventional M-code tracking.

The second technique computes both the in-phase and quadra-phase replica BOC M Code that after correlation with the incoming signal produces both the conventional BOC code correlation and a quadra-phase correlation that can be combined during acquisition to remove the ambiguity [60–62]. Figure 8.38 illustrates how these replicas are synthesized [62]. Note that there are five conventional in-phase BOC (B) replicas: early (E_B), narrow early (E_{BN}), prompt (P_B), narrow late (L_{BN}), and late (L_B). (The original M code tracking technique using the "bump-jump" code tracking technique calls these phases very early, early, prompt, late, and very late, respectively.) The shift register design provides 1/16th of a chip code phase increments that supports the nearly optimum 1/8th-chip separation between the narrow early and the narrow late code loop correlators for generating the conventional early-minus-late M code tracking error. There are also six similar named

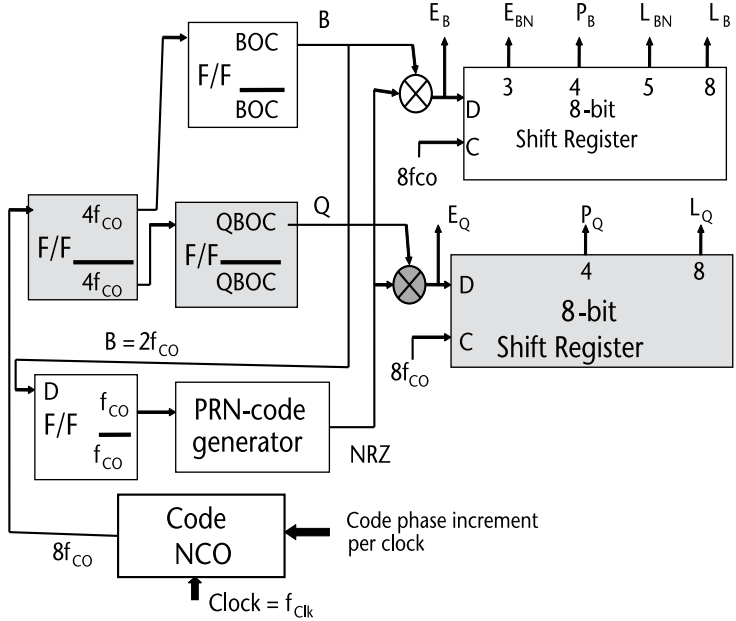


Figure 8.38 Synthesis of conventional (in-phase) BOC plus (shaded) quadra-phase BOC replica signals.

quadra-phase (Q) replicas synthesized: E_Q, E_{QN}, P_Q, L_{QN}, and L_Q, but typically only E_Q, P_Q, and L_Q are used to remove the M code BOC ambiguity. Figure 8.39 depicts the BOC and QBOC timing diagram for the replica code synthesis assuming zero Doppler. Figure 8.40 shows the early code wipe-off (following carrier wipe-off) using E_B and E_Q. The carrier wipe-off is performed only once, but this code wipe-off process in pairs is repeated using P_B, P_Q and L_B, L_Q and singularly for E_{BN} and L_{BN}. Each wipe-off produces an in-phase and quadra-phase output for a total of 16 separate signals that are integrated and dumped. The result of using a multimodal BOC signal without modification is called a multimodal BOC envelope (MBE) and the result of reconstructing an MBE into a unimodal correlation function to resolve the ambiguities is called a unimodal BOC envelope (UBE). For example, the early unimodal BOC envelope is formed by $E_{UBE} = \sqrt{I_{EB}^2 + I_{EQ}^2 + Q_{EB}^2 + Q_{EQ}^2}$ using the in-phase and quadra-phase signals produced by the E_B and E_Q code wipe-off. Figure 8.41 illustrates the cross-correlation powers that result from the correlation of the in-phase replica M code and the quadra-phase replica M code with the incoming M code signal as a function of replica code offset. The integrate-and-dump process following code wipe-off provides smoothing of this cross-correlation process. Figure 8.42 shows this filtering effect on the UBE of the M code versus replica code offset with the incoming M code signal. This produces the desired unambiguous M code correlation during the search process.

During tracking modes, the code loop is pulled in using the UBEs of the early and late signals, then transitions to high precision conventional M code correlation for tracking using either the traditional bump-jump technique or the unambiguous-aided ambiguous code-tracking scheme described in [61].

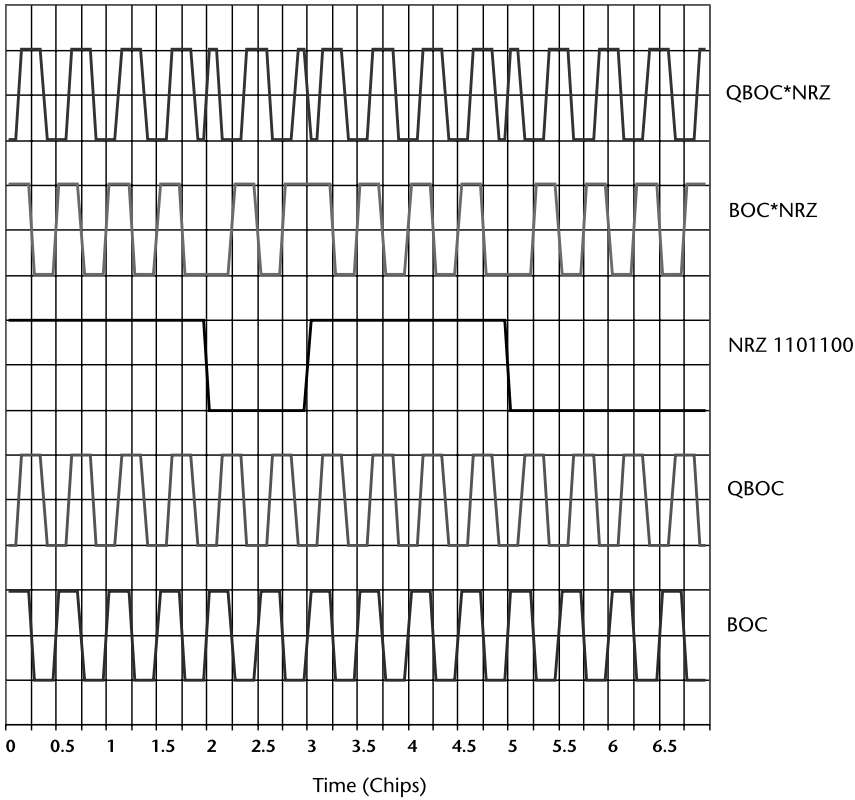


Figure 8.39 BOC and QBOC replica code synthesis timing diagram for zero Doppler.

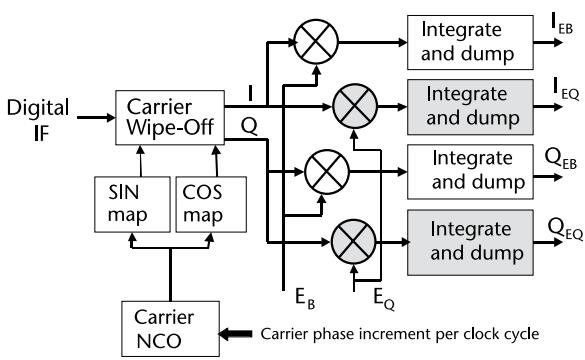


Figure 8.40 Code wipe-off of conventional (in-phase) Early BOC and (shaded) quadra-phase early BOC replica signals.

8.5.7 Vernier Doppler and Peak Code Search

When any search process terminates with success at finding the signal, the precision of the replica carrier and code estimates can be too coarse for immediate tracking loop closure, so a Vernier Doppler search process that reduces the Doppler estimate uncertainty is performed followed by a peak code search process that reduces the code phase estimate uncertainty. These refining processes are typically combined with the search process in the manner shown in the flow diagram of Figure 8.43.

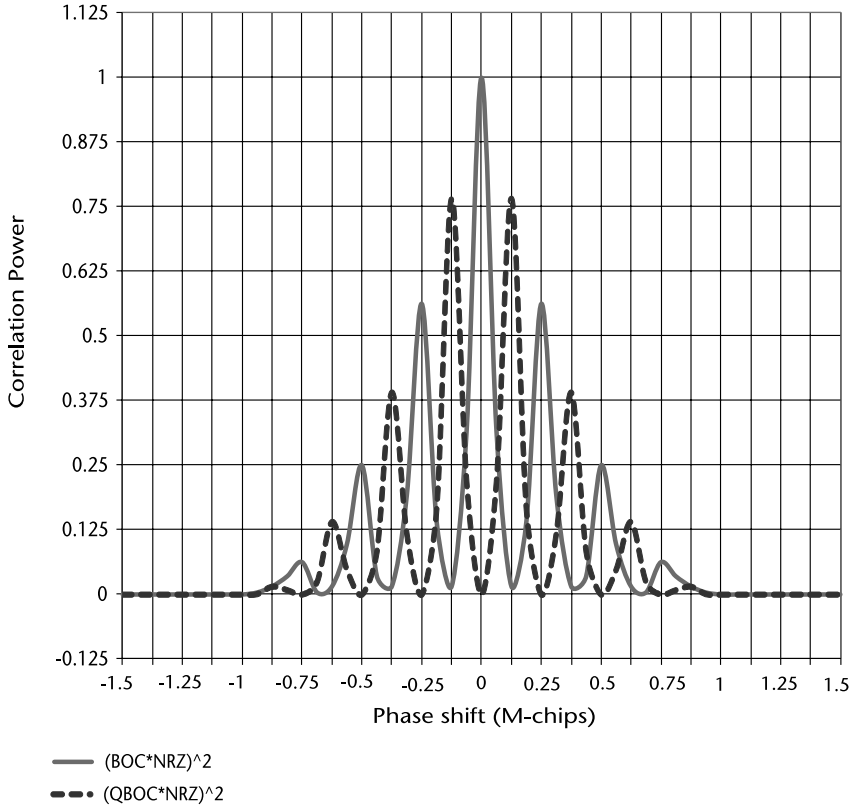


Figure 8.41 Cross-correlation powers of M code and quadra-M code.

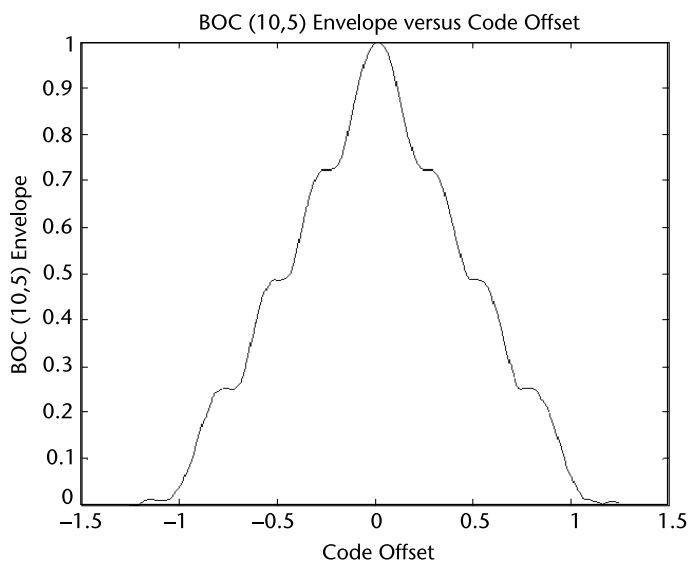


Figure 8.42 Filtered unimodal BOC envelope of M code versus replica code offset.

The flow diagram accommodates multiple target values of C/N_0 based on situational awareness as well as the ability to change values of uncertainty based on the navigation state with the end objective of successful code and carrier loop closure.

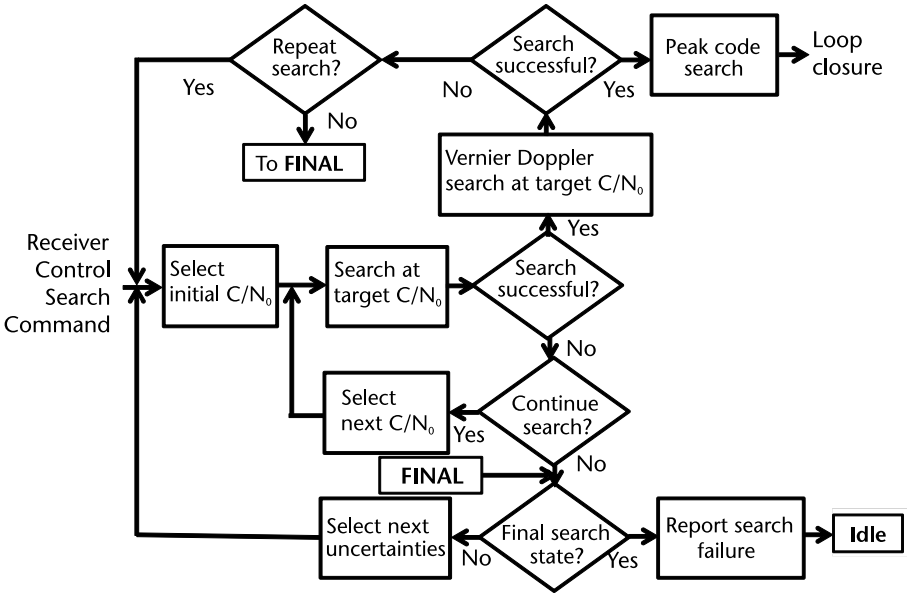


Figure 8.43 Receiver channel search logic including Vernier Doppler and peak code search.

Vernier Doppler search refines the carrier Doppler frequency uncertainty by using a hybrid Tong search to perform multiple passes with different integration times that depend on the integration time used by the preceding search. Receiver control passes these parameters to this process based on situational awareness, the PRN code being searched and its own state knowledge, including the preceding search uncertainties, the hybrid Tong resources available to the Vernier process and the target carrier loop filter frequency pull-in range. These are extremely fast searches since the replica code uncertainty range including the amount of phase change taking place during the total Vernier process is very small.

The peak code search locates the peak of the signal to a higher code chip resolution using the Vernier Doppler frequency (that in some cases also includes an estimate of the acceleration and jerk in the Vernier Doppler). It calculates and compares envelopes from several adjacent correlators, assumes the largest envelope corresponds to the prompt code, compares that to a threshold to ensure that the signal is present and, if successful, uses that code phase as the prompt signal during loop closure. Since the adjacent correlator spacing may be as coarse as the original search ½-chip correlation spacing, two sets of envelopes can be collected with the second set shifted in code phase by ¼-chip, and the largest envelope that exceeds the threshold is used as the prompt signal.

8.6 Carrier Tracking

The only parts of the carrier-tracking loop that were not described in Section 8.4 were the carrier discriminator and the carrier loop filter. Carrier tracking presents a paradox to the GNSS receiver designer. Designer A chooses to select the predetection integration time, the carrier discriminator and carrier loop filter (that completely characterizes the carrier tracking loop in a data modulated channel) so that

the carrier tracking will tolerate the specified maximum dynamic stress by choosing a short predetection integration time, and FLL discriminator feeding a wide noise bandwidth carrier loop filter. Designer B chooses to select those same parameters so that the carrier-tracking measurements meet the specified accuracy by choosing a long predetection integration time and a Costas PLL discriminator feeding a narrow noise bandwidth carrier loop filter. The paradox is that Designer A does not meet the accuracy specification and Designer B does not meet the dynamic stress specification. In practice, some design enhancements (like an inertial measurement unit to provide velocity aiding to the carrier tracking loop) or design innovations must be incorporated to resolve this paradox. A well-designed GNSS receiver channel should close its carrier tracking loop with a short predetection integration time, using an FLL and a wideband FLL loop filter. This design comparison assumes there is data modulation on the carrier so the loop closure process should systematically transition into a Costas PLL gradually adjusting the predetection integration time equal to the period of the data transitions while also gradually adjusting the carrier tracking loop bandwidth as narrow as the maximum anticipated dynamics permit. If the signal is a pilot signal (no data modulation), it should transition into a pure PLL (that is theoretically unconcerned about predetection integration time) and gradually adjust the carrier-tracking loop bandwidth as narrow as the maximum anticipated dynamics permit. Later, an FLL-assisted-PLL carrier-tracking loop will be described that automatically adapts to dynamic stress. With this added insight into carrier-tracking loops, the types and designs of carrier loop discriminators are described next. The loop filter design is described in Section 8.8.

8.6.1 Carrier Loop Discriminator

The carrier loop discriminator algorithm is always implemented using the prompt (i.e., on time) correlator I and Q signals. The algorithm used defines the type of tracking loop as a phase lock loop (PLL), a Costas PLL (which is a PLL-type discriminator that tolerates the presence of data modulation on the baseband signal), or a frequency lock loop (FLL). The PLL and the Costas PLL are the most accurate but are more sensitive to dynamic stress than the FLL that can be very robust in the presence of dynamic stress. The PLL discriminators achieve a 6-dB tracking threshold improvement in comparison with Costas PLL discriminators because the PLL dataless carrier permits tracking the full 360° (four-quadrant) range of the input signal while the Costas PLL can only operate over a 180° (two-quadrant) range because of the presence of data transitions. The PLL and Costas loop discriminators produce phase errors at their outputs. As a result, PLLs replicate almost the exact phase and frequency of the incoming SV carrier (converted to IF or baseband) to perform the carrier wipe-off function. The FLL discriminator produces a frequency error at its output. Because of this, there is also a difference in the architecture of the loop filter, described later. FLLs replicate almost the exact frequency of the incoming SV carrier to perform the carrier wipe-off function. For this reason, they are also called automatic frequency control (AFC) loops. Any frequency error in the replica signal causes the phase to rotate with respect to the incoming carrier signal but the FLL/AFC discriminator senses this and attempts to correct it in the feedback path.

8.6.1.1 PLL Discriminators

The PLL discriminator is used to track the prompt signal of a pilot (dataless) channel in phase lock. Table 8.19 describes two PLL discriminator algorithms, their output phase errors, and their characteristics. The four-quadrant arctangent (ATAN2) PLL discriminator algorithm is a maximum likelihood estimator, but the PLL approximation algorithm using the prompt Q signal normalized by a long term average of the prompt envelope has been proven experimentally to slightly outperform the theoretically optimal and more complex ATAN2 function. Figure 8.44(a) compares the phase error outputs of these two PLL discriminators assuming no noise in the prompt I and Q signals. Note that the ATAN2 discriminator is the only one that remains linear over the full input error range of ±180°. However, in the presence of noise, both of the discriminator outputs are linear only near the 0° region.

8.6.1.2 Costas PLL Discriminators

Any carrier loop that is insensitive to the presence of data modulation is usually called a Costas loop since Costas was the inventor of the first analog PLL discriminator that tolerated data modulation. Table 8.20 describes four Costas PLL discriminator algorithms, their output phase errors and their characteristics. Figure 8.44(b) compares the phase error outputs of these four Costas PLL discriminators assuming no noise in the prompt I and Q signals. As shown, the two-quadrant ATAN Costas discriminator of Table 8.20 is the only Costas PLL discriminator that remains linear over half of the input error range (±90°). In the presence of noise, all of the discriminator outputs are linear only near the 0° region. In an operational environment the PLL discriminator error signals are indeed periodic as indicated in Figure 8.44, but their amplitudes are severely attenuated beyond the phase limits of their pull-in ranges by the narrow bandwidths of their PLL tracking loops.

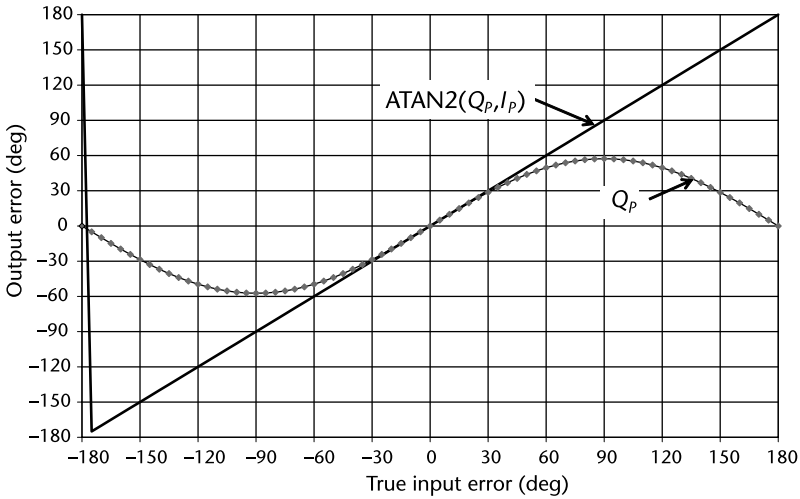
The Costas PLL characteristics are illustrated in Figure 8.45 where the phasor, A (the vector sum of I_P and Q_P), tends to remain aligned with the I-axis and switches 180° during each data (bit or symbol) transition.

8.6.1.3 FLL Discriminators

Table 8.21 summarizes three FLL discriminator algorithms, their output frequency errors and their characteristics.

Table 8.19 PLL Discriminators

Discriminator Algorithm	*Output Phase Error*	*Characteristics*
$ATAN2(Q_P, I_P)$	ϕ	Four-quadrant arctangent. Optimal (maximum likelihood estimator) at high and low SNR. Slope not signal amplitude dependent. High computational burden. Usually a table look-up implementation.
$\dfrac{Q_P}{Ave\sqrt{I_P^2 + Q_P^2}}$	$\sin\phi$	Q_P normalized by averaged prompt envelope. Slightly outperforms four-quadrant arctangent. Q_{PS} approximates ϕ to $\pm 45°$. Normalization provides insensitivity at high and low SNR. Also keeps slope not signal amplitude dependent. Low computational burden.



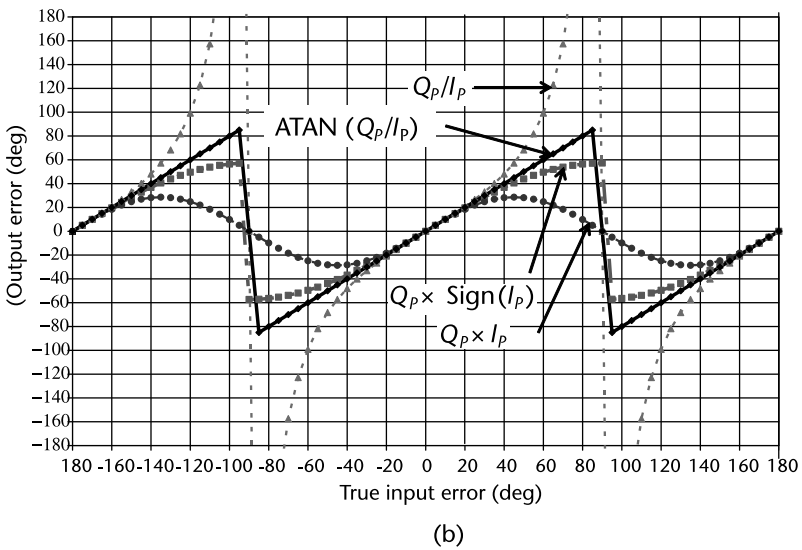


Figure 8.44 (a) Comparison of PLL discriminators and (b) comparison of Costas PLL discriminators.

Table 8.20 Common Costas Loop Discriminators

Discriminator Algorithm	*Output Phase Error*	*Characteristics*
$Q_P \times I_P$	$\sin 2\phi$	Classic Costas analog discriminator. Near optimal at low SNR. Slope proportional to signal amplitude squared A^2. Moderate computational burden.
$Q_P \times \text{Sign}\ (I_P)$	$\sin \phi$	Decision directed Costas. Near optimal at high SNR. Slope proportional to signal amplitude A. Least computational burden.
Q_P/I_P	$\tan \phi$	Suboptimal but good at high and low SNR. Slope not signal amplitude dependent. Higher computational burden. Divide by zero error at ±90°.
ATAN (Q_P/I_P)	ϕ	Two-quadrant arctangent. Optimal (maximum likelihood estimator) at high and low SNR. Slope not signal amplitude dependent. Highest computational burden. Usually a table look-up implementation.

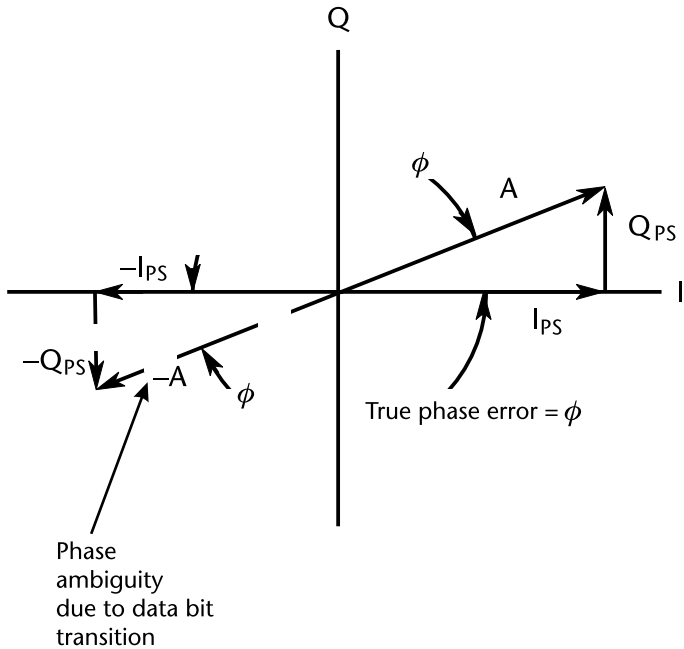


Figure 8.45 Costas I, Q phasor diagram depicting true phase error between replica and incoming carrier phase and the 180° phase changes due to data transitions.

Table 8.21 Common Frequency Lock Loop Discriminators

Discriminator Algorithm	*Output Frequency Error Characteristics*	
$\frac{cross}{(t_2 - t_1)}$	$\frac{\sin(\phi_2 - \phi_1)}{t_2 - t_1}$	Near optimal at low SNR. Slope proportional to signal amplitude squared A^2. Least computational burden.
where: $\text{cross} = I_{P1} \times Q_{P2} - I_{P2} \times Q_{P1}$		
$\frac{(cross) \times sign(dot)}{(t_2 - t_1)}$	$\frac{\sin[2(t_2 - t_1)]}{t_2 - t_1}$	Decision directed. Near optimal at high SNR. Slope proportional to signal amplitude A. Moderate computational burden.
where: $\text{dot} = I_{P1} \times I_{P2} + Q_{P1} \times Q_{P2}$		
$\frac{ATAN2(cross, dot)}{(t_2 - t_1)}$	$\frac{\phi_2 - \phi_1}{t_2 - t_1}$	Four-quadrant arctangent. Maximum likelihood estimator. Optimal at high and low SNR. Slope not signal amplitude dependent. Highest computational burden. Usually a table look-up implementation.

Note: Integrated and dumped prompt samples I_{P1} and Q_{P1} are the samples taken at time t_1, just prior to the samples I_{P2} and Q_{P2} taken at a later time t_2. For a data channel, these two adjacent samples should be within the same data bit or sample transition interval. The next pair of samples are taken starting $(t_2 - t_1)$ seconds after t_2 (i.e., no I and Q samples are reused in the next discriminator computation).

Figure 8.46 compares the frequency error outputs of these three discriminators assuming no noise in the prompt I and Q signals for 5-ms and 10-ms PIT. Figure 8.46(a) illustrates that the frequency pull-in range with a 5-ms PIT (reciprocal of 200-Hz bandwidth) has twice the pull-in range of Figure 8.46(b) with a 10-ms PIT

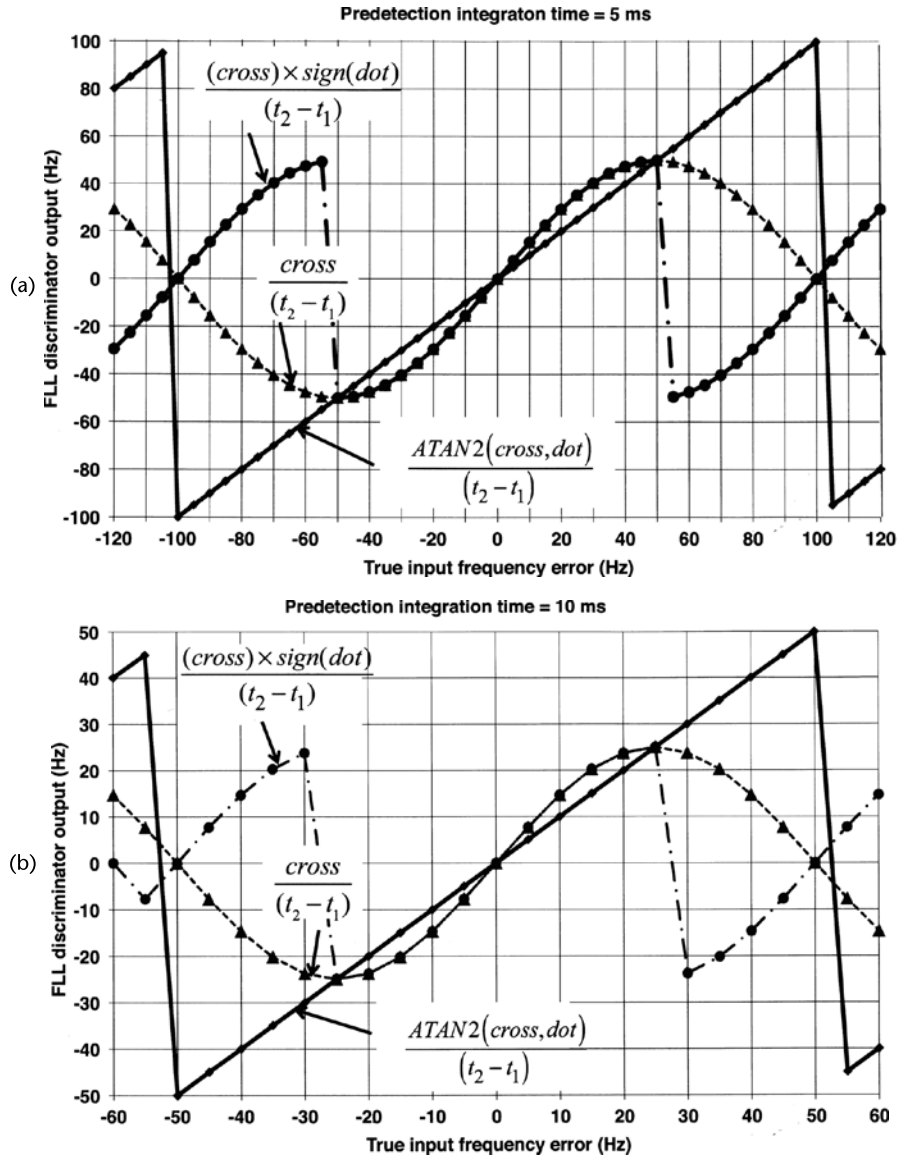


Figure 8.46 Comparison of FLL discriminators: (a) 5-ms predetection integration time, and (b) 10-ms predetection integration time.

(reciprocal of 100-Hz bandwidth). Note in both figures that the frequency pull-in range of the ATAN2 (cross, dot) FLL discriminator is half the reciprocal PIT bandwidth (±100 Hz for 5-ms PIT and ±50 Hz for 10-ms PIT) and that it has the widest pull-in frequency for a given PIT. The frequency pull-in range of the cross and (cross) × sign (dot) FLL discriminators have one fourth of the reciprocal PIT bandwidth (±50 Hz for 5-ms PIT and ±25 Hz for 10-ms PIT). The cross and (cross) × sign (dot) FLL discriminator error outputs are sine functions divided by the sample time interval $(t_2 - t_1)$ in seconds and were both divided by four to more accurately approximate the true frequency error output in their nearly linear regions.

In the presence of noise, all of the FLL discriminator outputs are linear only near the 0-Hz region. In an operational environment the FLL discriminator error

signals are indeed periodic as indicated in Figure 8.46, but their amplitudes are severely attenuated beyond the frequency limits of their pull-in ranges by the narrow bandwidths of their FLL tracking loops. In the presence of noise on the prompt I and Q signals the slopes of all of the FLL discriminator outputs tend to flatten as the noise levels increase so they are linear only near the 0-Hz error region.

The I, Q phasor diagram in Figure 8.47 depicts the change in phase, $\phi_2 - \phi_1$, between two adjacent samples of I_P and Q_P, at times t_1 and t_2. Any change in this frequency over this fixed time interval is proportional to the frequency error in the carrier-tracking loop. The figure illustrates that, in the case of a data channel, the FLL discriminator tolerates data bit or symbol transitions provided that the adjacent I and Q samples are taken within the same data bit or symbol interval. When these transition boundaries are unknown it is necessary to use very short predetection integration times so that most of the I,Q pairs do not straddle a transition boundary. Fortunately, this also increases the frequency pull-in range during FLL closure when this transition boundary uncertainty is most likely. It is also possible for the FLL loop to close with a false frequency lock in a high dynamic environment. Again, very short predetection integration times (wider pull-in range) are important for initial FLL loop closure. For example, if the code peak search dwell time was 1 ms or 2 ms, then the initial predetection integration time in FLL should be the same. Note that the FLL phasor amplitude, A, which is the vector sum of I_{PS} and Q_{PS} (i.e., the prompt envelope, rotates at a rate directly proportional to the frequency error between the replica carrier and the incoming carrier). When true frequency lock is actually achieved, the vector stops rotating, but it will stop at any angle with respect to the I axis. For this reason, coherent code tracking, as will be discussed in the following section, is not possible while in FLL because it depends on the I components being maximum (signal plus noise) and the Q components to be minimum (noise only) (i.e., in phase lock). It is possible to demodulate the SV data bit stream in FLL by a technique called differential demodulation. Because

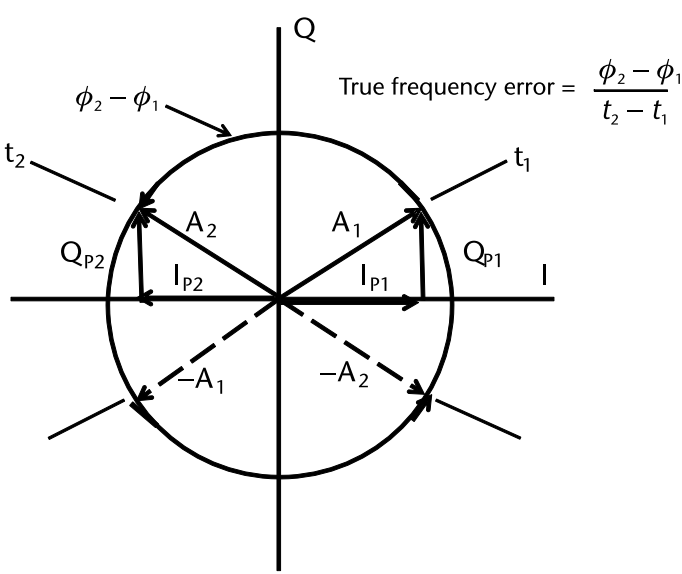


Figure 8.47 FLL I, Q phasor diagram depicting data or symbol transition tolerance if both samples are always taken within the same data bit or symbol transition boundaries.

the demodulation technique involves a differentiation (noisy) process, detecting the change in sign of the phasor in an FLL is noisier than detecting the sign of the integrated (lower noise) I_{PS} in a PLL. Therefore, for the same signal quality, FLL data detection has a much higher bit and word error rate than PLL data detection.

8.7 Code Tracking

The only parts of the code-tracking loop that were not described in Section 8.4 were the code discriminator and the code loop filter. The design of the programmable predetection integrators, the code loop discriminator and the code loop filter fully characterizes the receiver code-tracking loop. These three functions determine the most important two performance characteristics of the receiver code loop design: the code loop thermal noise error and the maximum line-of-sight dynamic stress threshold. It is shown later in this chapter that the code-tracking loop is the strong link and the carrier-tracking loop is the weak link in the determination of the GNSS receiver tracking threshold. Even though the carrier-tracking loop has a much lower tolerance than the code-tracking loop to interference that causes reduced C/N_0, it would invite disaster to attempt to aid the carrier loop with the code loop output. This is because, unaided, the code loop thermal noise is orders of magnitude larger than the carrier loop thermal noise. It was shown in Section 8.4 that it is always the carrier loop that aids the code loop and it will be shown that it is the ambiguous but precise carrier loop measurements that are used to improve the precision of the unambiguous (or at least easily resolved less ambiguous) code loop measurements. Chapter 9 describes how external velocity aiding typically from an IMU that has been calibrated by the synergism of being integrated with the GNSS receiver can temporarily maintain a sufficiently accurate estimate of the carrier wipe-off process that enables the code-tracking loop to sustain operation in the presence of interference. It is often overlooked that a known stationary GNSS receiver has virtually perfect velocity aiding capability without the need for an IMU. With this introductory insight into the virtues of the code-tracking loop, the code loop discriminator is described next. The loop filter design is described in Section 8.8.

8.7.1 Code Loop Discriminators

Table 8.22 shows the algorithms of four GNSS code loop discriminators and their characteristics. These are alternatively called delay lock loop (DLL) discriminators. These discriminators always use the early (E) and late (L) correlator phases and one of the versions also uses the prompt (P) signal. The fourth DLL discriminator is called a coherent dot product DLL. A more linear version can be implemented using only the E and L components, but the dot product slightly outperforms it. The coherent DLL provides superior performance when the carrier loop is in PLL. Under this condition, there is signal plus noise in the I components and mostly noise in the Q components. However, this high-precision DLL mode fails if there are frequent cycle slips or total loss of phase lock because the phasor rotates causing the signal power to be shared in both the I and Q components that consequently loses power in the coherent DLL. Successful operation requires a sensitive phase lock detector and rapid transition to the quasi-coherent DLL. All of the DLL discriminators can

Table 8.22 Code Discriminators

Discriminator Algorithm	*Characteristics*
$\frac{1}{2}\frac{E-L}{E+L}$ where $E=\sqrt{I_E^2+Q_E^2}, L=\sqrt{I_L^2+Q_L^2}$	Noncoherent early minus late envelope normalized by $E + L$ to remove amplitude sensitivity. High computational load. For 1-chip BPSK E-L correlator spacing, produces true tracking error within ±½-chip of input error (in the absence of noise). Becomes unstable (divide by zero) at ±1.5-chip input error, but this is well beyond code tracking threshold in the presence of noise.
$\frac{1}{2}\left(E^2-L^2\right)$	Noncoherent early minus late power. Moderate computational load. For 1-chip BPSK, E-L correlator spacing produces essentially the same error performance as ½ (E-L) envelope within ±½-chip of input error (in the absence of noise). Can be normalized with $E^2 + L^2$.
$\frac{1}{2}\left[\left(I_E-I_L\right)I_P+\left(Q_E-Q_L\right)Q_P\right]$ (dot product) $\frac{1}{4}\left[\left(I_E-I_L\right)/I_P+\left(Q_E-Q_L\right)/Q_P\right]$ (normalized with I_P^2 and Q_P^2)	Quasi-coherent dot product power. Uses all three correlators. Low computational load. For 1-chip BPSK E-L correlator spacing, it produces nearly true error output within ±½-chip of input (in the absence of noise). Normalized version shown second using I_P^2 and Q_P^2, respectively.
$\frac{1}{2}\left(I_E-I_L\right)I_P$ (dot product) $\frac{1}{4}\frac{\left(I_E-I_L\right)}{I_P}$ (normalized with I_P^2)	Coherent dot product. Can be used only when carrier loop is in phase lock. Low computational load. Most accurate code measurements. Normalized version shown second using I_P^2.

Note: The code loop discriminator envelopes may be noncoherently summed to reduce the iteration rate of the code loop discriminator and filter as compared to that of the carrier loop filter when the code loop is aided by the carrier loop or alternatively the discriminator outputs can be summed to reduce the iteration rate of the code loop filter. The rule-of-thumb limit is that total integration time must be less than one-fourth the DLL bandwidth. Note that this does not increase the predetection integration time for the code loop but does reduce noise. However, the code loop NCO must be updated every time the carrier loop NCO is updated even though the code loop filter output has not been updated. The most recent value of the code loop filter output is combined with the current value of carrier aiding.

be normalized. Normalization removes sensitivity to signal amplitude fluctuations that improves performance under rapidly changing C/N_0 conditions. Therefore, normalization helps the DLL tracking and threshold performance to be independent of automatic gain control (AGC) performance. However, normalization does not prevent reduction of the gain (slope) when C/N_0 decreases. As C/N_0 is reduced, the DLL slope approaches zero. Since loop bandwidth is roughly proportional to loop gain, then loop bandwidth approaches zero at low C/N_0. This results in poor DLL response to dynamic stress and can result in instability if a third-order DLL filter is used (never used with carrier aided code implementation). Carrier aiding (including externally provided carrier aiding) minimizes this problem, but the phenomena may produce unexpected DLL behavior at very low C/N_{0}-.

Figure 8.48 compares these four DLL discriminator outputs assuming a BPSK-R modulated signal with 1-chip spacing between the early and late correlators. This means that a 2-bit shift register is shifted at twice the clock rate of the code generator. Also assumed is an ideal correlation triangle (infinite bandwidth) and no noise on the incoming I and Q signals. For typical receiver bandwidths, the correlation

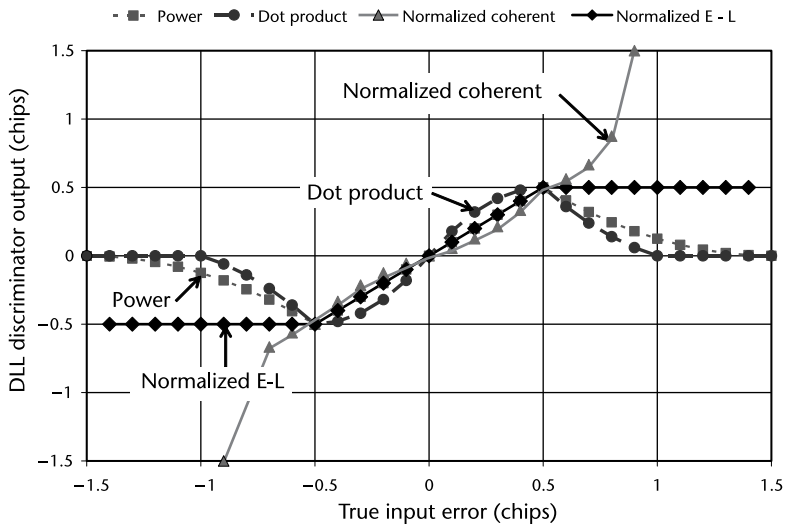


Figure 8.48 Comparison of delay lock loop discriminators.

peak tends to be rounded, the ramps on either side of the peak are nonlinear and the correlation amplitudes at ±½-chip from the correlation peak are slightly higher than for the infinite bandwidth case while the prompt correlation amplitude is slightly lower. These DLLs can be used with BOC signals but the E and L correlator spacing (and shift register design) must be optimized based on the autocorrelation, typically much shorter than ½-chip, and there may be more correlators involved to resolve ambiguity problems with the autocorrelation envelope. Specific design examples are provided in Section 8.7.3.

The normalized early minus late envelope discriminator is very popular because its noise-free output error is linear over a ±1-chip range and has a pull-in range to almost ±1.5-chip, but the dot product power discriminator slightly outperforms it. Early GPS receiver designs synthesized an $E - L$ replica code to save one correlator (i.e., only one complex correlator is required to generate the composite $E - L$ signal that can be normalized with the P signal, but linearized $E + L$ normalization requires separate E and L correlators).

8.7.2 BPSK-R Signals

Figure 8.49 illustrates the envelopes that result for three different replica code phases being correlated simultaneously with the same incoming BPSK-R modulated signal assuming infinite bandwidth. For ease of visualization, the incoming BPSK-R modulated signal is shown without noise. The three replica phases are separated by ½-chip and are representative of the early, prompt and late replica codes that are synthesized in a typical BPSK-R replica code generator, although narrower early to late correlator separations are used to reduce multipath error and measurement noise. Narrow correlators have reduced code tracking loop dynamic stress tolerance, but if carrier aided code tracking is used, the carrier loop removes most of the code loop dynamic stress thereby enabling both narrow correlator spacing and smaller code loop filter noise bandwidths after the code loop reaches steady state.

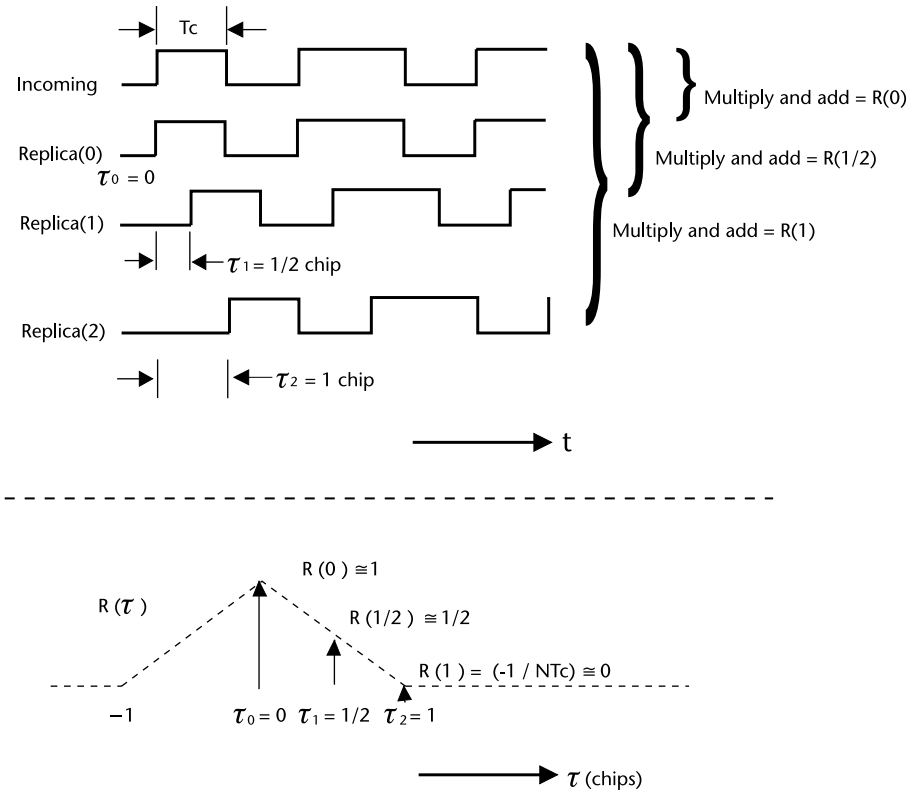


Figure 8.49 BPSK-R code correlation process for three different replica code phases.

Figure 8.50 illustrates how the early, prompt and late envelope amplitudes change as the phases of the replica code signals are advanced with respect to the incoming BPSK-R signal. For ease of visualization, only 1-chip of the continuous incoming PRN code and replica code phases is shown and the incoming signal is shown without noise. In reality, the incoming PRN code is buried in noise and a massive number of correlated products must be accumulated in each correlator phase to provide the necessary bandwidth reduction and resulting processing gain for each envelope amplitude to emerge out of the noise.

Figure 8.51 illustrates the normalized early minus late envelope discriminator error output signals that correspond to the four replica code offsets of Figure 8.50.

The BPSK-R signal closed code loop operation becomes apparent as a result of studying these replica code phase changes, the envelopes that they produce, and the resulting error output generated by the early minus late envelope code discriminator. If the replica code is aligned, then the early and late envelopes are equal in amplitude with no error generated by the discriminator. If the replica code is misaligned, then the early and late envelopes are unequal by an amount proportional to the code phase error between the replica and the incoming signal (within the limits of the discriminator pull-in range). The code discriminator senses the amount of error in the replica code and the direction (early or late) from the difference in the amplitudes of the early and late envelopes. This error is filtered and then applied to the code loop NCO where the output frequency is increased or decreased as

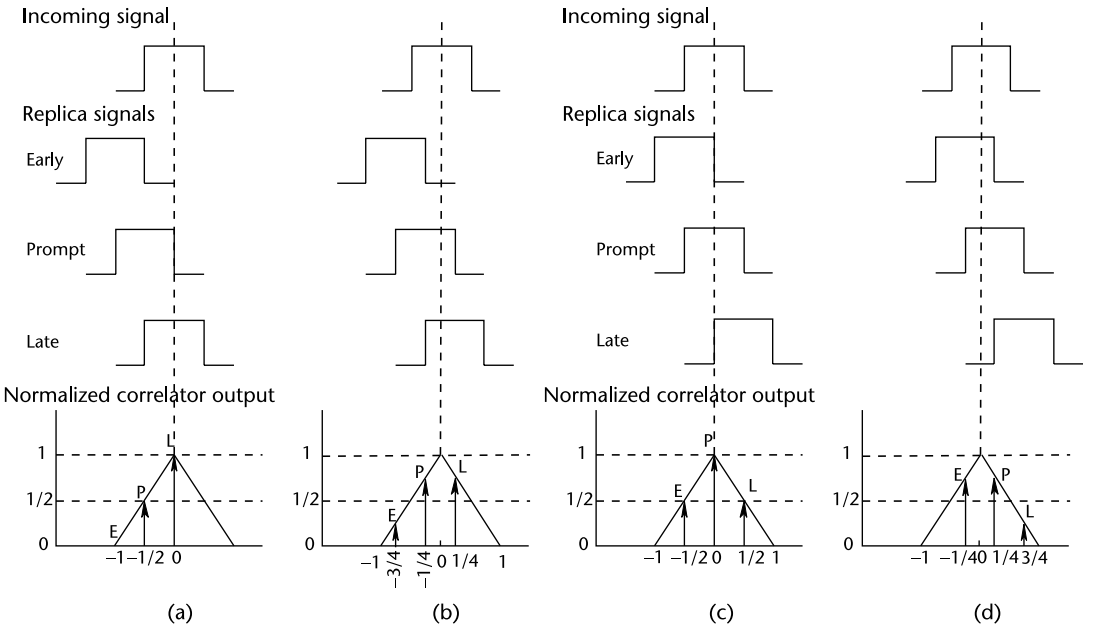


Figure 8.50 BPSK-R code correlation phases: (a) prompt replica code ½-chip early, (b) prompt replica code ¼-chip early, (c) prompt replica code aligned, and (d) prompt replica code ¼-chip late.

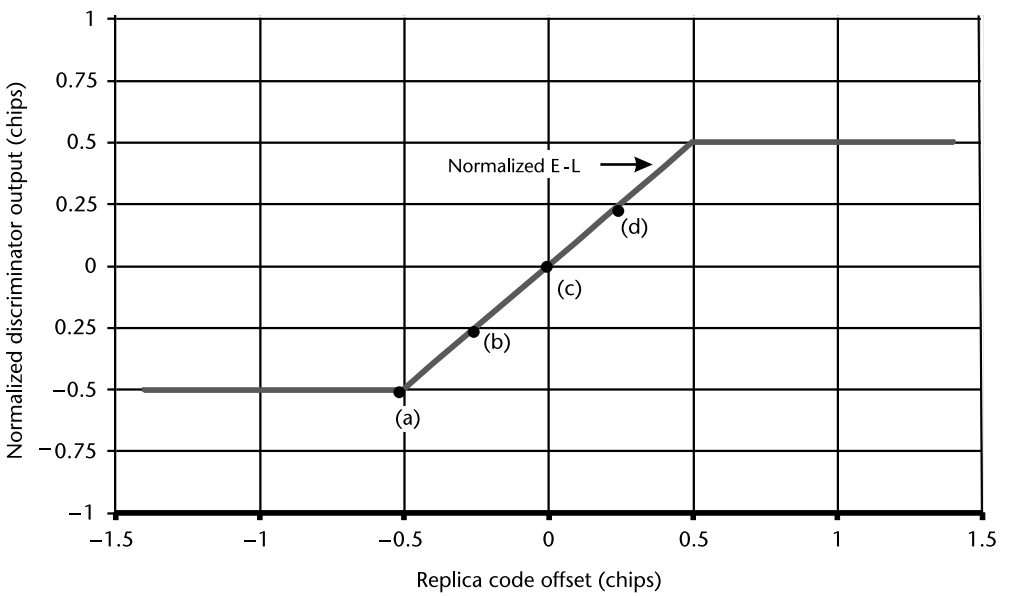


Figure 8.51 BPSK-R code discriminator output versus replica code offset.

necessary to correct the replica code generator phase with respect to the incoming SV signal code phase.

The discriminator examples given thus far have assumed that each channel of the receiver contains three complex code correlators to provide early, prompt, and late correlated outputs. In early generations of GPS receiver designs, analog

correlators were used instead of digital correlators because ADC technology was not fast enough to digitize the signals prior to correlation. There was strong emphasis on reducing the number of expensive and power-hungry analog correlators, so there were numerous code-tracking loop design innovations that minimized the number of correlators. The tau-dither technique time-shares the early and late replica code with one complex (I and Q) correlator and provides the required early minus late discriminator and the early plus late normalizing signal. This time-share technique suffers a 3-dB loss of code tracking threshold because only half the energy is available from the early and late signals. This loss of code tracking threshold is unimportant in an unaided receiver design because there is much more than 3-dB difference between code loop and carrier tracking loop thresholds. The extra margin in the code loop threshold only pays off for aided receivers. The TI 4100 multiplex GPS receiver [63, 64] not only used the tau-dither time-share technique, but also time-shared two analog correlators and one replica code generator and one carrier generator to simultaneously and continuously track using 2.5-ms dwells on the L1 P-code and L2 P-code signals of four GPS satellites in phase lock. It also simultaneously demodulated the 50-Hz navigation messages. Because the L2 tracking was accomplished by tracking L1-L2, this nearly zero dynamics signal permitted very narrow bandwidth tracking loops, and therefore suffered only a little more than 6 dB of tracking threshold losses instead of the expected 12 dB. Since the same circuits were time-shared (multiplexed) across all channels and frequencies, there was zero interchannel bias error in the TI 4100 measurements and there was virtually zero quantization error in the code and carrier measurements. The TI 4100 was the first commercial GPS receiver and the first high precision geodetic surveying receiver. The TI 4100 is the only commercial receiver in history that performed L1 and L2 P-code interferometric measurements. This was because it operated with normal P-code access using signals from the Block I GPS satellites that preceded the adoption of P(Y) code. As a result, the TI 4100 revolutionized the field of geodetic surveying by achieving "5 dimes" (5-mm) benchmark location precision over a 10,000-m baseline with less than 15 minutes of L1/L2 observations on four SVs. The TI 4100 was later adapted to operate with P(Y) code, but could only be purchased by authorized users. GPS P(Y) code codeless/semicodeless processing techniques were used in next-generation commercial high-precision receivers. These techniques are described in Section 8.6.3, but this technology will rapidly disappear with the emergence of multiple modernized commercial GNSS frequencies. This new era in commercial GNSS signals represents a major breakthrough for all high-precision commercial GNSS applications.

Modern digital GNSS receivers often contain many more than three complex correlators because digital correlators are relatively inexpensive (e.g., only one exclusive-or circuit is required to perform the 1-bit multiply function). The innovations relating to improved performance through the use of precorrelation ADCs, DSP technology, and more than three complex correlators include faster acquisition times (described earlier), multipath mitigation (described in Chapter 9) and a wider discriminator correlation interval that provides jamming robustness when combined with external (IMU) aiding [65]. However, there is no improvement in tracking error due to thermal noise or improvement in tracking threshold using multiple correlators. Reducing parts count and power continue to be important, so multiplexing is back in vogue for hardware-based digital components, but now

using faster than real-time digital multiplexing techniques with no loss in signal power. The speed of digital circuits has increased to the point that correlators, NCOs, and other high-speed baseband functions can be digitally multiplexed without a significant power penalty because of the reduction in feature size of faster digital components. The multiplexing is faster than the real-time digital sampling of the GNSS signals by a factor of N where N is the number of channels sharing the same device. Since there is no loss of energy, there is no loss of signal processing performance as was the case with the TI 4100 analog time division multiplexing and there is no interchannel bias error. In the case of software-defined functions, there is a significant amount of parallelism in modern DSP architectures and every reentrant function that is implemented in a DSP is equivalent to digital multiplexing.

8.7.3 BOC Signals

The ambiguity problem encountered with BOC signals and three techniques for removing the ambiguity were introduced in Section 8.5.6 in the context of the M-code BOC signal acquisition. The same techniques can be used for any modernized BOC signal that produces an ambiguous code discriminator error output. In the case of the interoperable MBOC signal, the code discriminator ambiguity is a small order effect that does not present a serious code discriminator ambiguity problem. As is the case for M code, if it is preferred to check (then correct if necessary) for ambiguity instead of preventing it, the ambiguity in other GNSS BOC signals can be detected with an additional very early and very late correlator, each correlator set at the peaks of the unwanted (ambiguous) correlation peaks. If there is a major change in the difference between these two signals, that error provides both the detection and direction for bumping the replica code phase back into the correct phase.

For more details on GNSS signals in general, [39] provided comprehensive information, in particular, describing discriminators for code tracking of BOC(1,1), BPSK-R(1) and TMBOC(6,1,4/33) GNSS signals.

8.7.4 GPS P(Y)-Code Codeless/Semicodeless Processing

Historically, P(Y) code has been broadcast on L2 by the GPS SVs on all operational SVs with AS activated. This has denied direct two-frequency operation to commercial (L1 C/A code) GPS users until the recent advent of L2C and L5 civil signals plus emerging multifrequency open service signals from other GNSS constellations. Precision surveying requires differential interferometry techniques (requiring carrier-phase measurements) that remove common mode bias errors to achieve centimeter-level precision over short baselines, but the ionospheric delay is not common mode over long baselines. Therefore, dual-frequency carrier-phase measurements are essential for precision surveying applications over long baselines. As a result, commercial GPS receiver designers pursued design techniques that could obtain L2 Y-code pseudorange and carrier-phase measurements without full signal access that cryptographic knowledge provides. These techniques are referred to as either codeless or semicodeless processing. Codeless techniques only utilize the known 10.23-MHz chip rate of the Y-code signal and the fact that [66] assured that the same Y-code signal is broadcast on both L1 and L2, whereas the semicodeless techniques further exploit a deduced relationship between the Y code and P code.

Since they operate without full knowledge of the Y code signal, the codeless and semicodeless designs operate at significantly reduced signal-to-noise ratios, which require the tracking loop bandwidths to be extremely narrow. This, in turn, reduces their ability to operate in a high dynamic environment without aiding. Fortunately, robust aiding is generally available from tracking loops operating upon the C/A code signal. Typical codeless and semicodeless receiver designs use L1 C/A code-tracking loops to effectively remove the line-of-sight dynamics from the L1 and L2 Y-code signals, and then extract the L1 to L2 differential measurements by some variation of a signal squaring technique, which does not require full knowledge of the replica code. Codeless techniques effectively treat the Y-code PRN as 10.23-Mbps data, which can be removed through squaring or by cross-correlation of the L1 and L2 signals. Semicodeless techniques exploit some known features of the Y code; see, for example, [67]. In addition to the signal-to-noise disadvantage mentioned earlier, codeless receivers suffer from other robustness problems. Although the parallel C/A-code processing provides access to the GPS navigation message, codeless processing of L2 does not allow decoding of the navigation data for the purposes of verifying that the desired SV is being tracked. Also, two SVs with the same Doppler will interfere with each other in the codeless mode; therefore, the scheme fails for this temporary tracking condition. However, modern semicodeless receivers provide relatively robust tracking of the L2 Y-code signal with assistance from the L1 C/A code. These concepts will become obsolete when the modernized GPS civil signals become available.

8.8 Loop Filters

The objective of the loop filter is to reduce noise to produce an accurate estimate of the original signal at its output. The loop filter order and noise bandwidth also determine the loop filter's response to signal dynamics. The only difference between carrier and code loop digital filter design for carrier tracking or for code tracking is the order of the loop filter and their bandwidths. The loop filter order and noise bandwidth, B_n, are determined based on trade-offs between the expected operating environment, various receiver component noise contributions and the desired precision of tracking the signal. The difficulty of the design has more to do with these trade-offs than with the digital loop filter design techniques that are virtually identical for the same digital loop filter order. Because the loop filter is part of a feedback loop, there are stability issues associated with the loop order for a desired noise bandwidth because there are predetection integration time and computation time that represent a delay in the closed loop. The stability issue for the digital loop filter design approach described next will be detailed in Section 8.8.5. As shown in the receiver block diagrams, the loop filter's output signal is effectively subtracted from the original signal to produce an error signal that is filtered and used to correct the carrier and code replica signals in a closed loop process.

There are many design approaches to digital filters. The design approach by Holmes [68] described herein draws on existing knowledge of analog loop filters, and then adapts these into digital implementations. The well-known shortcoming of this digital filter design is that the loop filter noise bandwidth (B_n) in units of hertz, slightly increases with predetection integration time (T) in units of seconds.

The consequence is that the dimensionless B_nT product does not remain constant for all T. Newer digital loop filter designs (e.g., Stevens and Thomas [69] and Thomas [70]) can achieve stability with a constant B_nT product that is larger than the stable value of B_nT that can be achieved with any loop filter based on analog design techniques. However, the Holmes [68] loop filter design stability criteria are predictable and suffice for the majority of GNSS receiver applications. Figure 8.52 shows block diagrams of first, second, and third-order analog filters [68]. Analog integrators are represented by $1/s$, the Laplace transform of the time-domain integration function. The input signal is multiplied by the multiplier coefficients, and then processed as shown in Figure 8.52. These multiplier coefficients and the number of integrators completely determine the loop filter's characteristics. Table 8.23 summarizes these filter characteristics and provides all the information required to compute the filter coefficients for first, second, and third-order loop filters. Only the filter order and noise bandwidth must be chosen to complete the design and this must be consistent with the loop filter stability criteria presented in Section 8.8.5.

Figure 8.53 depicts the block diagram representations of analog and digital integrators. The analog integrator of Figure 8.53(a) operates with a continuous time-domain input, $x(t)$, and produces an integrated version of this input as a continuous time-domain output, $y(t)$. Theoretically, $x(t)$ and $y(t)$ have infinite numerical resolution and the integration process is perfect. In reality, the resolution is limited

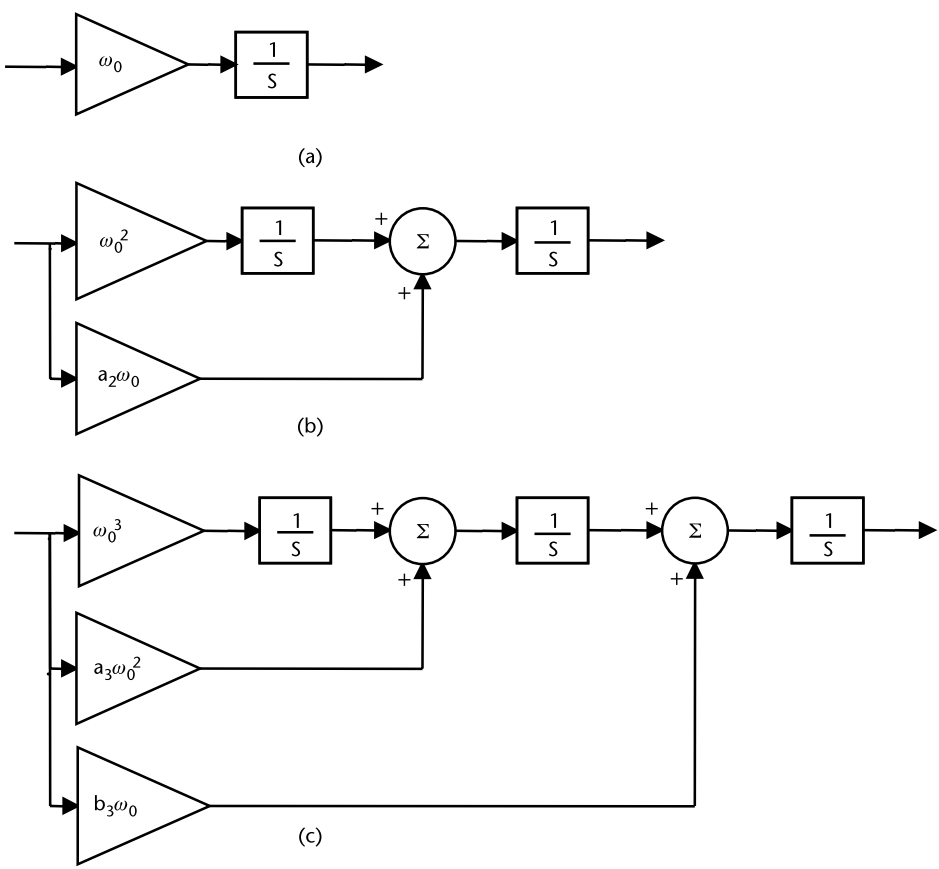


Figure 8.52 Block diagrams of: (a) first-, (b) second-, and (c) third-order analog loop filters.

Table 8.23 Loop Filter Characteristics

Loop Order	*Noise Bandwidth B_n (Hz)*	*Typical Filter Values*	*Steady State Error*	*Characteristics*
First	$\frac{\omega_0}{4}$	ω_0 $B_n = 0.25\omega_0$	$\frac{(dR/dt)}{\omega_0}$	Sensitive to velocity stress. Used in aided code loops.
Second	$\frac{\omega_0(1+a_2^2)}{4a_2}$	ω_0^2 $a_2\omega_0 = 1.414\omega_0$ $B_n = 0.53\omega_0$	$\frac{d^2R/dt^2}{\omega_0^2}$	Sensitive to acceleration stress. Used in aided code loops and aided and unaided carrier loops. Optimum damping factor $\delta = 0.707 = a_2/2$.
Third	$\frac{\omega_0(a_3b_3^2 + a_3^2 - b_3)}{4(a_3b_3 - 1)}$	ω_0^3 $a_3\omega_0^2 = 1.1\omega_0^2$ $b_3\omega_0 = 2.4\omega_0$ $B_n = 0.7845\omega_0$	$\frac{(d^3R/dt^3)}{\omega_o^3}$	Sensitive to jerk stress. Used in unaided carrier loops. Parameters provide fastest response to step function with minimal initial overshoot.

Source: [68]. Notes: (1) The loop filter natural radian frequency, ω_o, is computed from the value of the loop filter noise bandwidth, B_n, selected by the designer. (2) R is the line-of-sight range to the satellite. (3) The steady state error is inversely proportional to the nth power of the tracking loop bandwidth and directly proportional to the nth derivative of range, where n is the loop filter order.

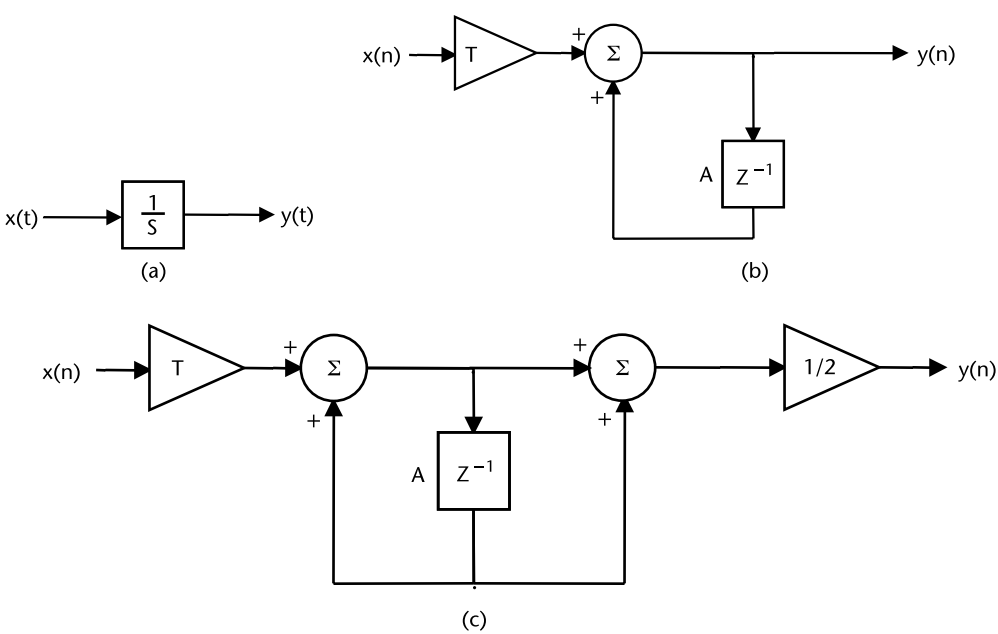


Figure 8.53 Block diagrams of: (a) analog, (b) digital boxcar, and (c) digital bilinear transform integrators.

by noise that significantly reduces the dynamic range of analog integrators. There are also problems with drift.

The boxcar digital integrator of Figure 8.53(b) operates with a sampled time-domain input, $x(n)$, which is quantized to a finite resolution, and produces a discrete integrated output, $y(n)$. The time interval between each sample, T, represents a unit delay, z^{-1}, in the digital integrator. The digital integrator performs discrete

integration perfectly with a dynamic range limited only by the number of bits used in the accumulator, A. This provides a dynamic range capability much greater than can be achieved by its analog counterpart and the digital integrator does not drift. The boxcar integrator performs the function $y(n) = T[x(n)] + A(n - 1)$, where n is the discrete sampled sequence number.

Figure 8.53(c) depicts a digital integrator that linearly interpolates between input samples and more closely approximates the ideal analog integrator. This is called the bilinear z-transform integrator. It performs the function $y(n) = T/2[x(n)] + A(n - 1) = 1/2[A(n) + A(n - 1)]$. The digital filters depicted in Figure 8.54 result when the Laplace integrators of Figure 8.52 are each replaced with the digital bilinear integrator shown in Figure 8.54(c). The last digital integrator is not included because the NCO that immediately follows provides this function. The NCO is equivalent to the boxcar integrator of Figure 8.53(b), but when combined with the predetection integration and dump function in the feedback loop, it is equivalent to the bilinear integrator of Figure 8.53(c) [71].

8.8.1 PLL Filter Design

The PLL filter design will typically be second-order for moderate dynamic applications and third-order for higher dynamic applications. The second-order PLL filter will have one digital bilinear transform integrator and the carrier NCO provides the second one. The third-order PLL will have two digital bilinear transform integrators and the carrier NCO provides the third one.

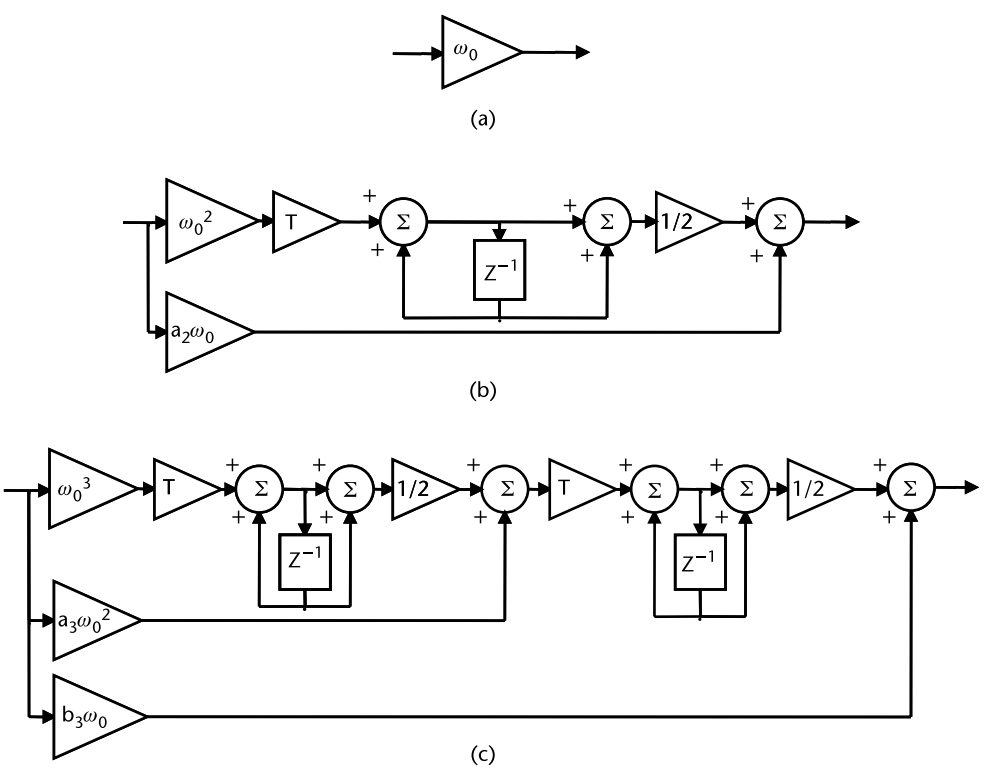


Figure 8.54 Block diagrams of (a) first, (b) second, and (c) third-order digital loop filters excluding last integrator (the NCO).

The Costas PLL is not defined by the digital loop filter, but by the presence of data or symbol modulation on the signal and the need for a Costas carrier loop discriminator. As will be seen, the Costas PLL encounters squaring loss that can only be reduced by increasing T. Herein lies the issue of extending the stable B_nT product. However, modernized GNSS signals have pilot channels that operate with a pure PLL discriminator that theoretically has zero squaring loss. Hence, the emphasis on increasing the stable B_nT product becomes a moot point, although increasing T does provide additional noise filtering. In either case there is no difference in the loop filter design. The only difference is the discriminators (that have already been described).

8.8.2 FLL Filter Design

The FLL filter design will typically be one order lower than the PLL filter design but it also requires one more integrator than its PLL counterpart because the FLL discriminator produces frequency error instead of phase error. Therefore, the first-order FLL will have one digital bilinear transform integrator and the NCO provides the second integrator. The second-order FLL will have two digital bilinear transform integrators and the NCO provides the third integrator.

8.8.3 FLL-Assisted PLL Filter Design

There are GNSS receiver designs that operate only in FLL and therefore cannot achieve the precision of their PLL counterparts and suffer much higher bit error rates in data demodulation. There are also GNSS receiver designs that cannot support FLL and therefore must close directly in PLL without the benefit of the extended FLL frequency uncertainty pull-in range nor can they revert to FLL operation in the presence of high dynamic stress. The major performance shortcomings of these latter design choices have been presented and will not be rationalized by further discussion of any cost or performance benefits derived from them. Instead, a synergistic solution to the GNSS receiver designer's carrier tracking loop dilemma of holding on in the presence of sudden changes in platform dynamic stress while also operating with the highest precision and preferred data demodulation mode of PLL most of the time. It is called an FLL-assisted-PLL [40].

Figure 8.55 illustrates two FLL-assisted-PLL loop filter designs [40]. Figure 8.55(a) depicts a second-order PLL filter with a first-order FLL assist. Figure 8.55(b) depicts a third-order PLL filter with a second-order FLL assist. If the PLL error input is zeroed in either of these filters, the filter becomes a pure FLL. Similarly, if the FLL error input is zeroed, the filter becomes a pure PLL. The best loop closure process (maximum frequency uncertainty pull-in capability) is to close in pure FLL, then apply the error inputs from both discriminators as an FLL-assisted PLL until phase lock is achieved, then convert to pure PLL until phase lock is lost. However, if the noise bandwidth parameters are chosen correctly, there is only about 1-dB loss in PLL carrier tracking threshold performance (due to FLL discriminator output noise) if both discriminators are continuously operated after initial FLL loop closure [40]. In general, the natural radian frequency of the FLL, ω_{0f}, is

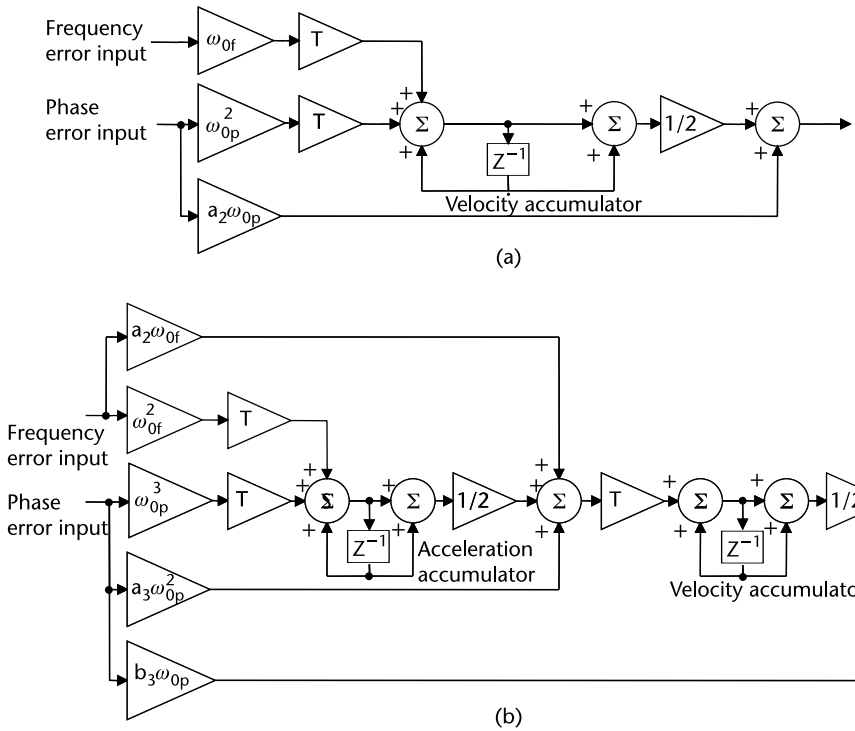


Figure 8.55 Block diagrams of FLL assisted PLL filters: (a) second-order PLL with first-order FLL assist, and (b) third-order PLL with second-order FLL assist.

different from the natural radian frequency of the PLL, ω_{0p}. These natural radian frequencies are determined from the desired loop filter noise bandwidths, B_{nf} and B_{np}, respectively. The values for the second-order coefficient a_2 and third-order coefficients a_3 and b_3 can be determined from Table 8.23. These coefficients are the same for FLL, PLL or DLL applications if the loop order and the noise bandwidth, B_n, are the same. Note that the FLL coefficient insertion point into the filter is one integrator back from the PLL and DLL insertion points. This is because the FLL error is in units of hertz (change in range per unit of time), whereas the PLL and DLL errors are in units of phase (range).

8.8.4 DLL Filter Design

The DLL filter design will typically be first order but there are design cases where second order is required (for example where loose-coupled external velocity aiding is temporarily operating the carrier loop in a high interference situation). The DLL loop filter B_n should always be very narrow (typically less than 1 Hz in steady state) since it should always be aided by the carrier loop and therefore has little or no dynamic stress to track. The first-order DLL has no digital bilinear transform integrators since the one integrator is provided by the code NCO. The second order DLL has one digital bilinear transform integrator and the second integrator is provided by the code NCO.

8.8.5 Stability

A Bode analysis technique is described in [71] and used to assure tracking stability in GNSS digital tracking loops designed with the Holmes [68] loop filters. The transfer functions for all components of the digital tracking loop are presented and used in this analysis technique. The composite transfer functions of first-, second-, and third-order tracking loops are determined. The stable value of the B_nT product is significantly reduced when there is computation delay within the tracking loop. Therefore, computation delay is modeled and utilized in the results reported in [71]. After detailed development of equations along with numerous block diagrams, the Bode analysis technique is ultimately used to determine the B_nT values corresponding to 0° (unstable) and 30° (stable) phase margins for all three loop orders and for the extreme two cases of zero (or on time) computation delay and T computation delay.

The key results from [71] that assure loop filter stability are presented herein beginning with a recap of carrier tracking signal processing using the functional block diagram of a GNSS receiver digital carrier-tracking loop shown in Figure 8.56. It depicts all of the processes that take place starting with the analog intermediate frequency (analog IF). After analog-to-digital conversion into digital IF, the complex carrier wipe-off process takes place followed by the prompt code wipe-off process. (The carrier wipe-off process starting with a baseband analog signal produces the identical results when the NCO IF bias is set to zero.) The resulting error signal is fed to the predetection filter where it is integrated and then dumped after a predetection integration time, T, into the phase detector. That error is fed to the PLL filter where it is integrated and fed to the carrier NCO. The output of the NCO is fed to a mapping function that converts the NCO representation of the replica carrier phase at IF into a complex cosine and sine replica of the incoming digital IF that is used to close the carrier-tracking loop. In Figure 8.56, the analog IF signal contains all of the in-view and in-band GNSS signals submerged in noise. Each analog IF signal (plus Doppler) has virtually the same in-band spectrum characteristic as it had in that same bandwidth at L-band, but the carrier frequency has been downconverted from L-band to the much lower IF and the composite signal plus noise has been digitized (hopefully with a minimum of aliasing). The digital carrier PLL is designed to track only the signal Doppler frequency of the selected GNSS signal, so there is a carrier frequency offset number (shown as NCO IF bias)

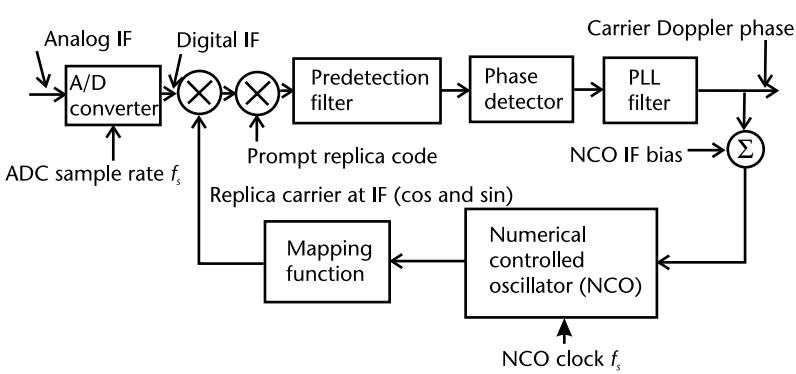


Figure 8.56 Block diagram of GNSS digital carrier-tracking loop.

that is added to each PLL Doppler frequency output value before it is fed into the NCO. Since the replica signal and the desired signal offset each other, they are not needed for the open loop transfer analysis, described next.

Figure 8.57 illustrates the open loop model of this carrier-tracking loop as implemented for the purpose of phase margin analysis by the Bode method [72] used for this stability analysis. This figure was also used for the purpose of determining the transfer function of a PLL. With appropriate changes of the labels in the figure, the same model can be used to determine the transfer function of a digital FLL and a digital DLL. For purposes of this Bode analysis, this model illustrates the equivalent effect of the carrier-tracking loop in Figure 8.56 when the loop has been opened and all signals external to the loop have been zeroed, as is required by the Bode analysis technique. Bode analysis is a linear analysis technique, so only small perturbations are considered and nonlinear effects are not taken into account in calculating phase margins, but the nonlinear effects tend toward improving the stability margin.

Note that Figure 8.57 does not depict the details of the actual replica I and Q signals emanating from a complex mapping function. These replicated complex (I and Q) signals perform carrier wipe-off of the digital IF signal in separate mixers. Nor does it depict the resulting I and Q error signals that are separately integrated and dumped by the predetection filter, where that complex result is fed to the phase detector. But the modeled effect and therefore the modeled transfer functions are the same. For example, the transfer functions of both the actual and the modeled mapping function are both assumed to be unity. Figure 8.57 also does not show the necessary prompt code wipe-off function shown in Figure 8.56 because the prompt code wipe-off transfer function is assumed to be unity. Assuming the transfer functions of all Figure 8.56 functions omitted in Figure 8.57 are indeed unity, then Figure 8.57 accurately models all of the processes that relate to the response and delay of a GNSS carrier-tracking loop under the Bode analysis condition, except for an additional important requirement: there will be additional real-time computation delay, t_c, between the predetection filter "dump" epoch and the epoch when the NCO actually receives that new frequency input. If t_c plus other processing delays in the total loop are shorter than or equal to a single real-time interval between

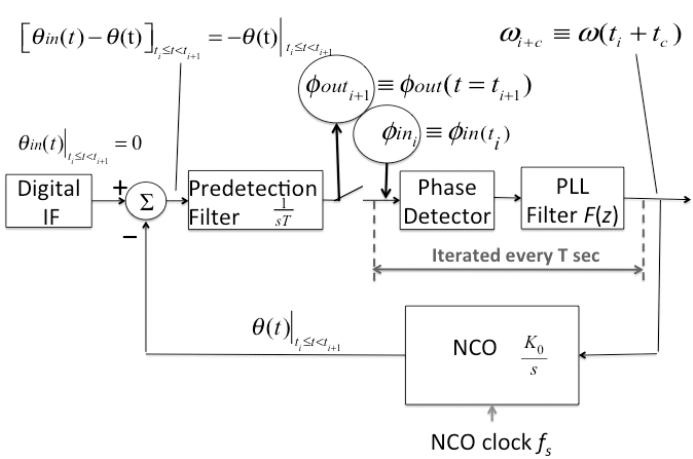


Figure 8.57 Model of GNSS digital carrier-tracking loop under Bode test condition.

NCO clock epochs (that are synchronous with the digital IF sample epochs), then those delays still provide an output by the time the next input signal sample arrives. For this case, the delays can be considered as not adding any additional delay to the loop. They fit within the total iteration time T, and thus can be considered as having the same result as an output which was immediately available with zero delay but still cannot be applied until the next epoch. Otherwise, the computation delay, t_c, must be included in the transfer function.

To clarify the use of Figure 8.57 in the context of the Bode analysis technique, all of the relevant closed loop functions are shown with the understanding that Bode analysis is strictly an open loop process. So the combination of the phase detector, PLL filter, the NCO, and the predetection filter is called the PLL open loop transfer function (G) and the loop is opened at some arbitrary point, in this case, between the predetection filter output and the phase detector input. Then a sine wave excitation signal, labeled ϕin_i, is input into the open loop circuit at that break. Any signals that are not inside the loop are irrelevant to the Bode open loop analysis, so they are set to zero during the analysis. Therefore, $\theta in_i(t)$ and the NCO IF bias are both set to zero. Under this test condition, the open loop signal can be analyzed or tested for either the amplitude or phase response of that sine wave excitation after it has traveled completely around the loop back to the break point. It is the phase shift response of that sine wave after it has traversed the loop (including any waiting time that may be involved at the output for the next time epoch to occur) that provides the phase curve portion of a Bode plot. The observation signal, labeled ϕout_{i+1}, is located at the output of the open loop at that break. Clearly, the transfer functions and thus the phase response of the circuit in closed loop and open loop operation are entirely different. But the phase margins discussed for phase locked loops (that are indeed closed loops) can be obtained from an open loop Bode analysis. The phase margin is obtained by comparing the open loop phase response to the open loop phase response that causes the loop to be unstable if the loop is closed.

The open loop analysis starts at the input of the phase detector with the error signal, ϕin_i , where the subscript i denotes the ith numbered time sample. The transfer function of the Phase Detector is assumed to be unity (i.e., the PLL discriminator translates the incoming phase error into the same output phase error with unity gain). (Note: This is not true under poor signal-to-noise ratio conditions, but this is a nonlinear effect that is not considered in this analysis.) This error signal then goes through the PLL filter, with transfer function $F(z)$, that is implemented in the digital z-domain, with time samples spaced T seconds apart. The output frequency error of the PLL filter is ω_i and is fed to the NCO to perform the last integration of the loop. However, the NCO and its mapping function have an iteration rate based on the NCO clock frequency, f_s, which is typically several orders of magnitude faster than the iteration rate of the PLL filter and is equal to and synchronous with the digital IF sample rate. Therefore, the combined NCO and mapping function of Figure 8.56 are modeled as a continuous analog integrator, $\frac{K_0}{s}$, called NCO in Figure 8.57, that converts the input frequency into phase at the output. The predetection filter in Figure 8.57 (that performs the integrate-and-dump function) is iterating at the digital IF sample rate that is also orders of magnitude faster than the PLL Filter update rate. Therefore, the predetection filter is modeled as an analog integrator

that integrates for T seconds, $\frac{1}{sT}$. The open loop analysis ends at the output signal of the predetection filter, ϕout_{i+1}. Thus, the analysis treats the PLL filter as a slow rate digital implementation, but the NCO and predetection filter are treated as continuous high rate analog components. The transfer function of each PLL component is derived in [71] based on this model. These are combined into the total open loop transfer function as follows:

$$G(z) = \left[\frac{T}{2}\frac{\left(1+z^{-1}\right)z^{-1}}{1-z^{-1}}\right]F_n(z) \tag{8.61}$$

where $F_n(z)$ is the transfer function of the nth order loop filter excluding the NCO.

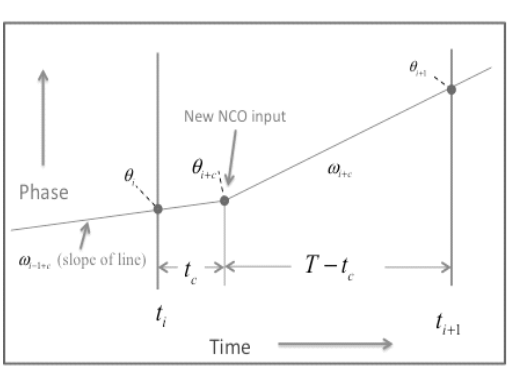
If $t_c \leq T$ is nonzero and significant, then (8.61) must be modified by multiplying by $z^{-\frac{t_c}{T}}$ as follows:

$$G(z) = \left[\frac{T}{2}\frac{\left(1+z^{-1}\right)z^{-1}z^{-\frac{t_c}{T}}}{1-z^{-1}}\right]F_n(z) \tag{8.62}$$

This is a simple calculation adjustment to make for Bode analysis purposes. A more complicated process is involved to insert this fractional delay in a simulation. The technique is illustrated in Figure 8.58.

The loop filter transfer functions, $F_n(z)$, denoting filter orders, $n = 1$ to 3 that exclude the NCO as shown in Figure 8.52, are as follows:

$$F_1(z) = \omega_0 \tag{8.63}$$



NOTATION

t_c = Computational Delay at NCO input

$t_i \equiv i*T$, i an integer

$\theta_{i+c} \equiv \theta(t_i + t_c)$

$\theta_{i+1} \equiv \theta(t_{i+1})$

$\omega_{i-1+c} \equiv \omega(t_{i-1} + t_c)$

$$\theta_{i+c} = \theta_i + \omega_{i-1+c} t_c$$

NCO output: $\theta_{i+1} = \theta_{i+c} + \omega_{i+c}(T - t_c)$

$= \theta_i + \omega_{i-1+c} t_C + \omega_{i+c}(T - t_C)$

Normalized Predetection Filter output: $\phi_{i+1} = \frac{1}{T}\left[\left(\frac{\theta_i + \theta_{i+c}}{2}\right)t_c + \left(\frac{\theta_{i+c} + \theta_{i+1}}{2}\right)(T - t_c)\right]$

$= 0.5\left[\theta_i\left(1 + \frac{t_c}{T}\right) + \theta_{i+1}\left(1 - \frac{t_c}{T}\right) + \omega_{i-1+c} t_c\right]$

Figure 8.58 Computation delay (t_c): NCO model with NCO and predetection filter output equations.

$$F_2(z) = a_2\omega_0 + \frac{\omega_0^2 T}{2}\left(\frac{1+z^{-1}}{1-z^{-1}}\right) \tag{8.64}$$

$$F_3(z) = b_3\omega_0 - \frac{a_3\omega_0^2 T}{2} + \frac{\omega_0^3 T^2}{4} + \frac{a_3\omega_0^2 T - \omega_0^3 T^2}{1-z^{-1}} + \frac{\omega_0^3 T^2}{\left(1-z^{-1}\right)^2} \tag{8.65}$$

The analyzed second-order filter of (8.64) uses values obtained from Table 8.23 of $\omega_0 = \frac{B_n}{0.53}$ and $a_2 = 1.414 = 2\delta$, where the damping factor $\delta = 0.707$, provides the fastest recovery time to a step function input with minimal initial overshoot response. The analyzed third-order filter of (8.65) uses values obtained from Table 8.23 of $\omega_0 = \frac{B_n}{0.7845}$, $a_3 = 1.1$ and $b_3 = 2.4$ that provide the fastest recovery time to a step function input with minimal overshoot response.

Figure 8.58 depicts the general case model of the NCO output phase due to the effect of computation delay when $t_c < T$. The effect is that the latency of the NCO updates results in part of the previous sample phase ramp extending into the new NCO phase ramp. The equations that account for this latency are shown for both the output of the NCO and the output of the predetection filter.

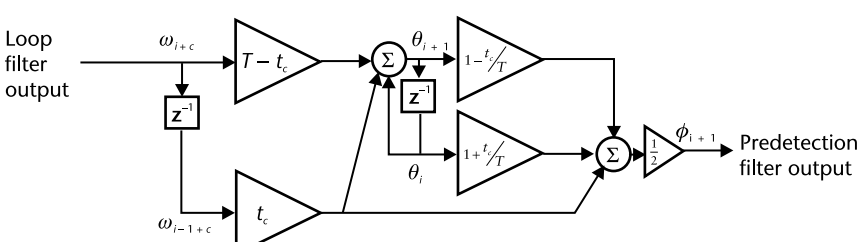
Figure 8.59 shows the model for the NCO and predetection filter that experience computation delay t_c. This model is used in digital simulations in the z-domain to account for this latency. The output equations for the NCO and the predetection filter are also shown.

The Bode analysis technique using the PLL open loop transfer function G(z) sets $z = e^{j\omega T}$, then finds the frequency, ω_{unity} where the absolute gain value is unity as described in the following equation:

$$\left|G\left(z \to e^{j\omega_{unity}T}\right)\right| = 1 \tag{8.66}$$

Using this value of , the phase margin is given by the following equation:

$$Phase\_Margin = angle\left(G\left(z \to e^{j\omega_{unit}T}\right)\right) - 180\,\text{deg} \tag{8.67}$$



$$\theta_{i+1} = \theta_i + \omega_{i-1+c}t_c + \omega_{i+c}\left(T - t_c\right)$$

$$\phi_{i+1} = 0.5\left[\theta_i\left(1 + \frac{t_c}{T}\right) + \theta_{i+1}\left(1 - \frac{t_c}{T}\right) + \omega_{i-1+c}\left(t_c\right)\right]$$

Figure 8.59 NCO and predetection filter simulation model with t_c computation delay.

The Bode analysis requires a program that solves (8.66) and (8.67). These results are used to plot the Bode *Phase_Margin* defined in (8.67) as a function of the predetection filter time, T, for a given noise bandwidth, B_n. Assuming a 30° phase margin design criteria, the 30° crossing point determines the corresponding maximum T to maintain this margin. The 0° crossing point determines T at the unstable threshold. Even though only one such plot is required because B_nT is approximately constant for small T and the same loop filter order, three plots are provided to validate this approximation. Figure 8.60 shows the Bode phase margin plots for B_n = 1 Hz for all three loop filter orders. For this case, the T reading in seconds at the 30° phase margin crossing is the B_nT dimensionless value desired. Similarly, Figure 8.61 shows these plots for $B_n = 2$ Hz and Figure 8.62 for $B_n = 4$ Hz. All three cases are for an on-time computation delay or effectively $t_c = 0$.

Table 8.24 summarizes the B_nT values for first-, second-, and third-order loops for 0° and 30° phase margins with computation delays of $t_c = 0$ and $t_c = T$. Figures 8.63, 8.64, and 8.65 provide quick B_n and T combination approximations for stable first-, second-, and third-order loops, respectively, using the Table 8.24 entries for 30° phase margin and for the extremes of computation delay between 0 and T.

Using closed loop simulations, Figures 8.66 and 8.67 validate the Bode analysis of the second- and third-order tracking loops, respectively, by demonstrating they remain stable if operated with a small amount of phase margin, but go unstable when the zero phase margin boundary is crossed.

The Bode analysis clearly demonstrates the deteriorating effect of computation delay on loop filter stability. It also shows some surprising results. For example, the third-order loop has about the same B_nT in the 30° phase margin region as the first-order loop and both are much better than the second-order loop.

Keep in mind that the Bode analysis technique does not predict the behavior when the tracking loops become nonlinear where longer T can be beneficial. It is prudent to use Monte Carlo simulations that model the nonlinear behavior of the tracking loops to fine-tune the value of T for a desired B_n in the presence of expected operating conditions that have caused the nonlinear conditions.

A loop filter parameter design example will clarify the use of the equations in Table 8.23 along with the B_nT values from Table 8.24. Suppose that the receiver

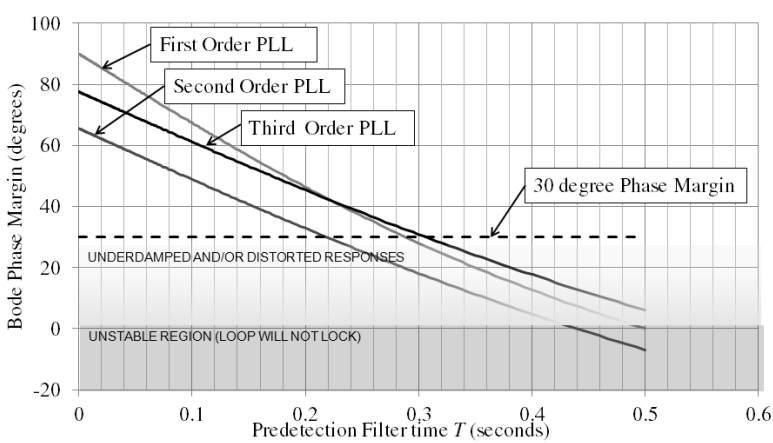


Figure 8.60 Bode phase margin plots for $B_n = 1$ Hz and t_c = on time.

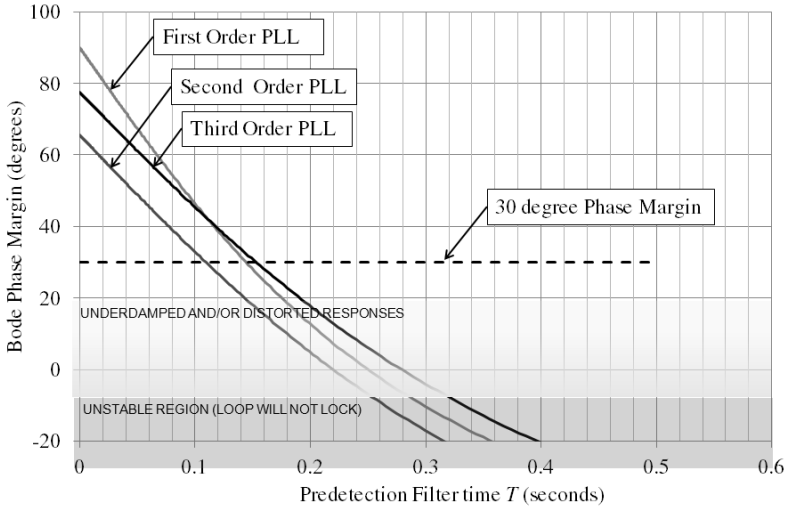


Figure 8.61 Bode phase margin plots for $B_n = 2$ Hz and $t_c =$ on time.

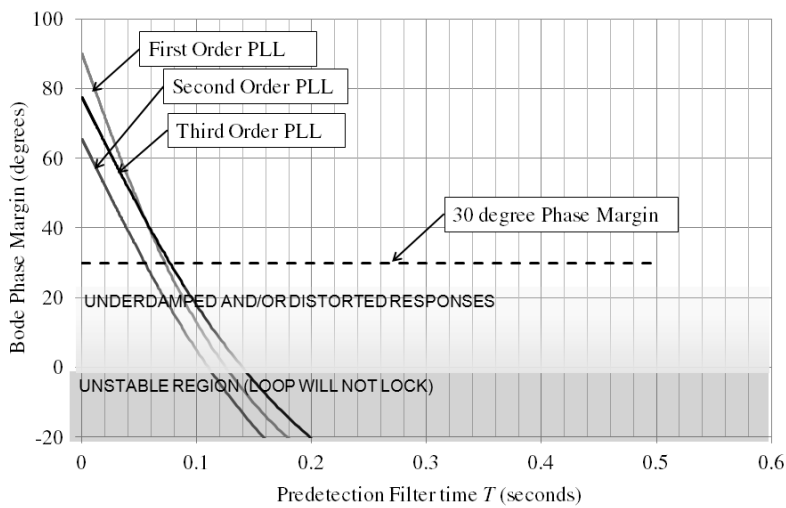


Figure 8.62 Bode phase margin plots for $B_n = 4$ Hz and $t_c =$ on time.

carrier-tracking loop will be subjected to high acceleration dynamics and will not be aided by an external navigation system, but must maintain PLL operation. A third-order loop is selected because it is insensitive to acceleration stress. To minimize its sensitivity to jerk stress, the noise bandwidth, B_n, is chosen to be the widest possible consistent with stability. In order to allow more loop filter computation time and to keep the NCO updates phased exactly on the T boundaries, a computation delay of exactly T is chosen. In this manner, after the slow functions computations are completed the values are buffered and then latched into the NCOs at exactly T second intervals. Table 8.24 specifies that $B_nT = 0.146$ for the third-order loop and T delay in the feedback path. Further assume that a steady state value of $T = 10$ ms will be used to match the symbol period. Then $B_n = 0.146/0.01 = 14.6$ Hz. Rounding this up to $B_n = 15$ Hz produces a $B_nT = 0.15$ that is sufficiently less than the

Table 8.24 First, Second-, and Third-Order Loops B_nT Values for 0° and 30° Phase Margins

Phase Margin (degrees)	*B_nT values (dimensionless)*					
	First-order loop		*Second-order loop*		*Third-order loop*	
	On time (s)	*T delay (s)*	*On time (s)*	*T delay (s)*	*On time (s)*	*T delay (s)*
0	0.500	0.207	0.440	0.201	0.558	0.245
30	0.289	0.134	0.218	0.107	0.306	0.146

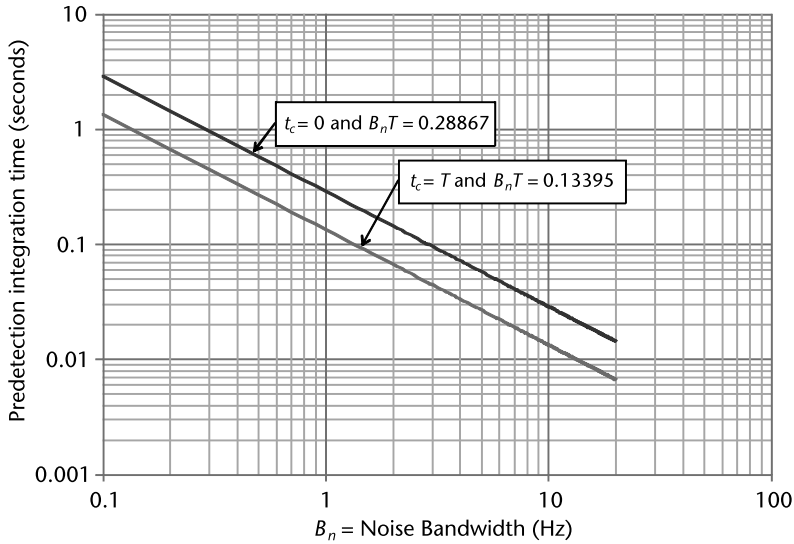


Figure 8.63 First-order DLL B_nT plots for 30° phase margin.

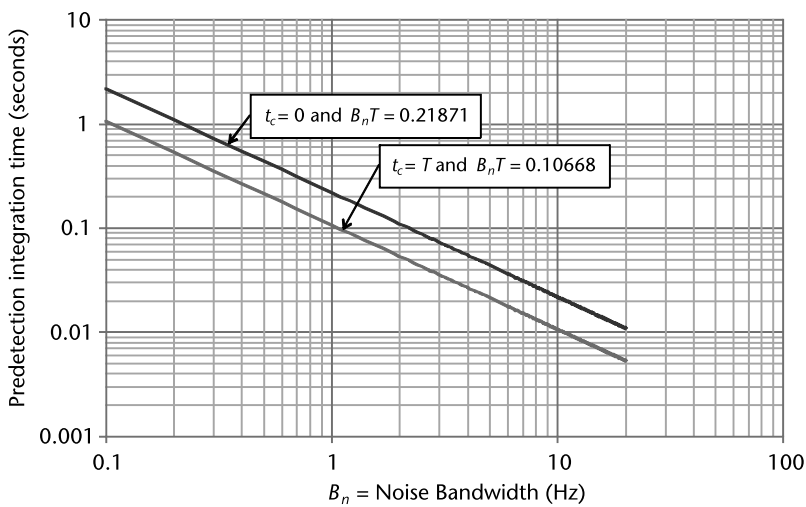


Figure 8.64 Second-order PLL B_nT plots for 30° phase margin.

unstable value of 0.245. The third-order PLL natural frequency from Table 8.23 is $\omega_0 = B_n / 0.7845 = 19.12$. The three multipliers shown in Figure 8.54(c) are computed using the a_3 and b_3 parameters from Table 8.23 as follows: $\omega_0^3 = 6989.78$,

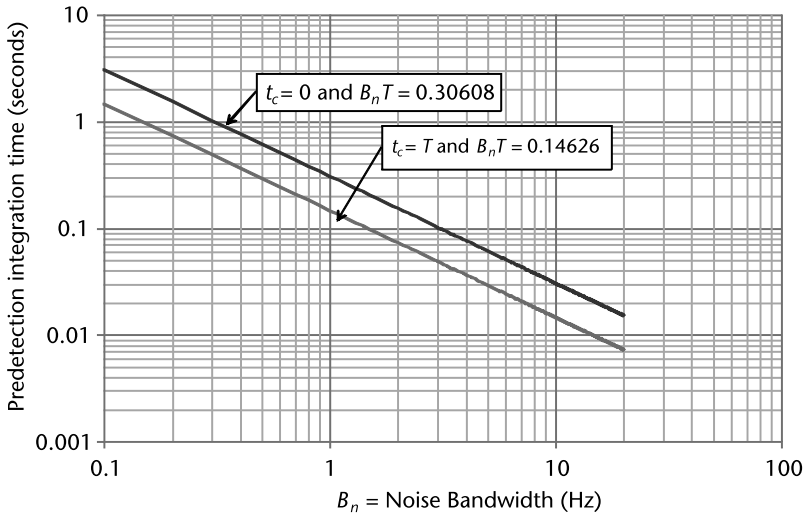


Figure 8.65 Third-order PLL B_nT plots for 30° phase margin.

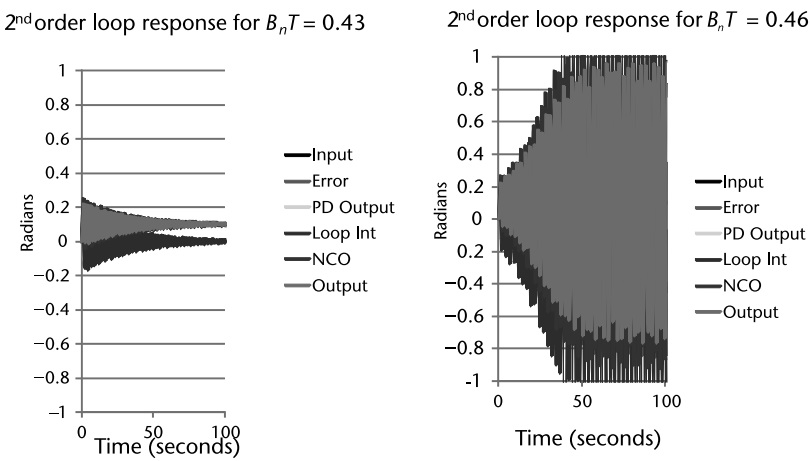


Figure 8.66 Second-order loop responses: stable (left), unstable (right).

$a_3\omega_0^2 = 1.1\omega_0^2 = 365.57$, $b_3\omega_0 = 2.4\omega_0 = 45.89$. This completes the third-order filter parameter design example. The remainder of the loop filter design is the implementation of the digital integrator accumulators to ensure that they will never overflow (i.e., that they have adequate dynamic range). The use of floating point arithmetic in modern microprocessors with built-in floating-point hardware greatly simplifies this part of the design process. Note that the velocity accumulator in the third order PLL of Figure 8.55(b) contains the loop filter estimate of line-of-sight velocity between the receiver and SV antenna phase centers. This estimate includes a self-adjusting bias component that compensates the carrier-tracking loop for the reference oscillator frequency error (i.e., the time bias rate error that is in common with all tracking channels). Similarly, the acceleration accumulator contains the loop filter estimate of line-of-sight acceleration that includes a self-adjusting bias component that compensates the carrier-tracking loop for the time rate of change

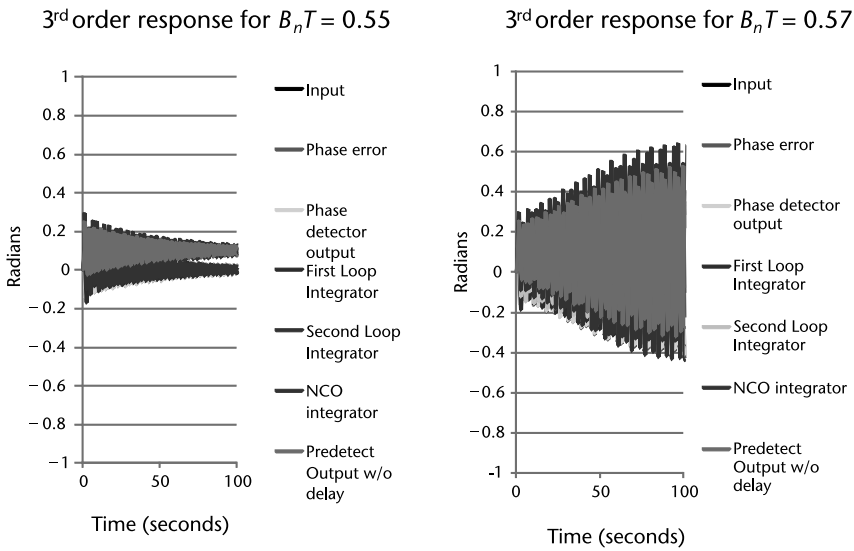


Figure 8.67 Third-order loop responses: stable (left), unstable (right).

of the reference oscillator frequency error. These accumulators should be initialized to zero just before initial loop closure unless good estimates of the correct values are known a priori. Also, they should be reset to their bias components (as learned by the navigation process) or to zero if unknown at the exact instance of injecting external carrier velocity aiding into the closed loop.

8.9 Measurement Errors and Tracking Thresholds

GNSS receiver measurement errors and tracking thresholds are closely related because the receiver loses lock when the measurement errors exceed a certain boundary. Because the code and carrier tracking loops are nonlinear, especially near the threshold regions, only Monte Carlo simulations of the GNSS receiver under the combined dynamic and signal-to-noise ratio conditions will determine the true tracking performance [73, 74]. However, rules of thumb can be used based on closed-form equations that approximate the measurement errors of the tracking loops. Numerous sources of measurement errors apply to each type of tracking loop. However, it is sufficient for rule-of-thumb tracking thresholds to analyze only the dominant error sources.

8.9.1 PLL Tracking Loop Measurement Errors

The dominant sources of phase error in a GNSS receiver PLL are thermal noise, oscillator imperfections, and dynamic stress. A conservative rule-of-thumb tracking threshold is that the overall 3-sigma error must not exceed 1/4 of the phase pull-in range of the PLL discriminator. The arctangent carrier phase discriminators are used as the basis for PLL pull-in range. As an example, for a pilot channel using a four-quadrant arctangent discriminator (PLL_P) whose phase pull-in range is 360°, the 3-sigma rule-of-thumb threshold is 90°. If there is data modulation, the PLL_D

two-quadrant arctangent discriminator must be used that has a phase pull-in range of 180° and the 3-sigma rule-of-thumb threshold is 45°. The PLL rule-of-thumb thresholds are then stated as:

$$\begin{aligned} 3\sigma_{PLL_P} &= 3\sigma_j + \theta_e \le 90\,\text{deg} \quad \text{(four-quadrant pilot)} \\ 3\sigma_{PLL_D} &= 3\sigma_j + \theta_e \le 45\,\text{deg} \quad \text{(two-quadrant data)} \end{aligned} \tag{8.68}$$

where σ_j = 1-sigma phase error from all sources except dynamic stress error and θ_e = dynamic stress error in the PLL tracking loop.

Equation (8.68) implies that dynamic stress error is a 3-sigma effect and is additive to the phase error. The phase error is the root sum square of every source of uncorrelated phase error such as thermal noise and oscillator noise. Oscillator noise includes vibration induced error and Allan deviation induced error. It also includes SV oscillator phase noise that has historically been so small as to be negligible. For example, IS-GPS-200 [67] specifies for the GPS C/A code and P(Y) code signals that "The phase noise spectral density of the unmodulated carrier shall be such that a phase locked loop of 10 Hz one-sided noise bandwidth shall be able to track the carrier to an accuracy of 0.1 radians rms." However, operational GPS SVs exhibit about an order of magnitude lower error than 0.1-radian (5.7°) rms to date and other GNSS SVs are similar in phase noise performance. This external source of noise error is not included in the foregoing analysis, but should be considered in very narrowband PLL applications.

Expanding on (8.68), the 1-sigma rule-of-thumb threshold for the PLL tracking loop for the two-quadrant arc-tangent discriminator is therefore:

$$\begin{aligned} \sigma_{PLL_P} &= \sqrt{\sigma_{tPLL_P}^2 + \sigma_v^2 + \theta_A^2} + \frac{\theta_e}{3} \le 30\,\text{deg} \\ \sigma_{PLL_D} &= \sqrt{\sigma_{tPLL_D}^2 + \sigma_v^2 + \theta_A^2} + \frac{\theta_e}{3} \le 15\,\text{deg} \end{aligned} \tag{8.69}$$

where σ_{tPLL_P} = Pilot PLL 1-sigma thermal noise in degrees, σ_{tPLL_D} = Data PLL 1-sigma thermal noise in degrees, σ_v = 1-sigma vibration induced oscillator error in degrees, and θ_A = Allan variance induced oscillator error in degrees.

8.9.2 PLL Thermal Noise

Often the PLL thermal noise is treated as the only source of carrier tracking error, since the other sources of PLL error may be either transient or negligible. The thermal noise error for PLL is computed as follows:

$$\begin{aligned} \sigma_{tPLL_P} &= \frac{360}{2\pi}\sqrt{\frac{B_n}{C/N_0}} \\ \sigma_{tPLL_D} &= \frac{360}{2\pi}\sqrt{\frac{B_n}{C/N_0}\left(1 + \frac{1}{2TC/N_0}\right)} \end{aligned} \quad \text{(deg)} \tag{8.70}$$

$$\sigma_{tPLL_P} = \frac{\lambda_L}{2\pi}\sqrt{\frac{B_n}{C/N_0}}$$
$$\sigma_{tPLL_D} = \frac{\lambda_L}{2\pi}\sqrt{\frac{B_n}{C/N_0}\left(1+\frac{1}{2TC/N_0}\right)} \quad \text{(m)} \tag{8.71}$$

where

σ_{tPLL_P} = 1-sigma error in a pilot channel PLL

σ_{tPLL_D} = 1-sigma error in a data channel PLL

B_n = carrier loop noise bandwidth (Hz)

C/N_0 = ratio of carrier power to noise power in a 1-Hz bandwidth (Hz)

T = predetection integration time (s)

λ_L = GPS L-band carrier wavelength (m)

= (299792458 m/s)/(1575.42 MHz) = 0.190293673 m/cycle for L1

= (299792458 m/s)/(1227.60 MHz) = 0.244210213 m/cycle for L2

= (299792458 m/s)/(1176.45 MHz) = 0.254828049 m/cycle for L5

$\frac{1}{2TC/N_0}$ = squaring loss in data channel

Clearly, the pilot channel offers two important advantages: (8.69) shows that the PLL tracking threshold is double its data channel counterpart and (8.70) shows that there is no squaring loss in the pilot channel. In fact, if the pilot channel and data channel carrier power is split 50/50 (as is the case with some GNSS signals), both lose 3 dB of power, but the pilot channel gains 6 dB of threshold (because the threshold range has doubled) for a net gain of 3 dB plus additional threshold improvement under low $(C/N_0)_{dB}$ conditions due to the absence of squaring loss. Note that none of the above equations include variables that relate to the underlying PRN code or the loop filter order. Also note that (8.70) is independent of carrier frequency because the error is expressed in units of degrees. The carrier thermal noise error standard deviation is strictly dependent on the carrier–to-noise power ratio, C/N_0, the noise bandwidth, B_n, and, in the case of a data channel (Costas PLL), the predetection integration time, T. The carrier-to-noise power ratio, C/N_0, is an important factor in many GNSS receiver performance measures. It is computed as the ratio of recovered power, C (in W), from the desired signal to the noise density N_0 in a 1-Hz noise bandwidth (in W/Hz). The piecewise equations for determining C/N_0 (expressed as $(C/N_0)_{dB}$ in units of dB-Hz) are described in Chapter 9. The standard deviation decreases if the $(C/N_0)_{dB}$ increases (e.g., the recovered signal power is increased or the noise level is decreased). Decreasing the noise bandwidth reduces the standard deviation. Increasing the predetection integration time in a Costas loop reduces the squaring loss, which, in turn, decreases the standard deviation.

It is a common misconception that some GNSS signals always produce more accurate carrier phase measurements than others in thermal noise owing to their modulation techniques. While signals with BOC modulation and/or overlay codes

have the potential for better code tracking accuracy and multipath error mitigation, the PLL thermal noise error is identical for any spreading waveform for the same $(C/N_0)_{dB}$ since PLL processing uses quantities after the spreading code has been stripped off. It is the received carrier power that makes the difference in PLL thermal noise error and if carrier smoothed code techniques or real-time kinematic (differential interferometric) techniques are used, the ultimate precision is obtained from the carrier tracking loop. Therefore, the most significant accuracy benefit of modernized signal designs is multipath mitigation.

It is another common misconception that the carrier loop measurement is a velocity measurement, when (8.71) shows that it is actually a range measurement (albeit an ambiguous one). The PLL thermal noise error is in units of range because it is part of the carrier Doppler phase measurement. As examples, a pilot channel PLL provides range measurements that are very precise within one wavelength, but the integer number of remaining wavelengths to the SV is unknown, while a data channel (Costas) PLL provides range measurements that are very precise within a half-wavelength, but the integer number of remaining half-wavelengths to the SV is unknown. The velocity is approximated using the change in carrier Doppler phase between two carrier range measurements over a short time. The carrier Doppler phase measurement must include an ambiguous count of integer wavelengths or half wavelengths. When a velocity measurement is made that takes the difference between two of these ambiguous range measurements and divides the result by the time interval, the ambiguity is removed.

Figure 8.68 illustrates the pilot channel and data channel (Costas) PLL thermal noise error plotted as a function of $(C/N_0)_{dB}$ for $B_n = 15$ Hz and 2 Hz assuming $T = 10$ ms for the Costas PLL. Note that even though the pilot PLL does not manifest squaring loss due to T, its loop stability is affected. Likewise, neither the pilot PLL nor the Costas PLL thermal noise error is affected by the loop filter order, but both are subject to the same B_nT rules for loop stability. For the case examples of Figure

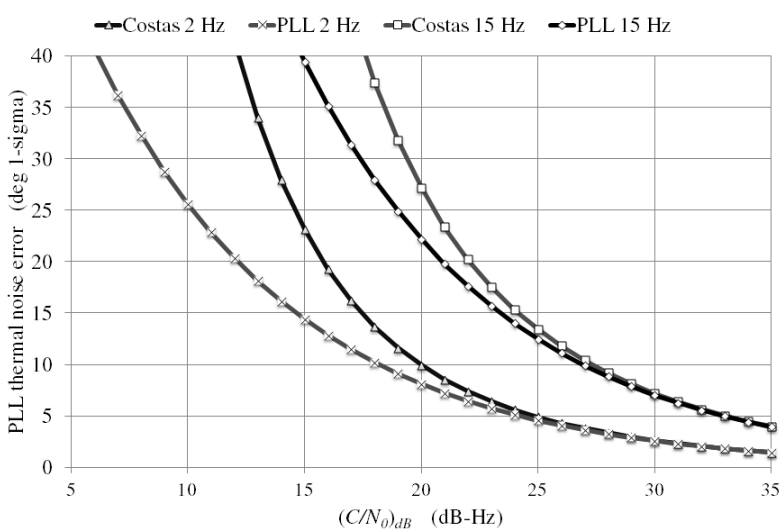


Figure 8.68 Pilot and data (Costas) PLL thermal noise error for Costas T = 10 ms.

8.68, the pilot PLL and Costas PLL remain stable for second- and third-order PLLs operating at $B_n = 15$ Hz and $T = 10$ ms ($B_nT = 0.15$), but the second-order PLL has less margin. The loop order dictates sensitivity to the same order of dynamics (first order to velocity stress, second order to acceleration stress, and third order to jerk stress) and the loop bandwidth must be wide enough to accommodate these higher-order dynamics. In general, when the loop order is made higher, there is an improvement in dynamic stress performance. Thus, the thermal noise can be reduced for the same minimum $(C/N_0)_{dB}$ by increasing the loop order and reducing the noise bandwidth while also improving the dynamic performance.

8.9.3 Vibration-Induced Oscillator Phase Noise

Vibration-induced oscillator phase noise is a complex analysis problem. In some cases, the expected vibration environment is so severe that the reference oscillator must be mounted using vibration isolators in order for the GNSS receiver to successfully operate a PLL. The equation for vibration induced oscillator error is:

$$\sigma_v = \frac{360 f_L}{2\pi} \sqrt{\int_{f_{\min}}^{f_{\max}} S_v^2(f_m) \frac{P(f_m)}{f_m^2} df_m} \quad \text{(deg)} \tag{8.72}$$

where

f_L = L- band input frequency in Hz;

$S_v(f_m)$ = oscillator vibration sensitivity of $\Delta f/f_L$ per g as a function of f_m where g is the acceleration due to gravity ≈ 9.8 m/s^2;

f_m = random vibration modulation frequency in Hz;

$P(f_m)$ = power curve of the random vibration in g^2/Hz as a function of g.

If the oscillator vibration sensitivity, $S_v(f_m)$, is not variable over the range of the random vibration modulation frequency, f_m, then (8.72) can be simplified to:

$$\sigma_v = \frac{360 f_L S_v}{2\pi} \sqrt{\int_{f_{\min}}^{f_{\max}} \frac{P(f_m)}{f_m^2} df_m} \quad \text{(deg)} \tag{8.73}$$

As a simple computational example, assume that the random vibration power curve is flat from 20 Hz to 2,000 Hz with an amplitude of 0.005 g^2/Hz. If $S_v = 1 \times 10^{-9}$ parts/g and f_L = L1 = 1,575.42 MHz, then the vibration-induced phase error using (8.73) is:

$$\sigma_v = 90.265 \sqrt{0.005 \int_{20}^{2000} \frac{df_m}{f_m^2}} = 90.265 \sqrt{0.005 \left(\frac{1}{20} - \frac{1}{2000} \right)} = 1.42° \text{ at L1}$$

8.9.4 Allan Deviation Oscillator Phase Noise

There are several stability measures for frequency sources. Allan variance is one suitable metric for analyzing the short term stability of GNSS receiver reference oscillators. The square root of Allan variance is referred to as Allan deviation that manifests itself in PLLs as phase error. The equations used to determine Allan deviation phase error are empirical. The equations are stated in terms of what the requirements are for the short-term stability of the reference oscillator as determined by the Allan variance method of stability measurement. The equation for short term Allan deviation for a second-order PLL is [75]:

$$\sigma_A(\tau) = 2.5\frac{\Delta\theta}{\omega_L\tau} \text{ (dimensionless units of } \Delta f/f) \tag{8.74}$$

where $\Delta\theta$ = rms error into phase discriminator due to the oscillator (rad), ω_L = L-band input frequency = $2\pi f_L$ (rad/s), and τ = short-term stability gate time for Allan variance measurement (s).

The equation for a third-order PLL is similar [75]:

$$\sigma_A(\tau) = 2.25\frac{\Delta\theta}{\omega_L\tau} \text{ (dimensionless units of } \Delta f/f) \tag{8.75}$$

If the Allan variance, $\sigma_A^2(\tau)$, has already been determined for the oscillator for the short-term gate time, τ, then the Allan deviation induced error in deg, $\theta_A = 360\Delta\theta/2\pi$, can be computed from the above equations. Usually $\sigma_A^2(\tau)$ changes very little for the short-term gate times involved. These gate times must include the reciprocal of the range of noise bandwidths used in the carrier loop filters, $\tau = 1/B_n$. A short-term gate-time range of 5 ms to 1,000 ms should suffice for all PLL applications. Rearranging (8.74) using these assumptions, the equation for the second-order loop is:

$$\theta_{A2} = 144\frac{\sigma_A(\tau)f_L}{B_n} \text{ (deg)} \tag{8.76}$$

Rearranging (8.75) using these assumptions, the equation for the third-order loop is:

$$\theta_{A3} = 160\frac{\sigma_A(\tau)f_L}{B_n} \text{ (deg)} \tag{8.77}$$

For example, assume that the loop filter is third-order with a noise bandwidth, B_n = 15 Hz, tracking the L1 signal, and the Allan deviation is specified to be $\sigma_A(\tau) = 100E - 10$ or better for gate times that include $\tau = 1/B_n = 67$ ms. The phase error contribution due to this source of error is $\theta_{A3} = 1.68°$ or less. Obviously, a reference

oscillator with a short-term Allan deviation characteristic that is more than an order of magnitude worse than this example will cause PLL tracking problems.

Figure 8.69 graphically portrays the sensitivity of a third-order PLL to changes in short-term Allan deviation performance of the reference oscillator, especially as the noise bandwidth, B_n, is narrowed. Intuitively, the designer attempts to improve the tracking threshold by reducing B_n expecting to reduce thermal noise error. However, as Figure 8.69 illustrates, the Allan deviation effect significantly increases its contribution to the error at narrower noise bandwidths. This effect is usually the primary source of aided GNSS receiver narrowband PLL tracking problems assuming that the external velocity aiding accuracy is not the limiting factor. As shown in Figure 8.69, a reference oscillator with an Allan deviation $\Delta f/f$ of less than 1.00E-09 will cause unreliable PLL operation in all circumstances. Therefore, the oscillator specification for Allan deviation is important for all GNSS receiver designs.

8.9.5 Dynamic Stress Error

The dynamic stress error is obtained from the steady state error formulas shown in Table 8.23. This error depends on the loop bandwidth and order. The maximum dynamic stress error may be slightly larger than the steady-state error if the loop filter response to a step function has overshoot, but the steady-state error formula usually suffices since the worst case (that the dynamic stress direction is in the line-of-sight to the SV) is assumed. There should be no more than about a 7% overshoot if the filter is designed for minimum mean square error, which is the case for the typical loop filter coefficients shown in the table. From Table 8.23, a second-order loop with minimum mean square error, the dynamic stress error is:

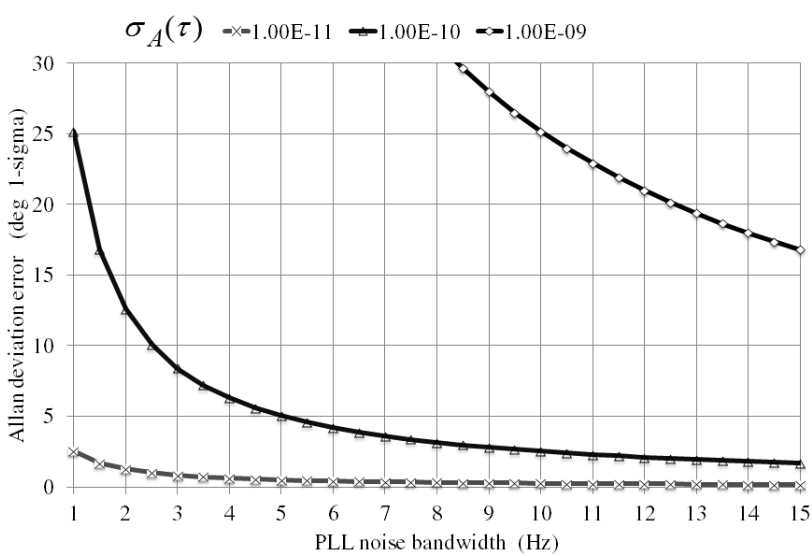


Figure 8.69 Allan deviation error in third-order PLL at L1.

$$\theta_{e2} = \frac{d^2 R/dt^2}{\omega_0^2} = \frac{d^2 R/dt^2}{\left(\frac{B_n}{0.53}\right)^2} = 0.2809\frac{d^2 R/dt^2}{B_n^2} \quad (\text{deg}) \tag{8.78}$$

where d^2R/dt^2 = maximum line-of-sight acceleration dynamics (deg/s^2).

From Table 8.23, a third-order loop with minimum mean square error, the dynamic stress error is defined as follows:

$$\theta_{e3} = \frac{d^3 R/dt^3}{\omega_0^3} = \frac{d^3 R/dt^3}{\left(\frac{B_n}{0.7845}\right)^3} = 0.4828\frac{d^3 R/dt^3}{B_n^3} \quad (\text{deg}) \tag{8.79}$$

where d^3R/dt^3 = maximum line-of-sight jerk dynamics (deg/s^3).

Note that (8.78) and (8.79) are 3-sigma errors. As an example of how this error is computed, suppose the third-order loop noise bandwidth is 15 Hz and the maximum line-of-sight jerk dynamic stress to the SV is 10 g/s = 98 m/s^3. To convert this to deg/s^3, multiply the jerk dynamics by the number of carrier wavelengths contained in 1m in units of deg/m. For L1, d^3R/dt^3= (98 m/s^3) × (360°/cycle) × (1,575.42×10^6 cycle/s)/ c = 185,398°/ s^3, where c = 299,792,458 m/s is the propagation velocity (speed of light). For L2, $d^3R/dt^3 = 98 \times 360 \times 1{,}227.60 \times 10^6/c$ = 144,466°/s^3. Using (8.79), the 3-sigma stress error for a 15-Hz third-order PLL is 26.52° for L1 and 20.67° for L2. These are well below the 90° PLL and 45° Costas PLL 3-sigma rule-of-thumb levels even though 10 g/s is a very high dynamic jerk stress level. However as observed in (8.79), for the same maximum jerk stress level the dynamic stress error increases by the reciprocal of noise bandwidth cubed, so narrow PLL bandwidths are highly vulnerable to jerk dynamic stress. This is the reason that external velocity aiding is used to enable narrow PLL bandwidths to continue to operate in the presence of dynamic stress.

8.9.6 Reference Oscillator Acceleration Stress Error

The PLL cannot tell the difference between the dynamic stress induced by real dynamics and the dynamic stress caused by changes in frequency in the reference oscillator due to acceleration sensitivity of the oscillator. The reference oscillator change in frequency due to dynamic stress is:

$$\Delta f_g = 360 S_g f_L G(t) \tag{8.80}$$

where S_g = g-sensitivity of the oscillator ($\Delta f/f$ per g), f_L = L-band input frequency (Hz), and $G(t)$ = acceleration stress in g as a function of time

When the units of Δf_g in (8.80) are deg/s, a velocity error is sensed by the loop filter as a result of the $G(t)$ component due to acceleration (g) sensed by the reference oscillator. For an unaided second-order carrier tracking loop, this acceleration induced oscillator error can be ignored because it is insensitive to velocity stress.

When the units of Δf_g are deg/s^2, an acceleration error as sensed by the loop filter as a result of the $G(t)$ component due to jerk stress (g/s). For an unaided third-order carrier tracking loop, this jerk induced oscillator error can be ignored because it is insensitive to acceleration stress. In reality, there will always be some level of dynamic stress that will adversely affect the tracking loop regardless of the loop filter order because there are always higher order components of dynamic stress when the host vehicle is subjected to dynamics. Nothing can be done about this for an unaided tracking loop except to align the least sensitive S_g axis of the reference oscillator along the direction of the anticipated maximum dynamic stress, but this is often impractical. For an externally aided tracking loop where the line-of-sight dynamic stress can be measured and S_g is known, it is prudent to model this acceleration stress sensitivity and apply the correction to the aiding. Note that, like all oscillator-induced errors, the error is common mode to all receiver tracking channels, so one correction applies to all aided channels.

8.9.7 Total PLL Tracking Loop Measurement Errors and Thresholds

Figure 8.70 illustrates the total PLL error as defined in (8.69) as a function of $(C/N_0)_{dB}$ for a third-order PLL including all effects described in (8.70), (8.73), (8.77), and (8.79) using a wide noise bandwidth of $B_n = 15$ Hz and a narrow noise bandwidth of $B_n = 2$ Hz, both with predetection integration times of $T = 10$ ms. The squaring loss for the Costas is apparent at the lower $(C/N_0)_{dB}$ levels. The value of T could have been increased at the narrower bandwidth consistent with the value of T that assures stability to improve tracking threshold, but the same values were used as in Figure 8.68 to show how much the other sources of error moved the curves to the left. Since the reference oscillator specification is for the highest performance device and the effective dynamic stress for the 2-Hz PLLs is extremely low assuming the highest performance external velocity aiding, there is not much movement due to the additional error contribution from these parameters. Many commercial GNSS receivers are unaided and use low-cost reference oscillators with Allan

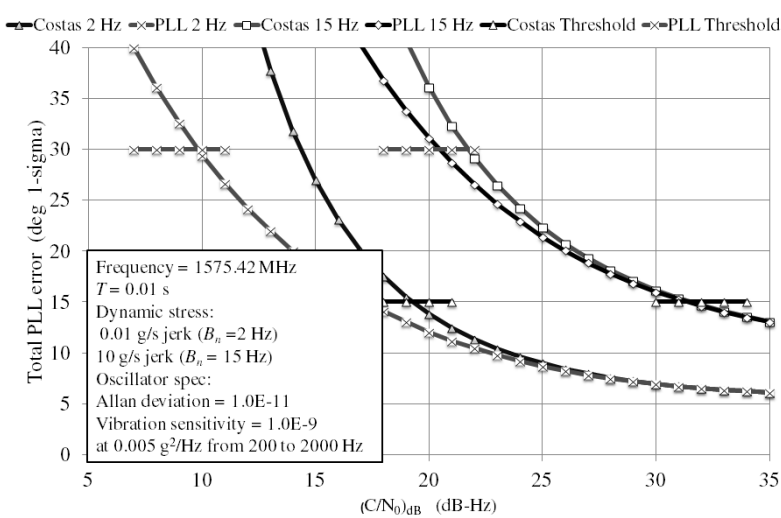


Figure 8.70 Total PLL error for third-order pilot and data (Costas) PLLs.

deviation performance of 1.00E-09 (or worse) with unspecified random vibration characteristics, requiring the use of second-order PLLs that operate at maximum stable bandwidths for the T required to demodulate data.

It is insightful to observe the dynamic range sensitivity as a function of B_n for the pilot and data channel PLL cases. This can be observed by rearranging (8.69) to solve for the dynamic stress error at threshold for a range of values of B_n with $(C/N_0)_{dB}$ as a running parameter:

$$\begin{aligned}\frac{\theta_e}{3} &= 30 - \sqrt{\sigma_{tPLL_P}^2 + \sigma_v^2 + \theta_A^2} \\ \frac{\theta_e}{3} &= 15 - \sqrt{\sigma_{tPLL_D}^2 + \sigma_v^2 + \theta_A^2}\end{aligned} \tag{8.81}$$

For example, (8.81) can be used to determine the maximum jerk stress (thresholds) as a function of B_n for third-order PLLs in dynamic stress units of jerk (g/s) as:

$$\begin{aligned}J_{P\theta e3\max} &= \frac{B_n^3}{2983.745784}\left(30 - \sqrt{\sigma_{tPLL_P}^2 + \sigma_v^2 + \theta_A^2}\right) \\ J_{D\theta e3\max} &= \frac{B_n^3}{2983.745784}\left(15 - \sqrt{\sigma_{tPLL_D}^2 + \sigma_v^2 + \theta_A^2}\right)\end{aligned} \quad \text{(g/s)}$$

All terms under the radical are computed in units of degrees for each value of B_n and with $(C/N_0)_{dB}$ held constant as a running parameter and the value of T used for the maximum value of B_n (consistent with the B_nT stability requirement). Figure 8.71 uses this technique to compare the jerk thresholds as a function of B_n for two third-order PLLs and two third-order Costas PLLs. The approximate $(C/N_0)_{dB}$ that corresponds to the 10 g thresholds for the Costas (data) PLL (32 dB-Hz) and for the pilot PLL (22.5 dB-Hz) were used for the running parameter $(C/N_0)_{dB}$ of two of the

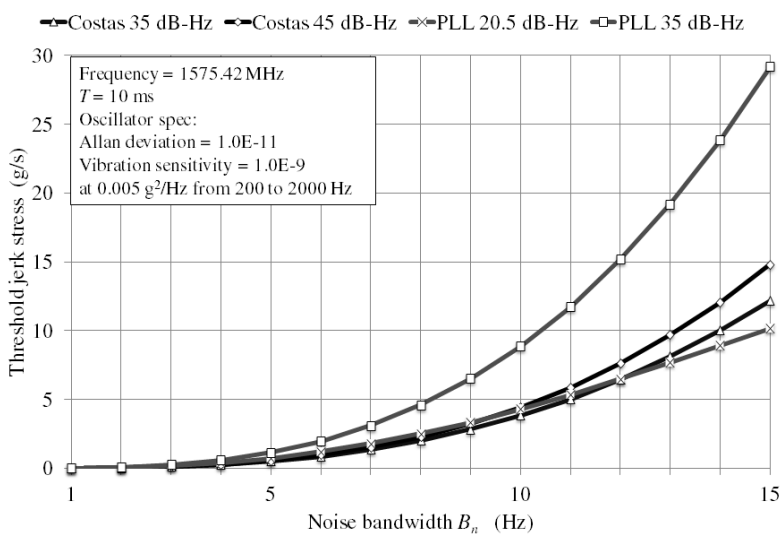


Figure 8.71 Jerk stress thresholds for third-order pilot and data (Costas) PLLs.

case examples to illustrate that they both meet at the same 10 g/s and B_n intercept in Figure 8.71, but the Costas PLL required 9.5 dB more carrier power than the pilot PLL to achieve the same dynamic stress threshold.

The two 2-Hz examples in Figure 8.70 would not show up in Figure 8.71 since their dynamic stress thresholds were only 0.01 g/s, so the other two examples in Figure 8.71 are at the same high $(C/N_0)_{dB}$ level (32 dB-Hz) to demonstrate that the pilot PLL always outperforms the Costas PLL dynamic stress threshold for the same T and will outperform it with an optimized Costas T that is consistent with loop stability.

Clearly, these PLL rule-of-thumb and approximation equations are excellent tools to quickly achieve the performance desired for a GNSS receiver design, but as a note of caution, the real-world thresholds are statistical in nature so there is randomness in the predicted threshold regions as to when the PLLs actually lose track [i.e., there are no "exact" thresholds, only statistical (rms) thresholds]. Also, worst-case assumptions are used for most of the contributing error components. Monte Carlo simulations are essential to the final design tune-up process. Representative operational simulations to more accurately predict GNSS receiver performance under the specified dynamic stress conditions should follow the Monte Carlo simulations.

8.9.8 FLL Tracking Loop Measurement Errors

The dominant sources of frequency error in a GNSS receiver FLL are thermal noise and dynamic stress. The rule-of-thumb tracking threshold is that the 3-sigma error must not exceed one fourth of the frequency pull-in range of the FLL discriminator. As observed in Figure 8.46 the four-quadrant ATAN2 FLL discriminator pull-in range is approximately $\pm 1/2T$ Hz. Therefore, the rule-of-thumb tracking threshold using this FLL discriminator is:

$$3\sigma_{FLL} = 3\sigma_{tFLL} + f_e \le 1/(4T) \tag{8.82}$$

where $3\sigma_{tFLL}$ = 3-sigma thermal noise frequency error and f_e = dynamic stress error in the FLL tracking loop.

The dynamic stress error in (8.82) is a 3-sigma effect and is additive to the thermal noise frequency error. The reference oscillator vibration and Allan deviation induced frequency errors are small order effects on the FLL and are considered negligible in every case where the FLL is robust (i.e., not ultranarrowband). The 1-sigma frequency error threshold would be 1/(12T) = 0.0833/T Hz.

The FLL tracking loop error due to thermal noise is:

$$\sigma_{tFLL} = \frac{1}{2\pi T}\sqrt{\frac{4FB_n}{C/N_0}\left[1+\frac{1}{TC/N_0}\right]} \ \text{(Hz)} \tag{8.83}$$

$$\sigma_{tFLL} = \frac{\lambda_L}{2\pi T}\sqrt{\frac{4FB_n}{C/N_0}\left[1+\frac{1}{TC/N_0}\right]} \ \text{(m/s)} \tag{8.84}$$

where

$F = 1$ at high C/N_0

$= 2$ near threshold

Note that there is no dependence on spreading code modulation design and loop order in (8.83). It is also independent of L-band carrier frequency if the error units are expressed in units of Hz.

Because the FLL tracking loop involves one more integrator than the PLL tracking loop of the same order n, the dynamic stress error is:

$$f_e = \frac{d}{dt}\left(\frac{1}{360\omega_0^n}\frac{d^n R}{dt^n}\right) = \frac{1}{360\omega_0^n}\frac{d^{n+1}R}{dt^{n+1}} \text{ (Hz)} \tag{8.85}$$

As an example of how the dynamic stress error is computed for a second-order loop ($n = 2$) using (8.85), assume the FLL design has a noise bandwidth $B_n = 2$ Hz and a predetection integration time $T = 5$ ms. From Table 8.23, $B_n = 0.53\ \omega_0$, so $\omega_0^2 = (2/0.53)^2 = 14.24$ Hz. If the maximum line-of-sight jerk dynamics is 10 *g*/s = 98 m/s^3, then this translates into $d^3R/dt^3 = 98 \times 360 \times 1{,}575.42 \times 10^6/c = 185{,}398°/s^3$ for L1. Substituting these numbers into (8.85) results in a maximum dynamic stress error of $f_e = 185{,}398/(14.24 \times 360) = 36$ Hz. Since the rule-of-thumb 3-sigma threshold is $1/(4\times 0.005) = 50$ Hz, the FLL noise bandwidth is acceptable for the 10 g/s level of maximum jerk dynamic stress. Figure 8.72 illustrates the FLL thermal noise tracking error and tracking thresholds assuming a second-order loop under 10 *g*/s jerk dynamics with typical noise bandwidths and predetection integration times.

Figure 8.73 illustrates the jerk stress thresholds for a second-order FLL using two samples at $T = 5$-ms intervals to form the FLL discriminator error every 10 ms. The FLL threshold jerk stress is plotted as a function of noise bandwidth B_n with C/N_0 as a running parameter using a technique similar to the one used for Figure

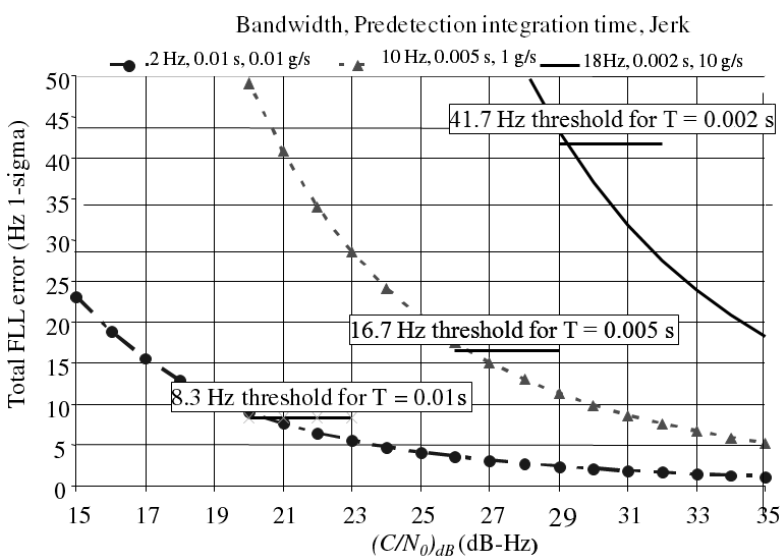


Figure 8.72 Total FLL error for second-order carrier loop.

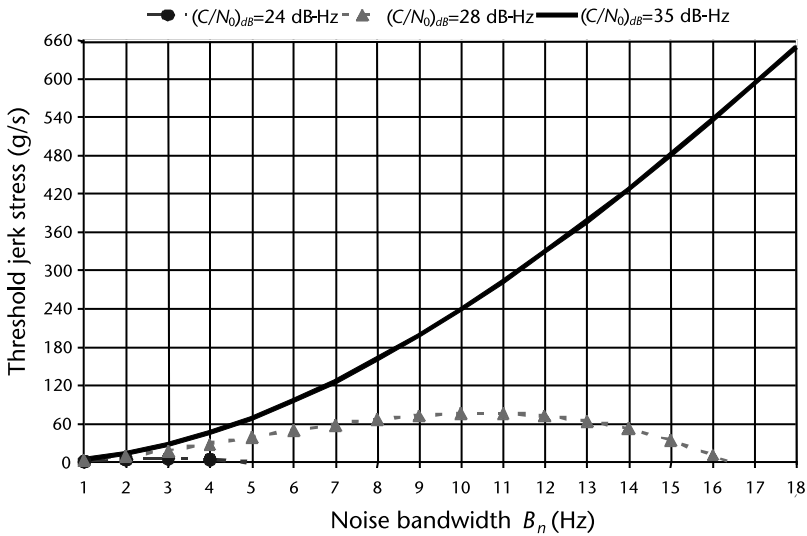


Figure 8.73 Jerk stress thresholds for second-order FLL with T = 5 ms.

8.71. Comparing the jerk thresholds in Figure 8.73 for the second-order FLL with those of Figure 8.71 for third-order PLLs, notice that the FLL has much better dynamic stress performance. For example compare the same running parameter $(C/N_0)_{dB}$ = 35 dB-Hz in Figure 8.73 at B_n =10 Hz where the FLL can tolerate up to 240 *g*/s while the Costas PLL at can only tolerate up to about 4 *g*/s and the pilot PLL can tolerate about 8 g/s. The spread is much smaller for weaker signal strengths and lower noise bandwidths. Both PLLs would have performed moderately better under dynamic stress if the predetection integration time had been reduced from 10 ms to 5 ms, as was the case for the FLL (even though its effective predetection bandwidth is actually 10 ms). This comparison reinforces the earlier statements that a robust GNSS receiver design uses an FLL as a backup to the PLL during initial loop closure and during high dynamic stress with loss of phase lock, but will revert to pure PLL for the steady state low to moderate dynamics in order to produce the highest accuracy carrier Doppler phase measurements. Also note that the maximum predetection integration time, *T,* for FLL in a data channel is half the bit rate or sample rate since two *T* samples within transition boundaries are required for the FLL discriminator. Very short values of *T* are used prior to the receiver learning where the data transitions are located to minimize the percentage of corrupted discriminator values.

8.9.9 Code-Tracking Loop Measurement Errors

When there is no multipath or other distortion of the received signal, and no interference, the dominant sources of range error in a GNSS receiver code tracking loop, usually called a delay lock loop (DLL), are thermal noise and dynamic stress. The rule-of-thumb tracking threshold for the DLL is that the 3-sigma value of the error due to all sources of loop stress must not exceed half of the linear pull-in range of the DLL discriminator. Therefore, the rule-of-thumb tracking threshold is:

$$3\sigma_{DLL} = 3\sigma_{tDLL} + R_e \le D/2 \quad \text{(chips)} \tag{8.86}$$

where σ_{tDLL} = 1-sigma thermal noise code tracking error (chips), R_e = dynamic stress error in the DLL tracking loop (chips), and D = early-to-late correlator spacing (chips).

A general expression for thermal noise code tracking error for a noncoherent DLL discriminator is [76]:

$$\sigma_{tDLL} \cong \frac{1}{T_c}\sqrt{\frac{B_n\int_{-B_{fe}/2}^{B_{fe}/2} S_S(f)\sin^2(\pi f D T_c)\,df}{(2\pi)^2\, C/N_0\left(\int_{-B_{fe}/2}^{B_{fe}/2} f S_S(f)\sin(\pi f D T_c)\,df\right)^2}} \times\sqrt{\left[1+\frac{\int_{-B_{fe}/2}^{B_{fe}/2} S_S(f)\cos^2(\pi f D T_c)\,df}{TC/N_0\left(\int_{-B_{fe}/2}^{B_{fe}/2} S_S(f)\cos(\pi f D T_c)\,df\right)^2}\right]} \tag{8.87}$$

where

B_n = code loop noise bandwidth (Hz);

$S_S(f)$= power spectral density of the signal, normalized to unit area over infinite bandwidth;

B_{fe} = double-sided front-end bandwidth (Hz);

T_c = chip period (s/chip) = $1/R_c$ where R_c is the spreading code rate.

As D becomes vanishingly small the trigonometric functions in (8.87) can be replaced by their first-order Taylor series expansions about zero, and this equation becomes:

$$\sigma_{tDLL} \cong \frac{1}{T_c}\sqrt{\frac{B_n}{(2\pi)^2\,(C/N_0)\int_{-B_{fe}/2}^{B_{fe}/2} f^2 S_S(f)\,df}\left[1+\frac{1}{T(C/N_0)\int_{-B_{fe}/2}^{B_{fe}/2} S_S(f)\,df}\right]} \quad \text{(chips)} \tag{8.88}$$

The term $\sqrt{\int_{-B_{fe}/2}^{B_{fe}/2} f^2 S_S(f)\,df}$ is called the root-mean-squared (rms) bandwidth of the signal, and is a measure of the sharpness of the correlation peak. Clearly, signals with larger rms bandwidths offer the potential for more accurate code tracking. In fact, the frequency-squared term in the rms bandwidth indicates that even very small amounts of high-frequency content in the signal can enable more accurate

code tracking if there is also a corresponding reduction in the early minus late correlator spacing, D. Intuitively, these high-frequency components produce sharper edges and more distinct zero crossings in the waveform, enabling more accurate code tracking.

The use of carrier-aided code (practically a universal design practice) effectively removes the code dynamics, so the use of narrow correlators (along with increasing the front-end bandwidth) is an excellent design trade-off for receivers using signals with relatively slow spreading symbol rates. For such signals, reducing the correlator spacing reduces the effects of thermal noise and multipath (see Section 9.5), but this also requires increasing the receiver front-end bandwidth that increases the vulnerability to in-band RF interference.

For a BPSK-R(n) modulation such as the GPS P(Y) code ($n = 10$), L5 ($n = 10$), C/A code ($n = 1$), or L2C ($n = 1$), the equations for the autocorrelation and power spectrum are:

$$\begin{aligned} R_{BPSK-R}(\tau) &= \begin{cases} 1-|\tau|/T_C & |\tau| \le T_c \\ 0, & \text{elsewhere} \end{cases} \\ S_{BPSK-R}(f) &= T_c \operatorname{sinc}^2(\pi f T_c) \end{aligned} \tag{8.89}$$

When using a noncoherent early-late power DLL discriminator for BPSK-R(n) modulated codes, the thermal noise code tracking error can be found by substituting (8.89) into (8.87). The result can be approximated by [77]:

$$\sigma_{tDLL} \cong \begin{cases} \sqrt{\dfrac{B_n}{2C/N_0} D\left[1+\dfrac{2}{TC/N_0(2-D)}\right]}, & D \ge \dfrac{\pi R_c}{B_{fe}} \\ \sqrt{\dfrac{B_n}{2C/N_0}\left(\dfrac{1}{B_{fe}T_c}+\dfrac{B_{fe}T_c}{\pi-1}\left(D-\dfrac{1}{B_{fe}T_c}\right)^2\right)\left[1+\dfrac{2}{TC/N_0(2-D)}\right]}, & \dfrac{R_C}{B_{fe}} < D < \dfrac{\pi R_c}{B_{fe}} \\ \sqrt{\dfrac{B_n}{2C/N_0}\left(\dfrac{1}{B_{fe}T_c}\right)\left[1+\dfrac{1}{TC/N_0}\right]}, & D \le \dfrac{R_c}{B_{fe}} \end{cases} \tag{8.90}$$

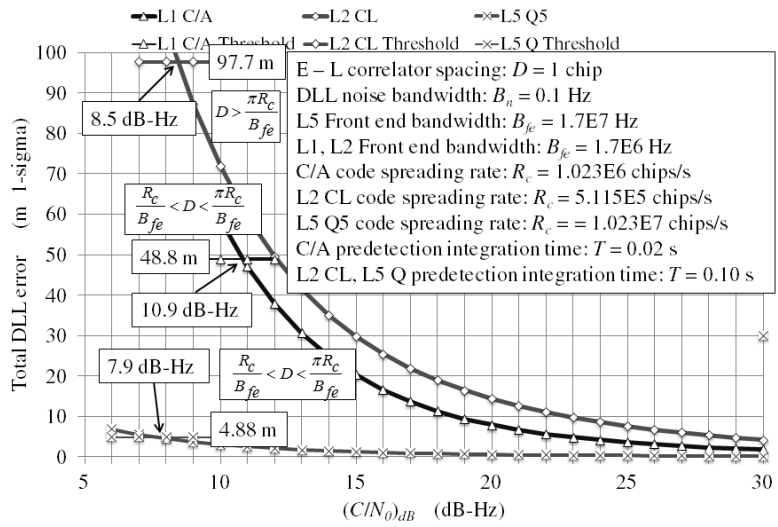
The part of (8.87) and (8.90) in brackets involving the predetection integration time, T, is called the squaring loss. Hence, increasing T reduces the squaring loss in noncoherent DLLs. When using a coherent DLL discriminator, the bracketed term on the right is equal to unity (no squaring loss) [77]. As seen in (8.90), the DLL error is directly proportional to the square root of the loop filter noise bandwidth (lower B_n results in a lower error that results in a lower C/N_0 threshold). Also, increasing the predetection integration time, T, results in a lower C/N_0 threshold, but with less effect than reducing B_n. Reducing the correlator spacing, D, also reduces the DLL error at the expense of increased code tracking sensitivity to dynamics. Narrowing D should be accompanied by increasing the front-end bandwidth B_{fe} to

avoid flattening of the DLL correlation peak in the region where the narrow correlators are being operated. In fact, (8.90) shows that there is no benefit to reducing D to less than the reciprocal of the front-end bandwidth (times the spreading code rate).

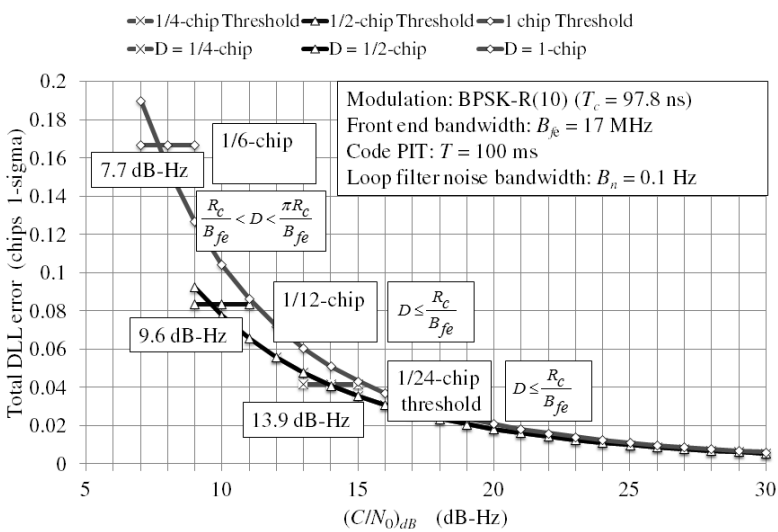
Note that the DLL error is independent of the DLL loop filter order in (8.90). To convert the DLL error from chips into meters, multiply (8.90) by $cT_c = c/R_c$. As examples, multiply (8.90) by c/1.023E6 = 293.05 m/chip for C/A code and the composite spreading code rate for L2C, or by c/1.023E7 = 29.305 m/chip for L5.

Figure 8.74(a) uses (8.90) to compare the GPS BPSK C/A code, L2 CL (pilot) and L5 Q5 (pilot) code accuracies and thresholds in units of meters. Although narrower correlators with corresponding wider front-end bandwidths could be used, a conventional one-chip correlator E-L spacing ($D = 1$) is used in all three cases. In the case of the L1 and L2 signals, the same front-end bandwidth, B_{fe} = 1.7E6 Hz, is assumed and a wideband front-end with B_{fe} = 1.7E7 Hz is assumed for the L5 signal. It is noteworthy that there is less than 0.1 dB of signal energy lost by using these bandwidths instead of the norm of $2R_c$ because there is no energy at the nulls, so if there are intended benefits to be derived by using a wider front-end bandwidth than $2R_c$, then it should be significantly wider (as is the case for the L2 front end). Since the pilot channels of L2 CL and L5 are used, then they both significantly benefit by the extended predetection integration time, T = 100 ms, that is used (and could be as long as the measurement time interval because of the carrier-aided-code feature that keeps the code tracking loop stable). The C/A code is limited to the 20-ms period of the legacy GPS navigation message data modulation. All of the assumptions are noted in Figure 8.74(a). Based on these assumptions, the C/A code and L5 Q5 plots used the middle equation of (8.90), while the L2 CL plot used the top equation. Observe the significant accuracy and threshold performance advantage of L5 Q5. Because of this superior accuracy performance, L5 will become a primary GPS signal for precision civil users. Obviously, the tracking threshold robustness of the GNSS receiver is limited by the carrier-tracking loop, so unless there is some form of external velocity aiding, the receiver will not achieve the tracking thresholds shown in this figure. (Refer to Section 9.2.3 for receiver tracking techniques that improve threshold.)

Figure 8.74(b) uses (8.90) to compare the DLL performance of the L5 Q5 signal for three correlator values of D = 1, ½ and ¼ chip) using a minimum receiver front-end bandwidth, B_{fe} = 1.7E7 Hz. As noted in this figure, reducing D to ½-chip and ¼-chip spacing for this bandwidth requires the use of the lower equation of (8.90). As a consequence, the code-tracking loop becomes insensitive to changes in the D value in this region. As a result, not only is there no accuracy payoff for reducing D from ½ to ¼ for this front-end bandwidth, but there is also a loss of code tracking threshold. Code tracking threshold loss is unimportant for an unaided GNSS receiver where the carrier-tracking threshold determines the receiver channel tracking performance, but there is no point in reducing D below ½-chip for this front-end bandwidth. This situation happened because both cases called for the lower equation of (8.90) for both the ½-chip and ¼-chip correlator spacing as observed in the figure. However, if B_{fe} were increased to 3.3E7 Hz then the top equation of (8.90) would be used for D = 1, the middle equation for D = ½ and the bottom equation for D = ¼, each choice producing a progressive payoff in accuracy with only slight loss of tracking threshold. So even though there is only a fraction



(a)

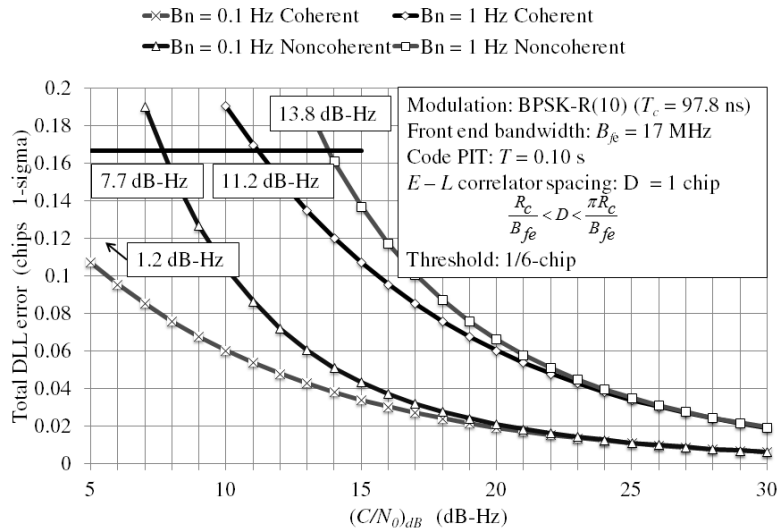


(b)

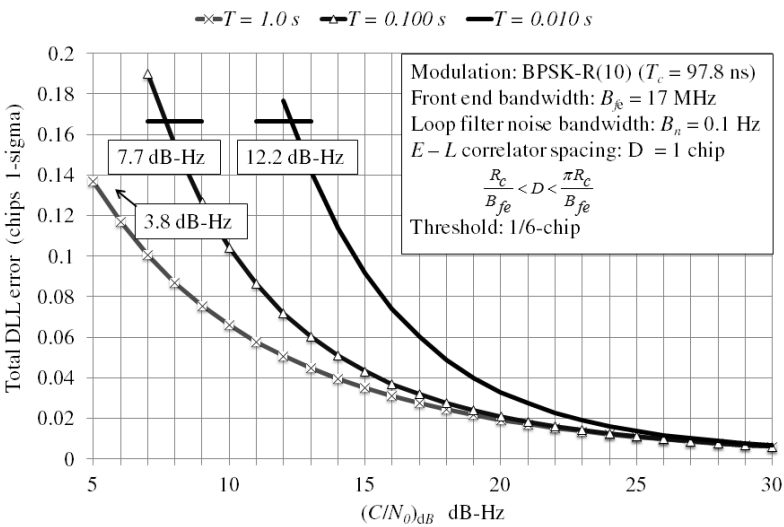
Figure 8.74 Delay lock loop error versus $(C/N_0)_{dB}$ for: (a) comparison of DLL accuracies and thresholds between L1 C/A, L2 CL and L5 Q codes, (b) comparison of L5 Q DLL error for different correlator spacing, (c) effect of noise bandwidth on L5 Q DLL tracking threshold, and (d) effect of predetection integration time on L5 Q DLL error.

of a decibel loss in main lobe power when B_{fe} = 1.7E7 Hz, there is a power increase from sidelobe power when B_{fe} = 3.3E7. These choices have to be traded between accuracy benefits versus increased vulnerability to interference.

Figure 8.74(c) uses (8.90) to demonstrate improved DLL accuracy and tracking threshold of the L5 Q signal by reducing code loop noise bandwidth. Since dynamic stress has been removed from the code-tracking loop using carrier-aided code (presented in Section 8.4.2.2), narrowing the code loop noise bandwidth is limited only by the quality of the velocity aiding. This figure also compares accuracy



(c)



(d)

Figure 8.74 (continued)

improvements of DLL coherent tracking (using the coherent code discriminator shown in Table 8.22). The coherent DLL mode can be activated when the carrier-tracking loop is in PLL, typically the normal carrier steady state tracking mode. Since the $(C/N_0)_{dB}$ meter (presented in Section 8.13.1) and phase lock detector (presented in Section 8.13.2.1) can be very sensitive and reliable state transition indicators, coherent DLL tracking can be the steady state code tracking mode until it becomes either prudent (from a low C/N_0 meter reading) or mandatory (from the pessimistic phase lock detector indicating loss of phase lock) to immediately transition into DLL noncoherent code tracking mode. Note that the design parameters

assumed for Figure 8.74(c) call for the use of the middle equation in (8.90) for all L5 Q case examples.

Figure 8.74(d) uses (8.90) to illustrate improved DLL accuracy and tracking threshold of the L5 Q signal by increasing the predetection integration time, T. Note that the lowest code-tracking threshold is achieved for $T = 1$ second. The code loop predetection integration time can be much longer than the carrier loop because of the carrier-aided-code feature and is unrestricted by data transitions for the L5 Q pilot channel. This is achieved with legacy GPS signals with only data channels by a technique called data wipe-off. This technique uses the GPS receiver's knowledge of the navigation message data bit stream (after 30 seconds of error-free data demodulation) to remove the 180° data transitions. This data wipe-off technique allows longer than 20-ms predetection integration times and, if properly implemented, achieves the 6 dB of additional $(C/N_0)_{dB}$ threshold improvement that a pilot channel provides. This is a short-term desperation DLL weak signal hold-on strategy for an externally aided legacy GPS receiver when the carrier is aided open loop. Data wipe-off also improves the PLL tracking threshold when the carrier loop is closed-loop aided, but not to the extent that the code loop tracking threshold is improved. Changes in any part of the SV navigation message data stream by a GPS control segment upload or autonomously by the SV will cause errors in data wipe-off, which, in turn, will cause deterioration in the tracking threshold. Alternatively, noncoherent integration is used to improve code-tracking threshold under these circumstances, but this does not achieve as much improvement as coherent integration. There is no longer a need for these legacy signal techniques with the introduction of pilot channels. Clearly, the use of a pilot channel is far more reliable than data wipe-off.

To check on the worst-case code loop stability for Figure 8.74(d), note that $B_n T = 0.1$ when $T = 1$ second and a code loop noise bandwidth, $B_n = 0.1$ Hz. From Table 8.24 for a first-order DLL loop and for computation delay T, the 30° phase margin requires $B_n T \leq 0.134$. Thus, the code DLL does meet this apparent stability requirement, but in fact the code tracking loop computation delay T is actually the same as the carrier loop T owing to the carrier-aided-code technique shown in Figure 8.18. There is even more code DLL phase margin. Obviously, the value for T can be longer for the code loop than for the carrier loop when carrier-aided-code is in operation, but keep in mind that both tracking loops are updated at the carrier loop rate.

The DLL tracking loop dynamic stress error is determined by:

$$R_e = \frac{d^n R/dt^n}{\omega_0^n} \text{ (chips)} \tag{8.91}$$

where d^nR/dt^n is expressed in chips/s^n and n is the same as the code loop order.

As an example of how the dynamic stress error is computed from (8.91), assume that the code loop is an unaided third-order C/A code DLL with $B_n = 2$ Hz and $D = 1$ chip. If the maximum line-of-sight jerk stress is 10 g/s, then this is equivalent to $d^3R/dt^3 = 98$ m/s^3/293.05 m/chip = 0.3344 chips/s^3. The third-order loop natural frequency, $\omega_0 = B_n/0.7845$ is obtained from Table 8.23. Substituting these numbers into (8.91) results in a maximum dynamic stress error of $R_e = 0.02$ chip, a 3-sigma

effect. Since the 3-sigma threshold is 1/2-chip, this would indicate that the DLL noise bandwidth is more than adequate for C/A code. If the receiver was P(Y) code, then $R_e = 0.2$ chip, which is still adequate. Note that carrier aided code techniques removes virtually all the dynamic stress from the code tracking loop. Therefore, so long as the carrier loop remains stable the code loop experiences negligible dynamic stress. Therefore, this effect is not included in the code loop tracking threshold analysis and would only be used if there is an open carrier loop with the carrier estimate being performed by an external source of velocity aiding. Any dynamic stress due to error in that aiding source would be analyzed using (8.91).

8.9.10 BOC Code Tracking Loop Measurement Errors

The modernized GPS M code was the first GNSS signal to use BOC modulation. It uses a sine-phased BOC(10,5) modulation technique to split the carrier spectrum. The power spectral density for a sine-phased BOC modulation is [78]:

$$S_{BOC_s}(f) = \begin{cases} T_c \text{sinc}^2(\pi f T_c) \tan^2\left(\dfrac{\pi f}{2f_s}\right) \text{ for } k \text{ even} \\ T_c \dfrac{\cos^2(\pi f T_c)}{(\pi f T_c)^2} \tan^2\left(\dfrac{\pi f}{2f_s}\right) \text{ for } k \text{ odd} \end{cases} \tag{8.92}$$

By substituting (8.92) into (8.87), an approximation for M code DLL error in the presence of thermal noise is [79]:

$$\sigma_{tM} = \begin{cases} \dfrac{1}{T_C}\sqrt{\dfrac{B_n}{(2\pi)^2 \dfrac{C}{N_0}\left(0.66B_{fe} - 7.7E6\right)^2}\left[1 + \dfrac{1}{\left(7.3E8B_{fe} - 0.96\right)T\dfrac{C}{N_0}}\right]}, 16 \text{ MHz} \le B_{fe} \le 24.5 \text{ MHz} \\ \dfrac{1}{T_C}\sqrt{\dfrac{B_n}{(2\pi)^2 \dfrac{C}{N_0}\left(0.007B_{fe} + 8.4E6\right)^2}\left[1 + \dfrac{1}{0.837T\dfrac{C}{N_0}}\right]}, 24.5 \text{ MHz} < B_{fe} \le 30 \text{ MHz} \end{cases} \tag{8.93}$$

where $\dfrac{1}{T_c} = R_c = 5.115\text{E}6$ in units of chips/s and B_{fe} inside the equations is in units of Hz.

To obtain the 1-sigma error in meters, multiply (8.93) by c/5.115E6 = 58.6105 m/chip. Note that the correlator spacing, D (in M-chips), does not appear in (8.93), but this approximation is restricted to E-L spacing of ¼-chip or less of an M code chip as defined by the $R_c = 5.115$ Mcps M code spreading rate. The rule-of-thumb tracking threshold for M code DLL tracking threshold is identical to (8.86).

M code has a pilot channel provision called time division data multiplexing (TDDM) that permits extended predetection integration times. TDDM is implemented so that every other code bit is dataless, thereby losing 3 dB of power in

both the pilot and data channels in return for a net 3 dB improvement in carrier tracking threshold using the pilot channel. Reduce $(C/N_0)_{dB}$ by 3 dB so that $C/N_0 = 10^{\{[(C/N_0)_{dB}-3]/10\}}$ in both places in (8.93) to account for this effective loss when M code is in TDDM mode. Note that in the TDDM mode, the carrier-aided-code loop can coherently integrate its pilot input up to the pseudorange measurement period and that value of T is used in (8.93) to estimate the code measurement (pseudorange) error. The effective T of the code-tracking loop remains the same as the carrier loop update rate because both are sent back to their respective NCOs at the same rate (by virtue of the carrier-aided code technique). The carrier loop T must be maintained at the value that supports the dynamic stress requirement.

Figure 8.75 compares the accuracies and thresholds of M code to P(Y) code assuming both use the same DLL noise bandwidth, $B_n = 0.1$ Hz, and both share the same wideband front end, $B_{fe} = 3.0\text{E}7$ Hz, so the lower equation of (8.96) is used for M code with D = typical M code value of 1/8 chip, but the DLL error is converted to meters. Two examples of M code are presented, the first operating in TDDM mode with $T = 0.3$ second (so the 3-dB loss in $(C/N_0)_{dB}$ is taken into account) and the second operating in non-TDDM mode with T limited to 0.01 second by the fastest M code symbol rate of 100 Hz. The P(Y) code example uses D = typical value of 1 chip and T limited to 0.02 second by the 50-Hz data rate of the legacy GPS signals. The P(Y) code assumptions dictate the use of the middle equation of (8.90), but the DLL error is converted into meters. Note that M code accuracy is superior to P(Y) code in both cases, but its rule-of-thumb tracking threshold is a lower value than the P(Y) because D is smaller for M code. As a result, P(Y) code threshold is slightly better than the non-TDDM M code example, but the TDDM M code example outperforms the P(Y) code threshold. This TDDM M code threshold can be further improved by extending the coherent integration time.

Unlike the modernized civilian GPS signals, the M code does not have two different replica codes, one for the pilot channel and another for the data channel. Instead, the pilot and data channels are time multiplexed onto the carrier signal,

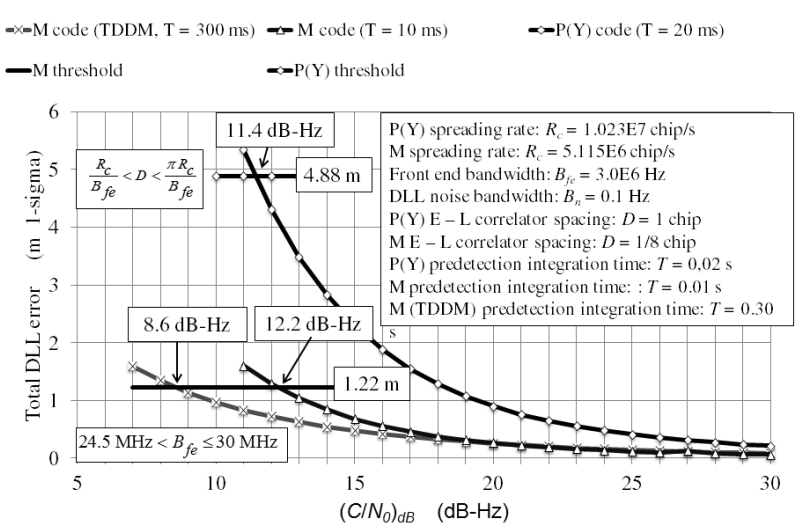


Figure 8.75 DLL accuracies and thresholds comparisons among M code (TDDM), M code, and P(Y) code.

each synchronized by the even or odd bits of the replica code generator. So the TDDM feature in M code introduces three levels (+1, 0, –1) of signal states instead of the usual two (+1, –1) signal states. During TDDM operation in the receiver, the data demodulation interval is selected based on a DATA = "true" signal from the replica code generator, the data interval is selected by that state and the value present will be either +1 or –1 corresponding to a 1 or 0, respectively, in that interval. When DATA = "false," then PILOT = "true" and the pilot interval is selected by that state, where the value will always be 0 in that interval. So after carrier and code wipe-off have been performed, the pilot and data channels have to be time demultiplexed on a tri-level basis and directed to their respective tracking and demodulation processes.

Reference [39] provides additional insight plus numerous graphs depicting the DLL code tracking error and threshold performance of a large number of BPSK and TMBOC signals using both noncoherent and coherent early minus late processing (abbreviated as NELP and CELP, respectively) tracking schemes. Keep in mind that the carrier- tracking loop must be in phase lock at all times that the code-tracking loop is operated coherently.

Reference [80] describes a GNSS receiver baseband architecture appropriately called a double estimator that converts any received BOC signal into a BPSK signal using three tracking loops in a manner that substantially retains the BOC code measurement accuracy. The design is different from the technique described in [59] because it not only removes the subcarrier frequency, but it also tracks it using what is called a subcarrier lock loop (SLL). The SLL square wave wipe-off is preceded by conventional carrier wipe-off and followed by conventional BPSK DLL code wipe-off. The SLL operates as a precise code tracking loop that tracks the M square wave component of the incoming BOC(M,N) signal while the DLL tracks the N component PRN code (with the subcarrier removed) using conventional BPSK E, P, and L correlators. The precise SLL tracking loop has a significant ambiguity to overcome in order to provide an unambiguous pseudorange measurement, but this is accomplished with aiding from the coarse DLL tracking loop and cooperatively using multiple channels along with the LAMBDA method [59] that was developed for optimally resolving the ambiguity in real-time kinematic applications. The scheme can experience difficulties with locking onto sidelobes under certain circumstances, but detection and correction can also be achieved based on multiple channel cooperation.

Fortunately, the internationally harmonized signal BOC(1,1) correlation envelope manifests very small sidelobes, so false code-lock tracking is unlikely when using conventional DLL E-L tracking loop discriminators.

8.10 Formation of Pseudorange, Delta Pseudorange, and Integrated Doppler

Contrary to popular belief, the natural measurements of a GNSS receiver are not pseudorange or delta pseudorange [30]. This section describes the natural measurements of a GNSS receiver and describes how they may be converted into pseudorange, delta pseudorange, and integrated carrier Doppler phase measurements. The natural measurements are replica code phase and replica carrier Doppler phase (if

the GNSS receiver is in phase lock with the satellite carrier signal) or replica carrier Doppler frequency (if the receiver is in frequency lock with the satellite carrier signal). The replica code phase can be converted into satellite transmit time that is used to compute the pseudorange measurement. The replica carrier Doppler phase or frequency is used to compute delta pseudorange measurements. The replica carrier Doppler phase measurements can also be used to compute integrated carrier Doppler phase measurements that are required for ultraprecise (differential) static and kinematic interferometry applications.

The most important concept presented in this section is the measurement relationship between the replica code phase state in the GNSS receiver and the satellite transmit time. Any error between the receiver replica code correlating with the incoming code is an error in the time transfer because the received signals themselves are buried in noise and cannot be read directly. The only measurements that can be read are the receiver's replica code DLL phase state and the one-cycle or half-cycle ambiguous standing wave phase of the carrier PLL phase state and these are measurements of each SV transmit time. Uncompensated ionospheric and tropospheric delays, relativistic effects, and multipath error also contribute to the received time error. The first principle of satellite navigation is that all GNSS SVs are transmitting PRN codes that are clocked by an atomic time standard and the PRN codes and carrier frequencies are kept synchronous with precise time. This first principle results in the transmit times of all SVs being maintained with respect to each SATNAV system's timescale that is ultimately steered to UTC (see Section 2.7.2). Since GNSS receivers provide the most accurate means of worldwide time transfer, it is essential that every GNSS receiver design ultimately achieves and maintains a monotonically increasing time (often time of week, with 1-week ambiguity) that is ultimately converted to UTC including the date for its users. The SV navigation message data provide the means to resolve the ambiguity in the PRN code length into the 1-week ambiguity and to convert that into UTC, but it is the receiver design responsibility to assign and maintain a monotonically increasing time that is synchronous with its replica code. The transmit times from four or more SVs is converted into pseudoranges using an estimated receive time (that typically is in common with all measurements) from which three-dimensional position and a time bias are determined by the navigation measurement incorporation process. Based on the known SV orbit geometry, that estimated receive time should never be less accurate than about 20 ms beginning with the acquisition of the first SV and the knowledge of its transmit time. When the estimated receive time is corrected with the time bias, true time of week is obtained. Time-keeping is maintained on the SV (under synchronization scrutiny by its respective control segment), so that its start boundaries are synchronized to the time of week. Each SV also keeps track of the international standard of time that includes leap seconds and corresponds to the time of week. This information is provided in the navigation message so that the GNSS receiver can obtain the unambiguous UTC from this message and synchronize it to its time of week.

The receiver time measurement begins with its ambiguous SV transmit time that is modulo one PRN code period. For example, the unencrypted GPS P code period is exactly one week, so its corresponding GPS time relationship is ambiguous over one week. Most GNSS PRN codes have much shorter periods. For example, the GPS C/A code period is only 1 ms, but there is a handover word in every SV data

message subframe that can be used to increase the transmit time ambiguity from 1 ms to 1 week. The original intent of this data was to provide the transmit time information necessary to handover from C/A code to P(Y) code. There is a similar process involved for every GNSS SV ambiguous PRN code period to increase the ambiguity to a larger period (e.g., 1 week). Every GNSS receiver is vulnerable to false ambiguity resolution and, under weak signal acquisition conditions, an ambiguity error will occur. When the error does occur, it causes serious range measurement errors, which, in turn, result in severe navigation position errors. The use of overlay codes in modern GNSS signals provides a significant improvement to this problem because they eliminate the need for the less reliable and time-consuming bit synchronization and frame synchronization required for signals that do not have this feature, such as C/A code.

8.10.1 Pseudorange

The definition of pseudorange to SV_i, where i is the satellite identification index, is as follows:

$$\rho_i(n) = c\left[T_R(n) - T_{Ti}(n)\right] \text{(m)} \tag{8.94}$$

where

c = speed of light = 299,792,458 (m/s);
$T_R(n)$ = receive time corresponding to epoch n of the GNSS receiver's clock (s);
$T_{Ti}(n)$ = transmit time based on the SV_i clock (s) observed at the receive time (s);

$T_{Ti}(n)$ is the natural measurement of the SV_i observable because it represents the replica code state (in chips converted to seconds) at receiver epoch n as interpreted by the receiver's linear time representation of its replica code state at this epoch. This state includes the integer replica code state time and the precise fraction of the replica code chip as measured by a software code accumulator that is updated every predetection integration time, T, when the code NCO is updated by the code tracking loop. (The design of the code accumulator is described later.) The code accumulator knows the current propagation rate of the NCO, so it can precisely predict the fractional state of the NCO at any time in between T.

To visualize this pseudorange measurement process, Figure 8.76 depicts GNSS satellite SV_i transmitting its PRN spreading code epochs, PRN_i, starting at the GNSS end of week. The receiver navigation process requests the measurement from all channels at the same receiver epoch n that is always scheduled to occur at some receiver set time, called fundamental time frame (FTF) in the figure. The FTF time is a monotonically increasing counter that has an accompanying receive time associated with it for pseudorange computation purposes. Visualize that the FTF is asynchronous with the multiplicity of incoming SV signals being tracked by multiple channels since they are all tracking different transmit times at the same receive time epoch n. For this reason, Figure 8.76 purposely skews the measurement time of the SV_i observable shown as Transmit time $(n) = T_{Ti}(n)$ in the figure to emphasize that there is usually time skew with respect to any code chip transition boundary.

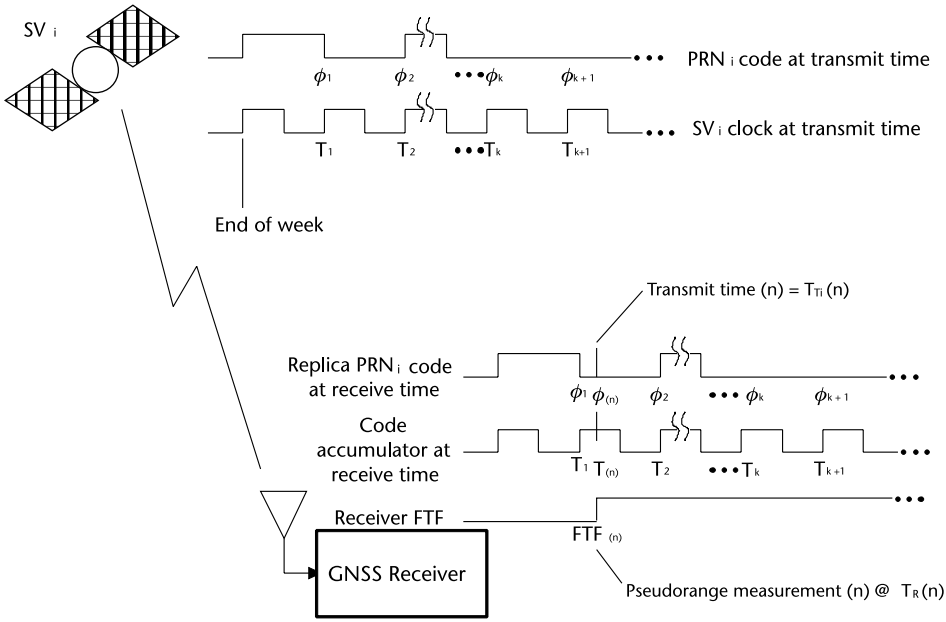


Figure 8.76 Relationship of satellite transmit time to pseudorange measurements.

$T(n)$ is the receiver epoch time in Figure 8.76 corresponding to $FTF(n)$, the scheduled measurement time request. Note in Figure 8.76 that corresponding to each chip of the PRN_i code is a linear SV_i clock time. Every epoch in the PRN code that is transmitted by SV_i is precisely aligned to the time of week as maintained inside SV_i's time-keeping hardware but obviously not transmitted. When this transmitted code reaches the GNSS receiver whose replica PRN code is successfully correlating with it, the phase offset of the replica code with respect to the beginning of the GNSS week represents the transmit time of SV_i. The pseudorange derived from this measurement corresponds to a particular receive time epoch (epoch n) in the receiver. Below $FTF(n)$ is the notation Pseudorange measurement (n) @ T_R (n) that corresponds to the computation shown in (8.94).

Unfortunately, the common misunderstanding by the navigation process experts that pseudorange is the natural receiver measurement results in their specification to the baseband receiver experts that they send pseudorange measurements to the navigation process. Instead, it is prudent to pass the observable T_{Ri} (n) (along with a receive time tag) to the navigation process because the first thing the navigation process needs for position measurement incorporation is the SV_i transmit time corrected to true GNSS time. The SV_i transmit time is lost if this artificial computation is performed by the receiver baseband control process. This method forces the navigation process into a wasteful iterative process of computing the SV_i transmit time.

Highly sophisticated receivers implement vector tracking of the SVs instead of scalar tracking described herein. This eliminates the pseudorange observable problem because ideally either the raw I and Q measurements or the discriminator outputs are sent to the navigation process as measurements for Kalman filtering by the navigation process. Thus, the navigation process dynamically changes the noise bandwidth of the tracking loops in an optimal manner plus it provides cross-aiding

between channels. However, the increased navigation processing load usually requires some compromises in the practical implementation of this scheme.

Typically, the receiver will take a set of measurements at the same receive time epoch. This process is why the receive time is not identified with any particular SV PRN number in (8.94). When the receiver navigation process schedules a set of measurements, it does this based on its own internal clock (set time) that contains a bias error with respect to true GNSS time. Eventually the navigation process learns this bias error as a by-product of the navigation solution. The SV transmit time also contains a bias error with respect to true system time, although its control segment ensures that this is maintained at a very small offset value. A correction to this offset is transmitted to the receiver by SV_i as clock correction parameters via the navigation message. However, none of these corrections are included in the pseudorange measurement of (8.94). These corrections and others are determined and applied by the navigation process.

8.10.1.1 Pseudorange Measurement

From (8.94), it can be concluded that if the receiver baseband control process can extract the SV transmit time from the code-tracking loop, then it can compute a pseudorange measurement. The precise transmit time measurement for SV_i is equivalent to its code phase offset with respect to the beginning of the system time week. There is a one-to-one relationship between the SV_i replica code phase and the monotonically increasing time of week. Thus, for every fractional and integer chip advancement in the code phase of the replica code generator since the initial (reset) state at the beginning of the week, there is a corresponding fractional and integer chip advancement in the time of week. In the following discussion, the fractional and integer chip code phase is called the code state and the receiver baseband control process time keeper that contains the GNSS time corresponding to this code state is called the code accumulator.

The replica code state corresponds to the receiver's best estimate of the SV transmit time. The receiver baseband control process knows the code state because it sets the initial states during the search process and keeps track of the changes in the code state thereafter. The receiver baseband code tracking loop process keeps track of the GNSS transmit time corresponding to the phase state of the code NCO and the replica PRN code generator state after each code NCO update. It does this by discrete integration of every code phase increment over the interval of predetection integration time, T, since the last NCO update and adds this number to the code accumulator. The combination of the replica code generator state (integer code state) and the code NCO state (fraction code state) is the precise replica code state. Since the code phase states of the replica code generator are pseudo random, it would be impractical to read the code phase state of the PRN code generator and then attempt to convert this nonlinear code state into a linear GNSS time state, say, by a table look-up. There are usually too many possible code states, especially for signals with encrypted codes.

A very practical way to maintain the GNSS time in a GNSS receiver is to use a separate code accumulator in the GNSS receiver baseband control process and to synchronize the replica PRN code generator to the phase state of this accumulator. Figure 8.77 illustrates a high level block diagram relationship between the replica

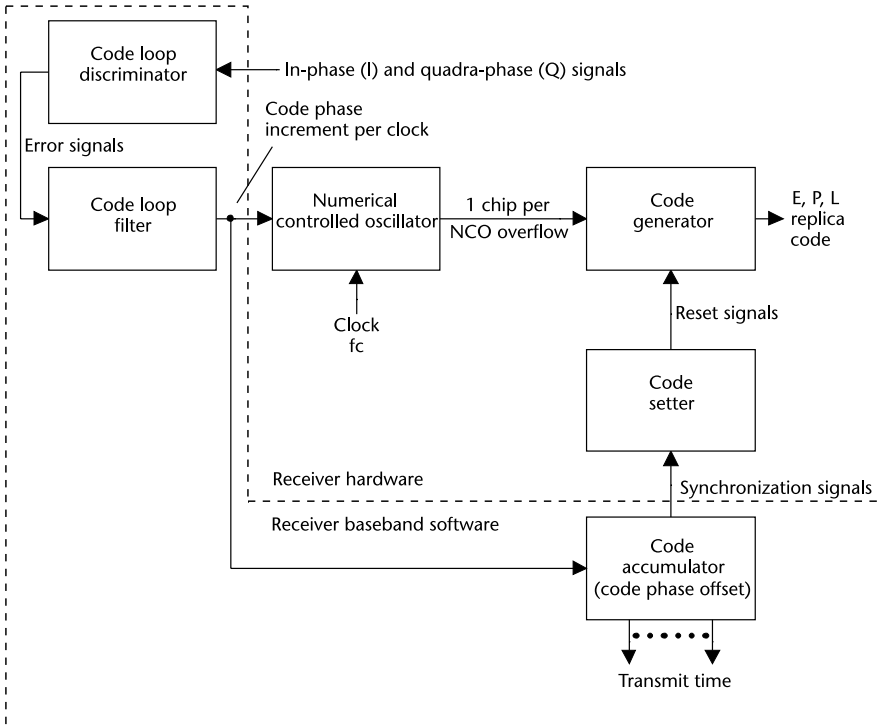


Figure 8.77 Relationship between the replica code generator and code accumulator.

code generator (as shown in Figures 8.13 or 8.14) and the code accumulator. It assumes that the fast functions of Figures 8.13 and 8.14 (specifically the replica code generator and code NCO) are implemented in hardware. The code setter (not shown in Figures 8.13 and 8.14) is also assumed implemented in hardware even though it is a slow function in terms of its period, *T*, but must be synchronous (fast) during the code-setting process. However, the same concept can be implemented in a software-defined receiver assuming it can support the code NCO design (normally not the case with current SDRs). The remaining functions are slow functions implemented in software. These are separated from the fast functions by a dotted line.

A typical GNSS navigation measurement incorporation rate is once per second, but for some applications much faster and it is possible for the receiver to schedule these measurements exactly on GNSS time rather than on set time (especially for more precise differential measurements between remotely located GNSS receivers). A typical GNSS receiver fundamental time frame (FTF) for scheduling measurements and synchronizing routines associated with data bit or symbol transitions is 10 ms. The receiver baseband control process schedule for updating the code and carrier NCOs is based on the current predetection integration time, *T*, usually some integer relationship to the FTF, such as 20, 10, 5, 2, or 1 ms. Assuming that the FTF is 10 ms, the receiver measurement process maintains a set time counter that will be called the FTF counter, typically a 32-bit counter that counts in 10-ms increments derived from the receiver's reference oscillator. The FTF counter is set to zero at power up, counts up, rolls over, counts up, and so forth. The FTF counter provides set time to every receiver process. Assuming that the navigation measurement incorporation rate is 1 Hz, the navigation process will schedule measurements to

be extracted from the code and carrier-tracking loops every hundred FTFs. This arrangement assumes the desired measurements are based on the receiver's time epochs. If based on GNSS time epochs, the navigation process requests the measurements with a bias value attached to the FTF so that the sum of the two corresponds to the exact 1-second epoch of the appropriate SATNAV system time. Those measurements can be output on set time with the offset bias time included with the measurement or there can be a steered clock that generates the output epoch and measurement to closely align in real time with the 1-second epoch. When the receiver baseband control process extracts the measurements from the code and carrier tracking loops, it time tags the measurements with the FTF count. The navigation process assigns and maintains a GNSS receive time corresponding to the FTF count. The receive time initialization can be the first SV's transmit time plus a nominal propagation time of, say, 76 ms (for MEO), if the navigation process does not know the GNSS time accurately. This nominal value will set the initial receive time accuracy to within 20 ms.

When a pseudorange measurement is scheduled on FTF(n), the receiver baseband code tracking loop process extracts the SV_i transmit time from its code accumulator and propagates this time forward to FTF(n). The result is the SV_i transmit time with a measurement resolution of 2^{-N} of a code chip, where N is the number of bits in the code NCO. If the code NCO uses a 32-bit register, this measurement resolution is less than a quarter of a nanochip that makes the code measurement quantization noise negligible. As stated earlier, the navigation process computes the pseudorange from the SV_i transmit time measurement using (8.94) and time tag $FTF(n)$, but only after the navigation process applies the clock correction (including relativity correction), then uses the corrected SV_i transmit time to compute the SV_i position. The corrected pseudorange is then incorporated into the navigation filter. (Satellite clock and relativistic corrections are discussed in Chapter 10.)

8.10.1.2 Measurement Time Skew

Figure 8.20 illustrates the time skew, T_s that exists between the SV data or symbol transition boundaries and the receiver FTF epochs. The GNSS control segments ensure that every SV transmits every epoch closely aligned to true GNSS time (e.g., the GPS SV clocks are aligned to within 1 ms of true GPS time). Therefore, all of the SV data transition boundaries are approximately aligned to true GNSS time at transmit time. However, at the GNSS receiver the SV data transition boundaries are, in general, skewed with respect to each other and with respect to the receiver's FTF boundary. This is because the SVs are at different ranges with respect to the user GNSS receiver antenna phase center. The user GNSS receiver must adjust the phases of its integrate and dump boundaries in order to avoid integrating across the SV data bit transition boundaries. The time skew, T_s (labeled Offset to SV_i data/symbol transitions in Figure 8.20), is different for each SV being tracked and it also changes with time because the ranges to the SVs change with time. Therefore, the epochs corresponding to the end of each replica code generator period are skewed with respect to each other and by T_s with respect to the FTF. As a result, the integrate and dump times and the updates to the code and carrier NCOs are performed on a changing skewed time phase with respect to the FTF time phase, but the receiver baseband control process learns and controls this time skew in discrete

phase increments. The code accumulator is normally updated on the skewed time schedule that matches the code NCO update schedule. Therefore, if all of the GNSS receiver measurements of a multiple channel GNSS receiver are to be made on the same FTF, the contents of the code accumulator, when extracted for purposes of obtaining a measurement, must be propagated forward by the amount of the time skew between the code NCO update events and the FTF.

8.10.1.3 Maintaining the Code Accumulator

The following code accumulator was originally designed for setting the initial code generator and NCO phase states (and maintaining them) of the GPS C/A and P(Y) replica code generators [29]. The same design is readily adaptable to any GNSS PRN code because the original design was based on synchronizing code measurements to the monotonically increasing GPS time of week and for extracting precise SV transmit time measurements at any navigation process measurement incorporation rate scheduled on any desired set time. Three counters, Z, X1, and P, are used to maintain the code accumulator. The Z counter (19 bits minimum but typically a 32-bit register in software) accumulates in GPS time increments of 1.5 seconds and then is reset one count short of the maximum Z-count of 1 week = 403,200. Hence, the maximum Z-count is 403,199. The X1 counter (24 bits minimum but typically a 32-bit register in software) accumulates the 1.5-second basic timing unit of GPS in time increments of the highest spreading code rate [i.e., $T_c = 1/(10 \times 1.023 \times 10^6) = 97.8$ ns] and then is reset one count short of the maximum X1-count of 1.5 seconds = 15,345,000. Hence, the maximum X1-count is 15,344,999. The P-counter is the same size, 2^N, as the code NCO accumulator, typically $N = 32$ bits. The P counter input from the code tracking loop is adjusted so that it overflows in time increments (excluding Doppler effect) of 97.8 ns. One example of how this is accomplished for all PRN code spreading code rates is to use a constant code loop output bias of $R_c = 1/T_c$ of 10.23 Mcps and then divide the P counter output by 1, 2, 5, or 10 as appropriate for the actual PRN code being tracked. This division is typically performed as part of the replica code generator that is typically implemented in hardware (e.g., divide by 10 for $R_c = 1.023$ Mcps). Note that this division also perfectly compensates for the Doppler effect on the PRN spreading code being used. Using the constant 10.23-Mcps code bias scheme and assuming that the code NCO and code accumulator are updated every T seconds, and that there is exactly T seconds of time delay in the loop filter feedback process so that the code NCO is updated exactly on T second boundaries, the algorithm for maintaining the entire code accumulator is as follows (note that the equals sign in the algorithm means "is replaced with"):

$$
\begin{aligned}
&P_{temp} = P + f_S \Delta\phi_{co} T \\
&P = \text{fractional part of } P_{temp} \text{ (chips)} \\
&X_{temp} = \left(X1 + \text{whole part of } P_{temp}\right) \big/ 15{,}345{,}000 \\
&X1 = \text{remainder of } X_{temp} \text{ (chips)} \\
&Z = \text{remainder of } \left[\left(Z + \text{whole part of } X_{temp}\right) \big/ 403{,}200\right] (1.5 \text{ s})
\end{aligned}
\tag{8.95}
$$

where P_{temp} = temporary P register, f_s = code NCO clock frequency (ADC sample rate) (Hz), $\Delta\phi_{co}$ = code NCO phase increment per sample, and T = time between code NCO updates (predetection integration time) (s).

The above definition of $\Delta\phi_{co}$ contains two components, the code NCO bias and the code loop filter Doppler correction (both appropriately scaled to the R_c of the replica code generator). A GNSS receiver cannot incorporate measurements until the ambiguity in the replica code state is resolved. The receiver baseband control process requires additional information from the navigation data message to remove this ambiguity and when that information is received it is placed into the code accumulator, but this does not mean that the code accumulator cannot control the replica code generator during and after the open loop signal acquisition has found the signal and the code loop closure process is successful. The signal acquisition control process is also a slow function that uses the code setter and portions of the code accumulator to control the open loop search as well as the loop closure process, so the information transfer is transient free.

Note in Figures 8.13 and 8.14 that the code NCO synthesizes a code shift register clock rate that is $\hat{f}/\delta$ where $\hat{f}$ is the code generator spreading code chip rate and δ is the separation between the E, P, and L replica code phases in chips. The E – L code correlator spacing is D in chips, so $D = 2\delta$. For a typical E – L spacing of D = 1 chip, the shift register is running twice as fast as the replica code generator. This timing generates phase shifted E and L replica codes that are used for error detection in the code discriminator. The P counter tracks the fractional part of the code phase state that is contained in the code NCO state at any instant in time, but the P counter is only updated every T seconds because that is the code NCO update rate.

Also note in Figures 8.13 and 8.14 that the set time sync that advances or retards the dump phases so that they are approximately aligned with the incoming SV signal data or symbol transition boundaries is controlled by the code accumulator. The accumulator does this by keeping the dump phases approximately aligned with its X1 transitions. Each one of these advance or retard commands results in a small change in the predetection integration time, T, that must be accounted for in the code and carrier tracking loops for that particular T. Once the data transition boundary is aligned, these advance and retard commands are rare because the data boundary phase will only change about 20 ms from the MEO SV at zenith to its rising or setting tracking horizon.

8.10.1.4 Obtaining a Measurement from the Code Accumulator

To obtain a measurement, the code accumulator must be propagated to the nearest FTF(n). This results in the set of measurements $P_i(n)$, $X1(n)$, and $Z_i(n)$ for SV_i. When converted to time units of seconds, the result is $T_{Ti}(n)$, the transmit time of SV_i at the receiver time epoch n. This is done very much like algorithm (8.95) except the time T is replaced with the skew time, T_S, and the code accumulator is not updated. The algorithm for the transmit time measurement at FTF(n) is (note that the equals sign in the algorithm means "is replaced with"):

$$
\begin{aligned}
&P_{temp} = P + f_S \Delta\phi_{co} T_S \\
&P_i(n) = \text{fractional part of } P_{temp} \text{ (chips)} \\
&X_{temp} = \left(X1 + \text{whole part of } P_{temp}\right) \Big/ 15,345,000 \\
&X1_i(n) = \text{remainder of } X_{temp} \text{ (chips)} \\
&Z_i(n) = \text{remainder of } \left[\left(Z + \text{whole part of } X_{temp}\right) \Big/ 403,200\right] (1.5 \text{ s})
\end{aligned}
\tag{8.96}
$$

Algorithm (8.96) produces no error due to the measurement propagation process for the code accumulator measurements because the code NCO is running at a constant rate, $\Delta\phi_{co}$ per clock sample, during the propagation interval and the baseband process knows this rate. Assuming double precision floating point computations are used, the following equation precisely converts the code accumulator measurements into SV_i transmit time:

$$
T_{Ti}(n) = \left[P_i(n) + X1_i(n)\right] \Big/ \left(10.23 \times 10^6\right) + Z_i(n) \times 1.5 \text{ (s)} \tag{8.97}
$$

The equation for computing the pseudorange to SV_i using (8.97) is:

$$
\rho_i(n) = \left[T_R(n) - T_{Ti}(n)\right] c \quad \text{(m)} \tag{8.98}
$$

where $T_R(n)$ = receive time of week for all SV measurements (s).

The receive time of week is maintained by the navigation process at the same resolution as the transmit time and rolls over once a week and is updated on set time FTF(n) epochs. It can be an arbitrary time of week and only becomes precise GNSS time when the navigation process time bias estimation is added to it. A good practice for MEO SVs is for the navigation process to initialize $T_R(n)$ to within about 20-ms accuracy using the transmit time of the first SV acquired (that is fairly accurate even before the navigation process corrects it) and adding the range of that SV (converted into time units) to the SV transmit time by assuming the SV is at a 45° elevation angle. Note that it is the error in the receive time estimate that is primarily responsible for the "pseudo" part of pseudorange even though its maximum error is bounded to a reasonable value by this initialization procedure.

8.10.1.5 Synchronizing the Replica Code Generator to the Code Accumulator

Synchronizing the replica code generator to the code accumulator is the most complicated part of the replica code control and measurement process. This complication arises primarily because it is the responsibility of the code accumulator and code setter to translate the random sequence taking place in the replica code generator into the linear time sequence taking place in the code accumulator while also in total control of the replica code generator matching the incoming PRN code. Fortunately, there are predictable reset timing events in every replica code generator that permit them to be synchronized to the code accumulator. The first thought might be to design the replica code generator such that it provides the linear time sequence that is synchronized by the reset epochs and read by the receiver baseband control process. However, the slow function control part of the receiver is where the

ambiguity in the PRN code is obtained from the navigation message data and that handover complexity must reside there. Also, this is where the code measurement extraction process takes place.

The replica code generator is operating on the SV receive time schedule, so the code accumulator and its associated code tracking loop functions must be operating on a schedule that closely matches the data or symbol transition boundaries of the SV being tracked. A proven design uses a code setter in the fast function replica code generator (typically hardware), maintains the code accumulator as a slow function, and periodically synchronizes the replica code generator to the code accumulator. An example of a code setter design technique using a replica C/A code generator is illustrated in Figure 8.78.

Referring to Figure 8.78, the code setter resets the G1 and G2 registers in the replica C/A code generator at the same time. The code setter principle of operation is that the slow function code accumulator can predict the precise time offset from the current state of the code accumulator at the next predetection integration time boundary and places that offset into the code setter. That offset is synchronously latched into the code setter at that boundary by the fast function clock (i.e., by the ADC sample frequency). The code setter then begins to count that offset at the same frequency being fed to G1 and G2 of the replica C/A code generator. A carry is produced when that offset becomes zero causing the G1 and G2 registers to be synchronously set to their starting (reset) points. The phase of the replica C/A code generator is correctly matched to the code accumulator from that point onward (if there are no glitches).

There are other design features that should be addressed. Note that there is a difference from previous functional block diagrams in the input frequency from the code NCO (shown as $2\hat{f}$ in the figure) to the input of this functional block diagram, but that input does not go directly to the replica C/A code generator. Also, the input signal is actually a Doppler-compensated 20.46-MHz NCO output that was created as a 10.23-MHz NCO frequency by the code tracking loop in combination with the universal code accumulator design. The doubled frequency is obtained by tapping the NCO one stage back, in the same manner as for the code shift register when the code E, P, L spacing is ½-chip. Closer inspection reveals that

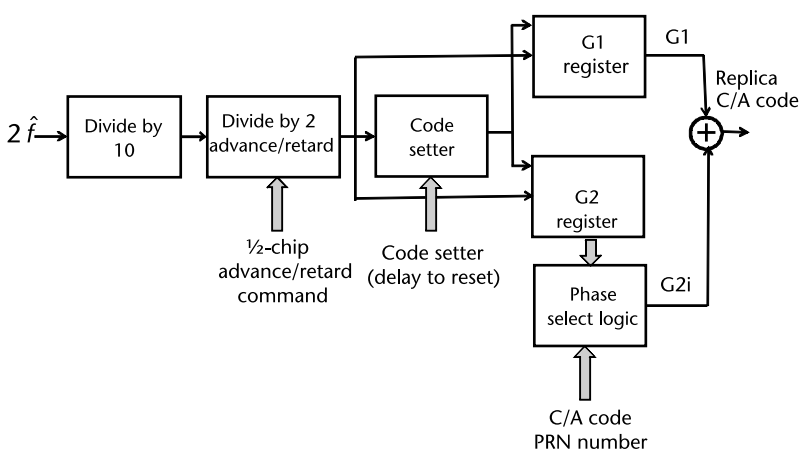


Figure 8.78 Code setter scheme for C/A code generator.

the replica C/A code generator normally receives the desired Doppler compensated 1.023-MHz frequency.

The first input stage of Figure 8.78 provides the divide-by-10 requirement for the universal 10.23-MHz code accumulator to code NCO output design convention. The second stage provides a steady state divide-by-2 function that produces the desired Doppler compensated 1.023-MHz clock for the G1 and G2 registers. The second stage also provides the means for an externally controlled ±½-chip advance (add an extra clock) or retard (remove an extra clock) phase change to the replica C/A code generator during Vernier searches. The C/A code PRN number is sent to the phase select logic (below the G2 register) as part of the channel initialization procedure. The phase select logic translates this PRN number to select the correct G2 tap delay (and extended PRN number features) for the desired C/A PRN number. The output of that tap combination on the G2 register (or equivalently, the delay added to the G2 register) is added to G1 register output to synthesize the replica C/A code.

With the code setter design described above, the C/A code setup process works as follows. In accordance with a future time delay equal to a fixed number of code NCO reference clock cycles later, the code accumulator value for that future time is loaded into the code setter. This value matches the desired C/A code time after the scheduled time delay. The value for the code setter is computed just as though the 1,023 state C/A code generator had the same linear counting properties as a 1,023-bit counter. The code setter begins counting on the scheduled time delay, starting with the loaded count value. The code setter sets the G1 and G2 registers when the counter produces a carry output that sets the G1 and G2 registers to their initial states. As a result, the C/A replica code generator phase state matches the code accumulator GPS time state and is synchronized to the code accumulator thereafter. When the receiver is tracking the SV after initialization, the code setter process can be repeated as often as desired without altering the C/A replica code generator phase state, because both the code accumulator and the code generator are ultimately synchronized by the same reference clock, the code NCO clock. If the receiver is in the search process, the C/A code advance/retard feature provides the capability to add or remove clock cycles in half-chip increments. The code accumulator must keep track of these commanded changes. If the receiver can predict the satellite transmit time to within a few chips during the search process, it can use the code setter to perform a direct C/A code search. This condition is satisfied if the receiver has previously acquired four or more satellites and its navigation solution has converged. Ordinarily, all 1,023 C/A code chips are searched.

Some commercial C/A code receiver designs do not use a code NCO, but instead propagate the code generator at the nominal spreading code chip rate between code loop updates, tolerating the error build up due to code Doppler and ionospheric delay changes. Instead of the code NCO, a counter with a fractional chip advance/retard capability is used to adjust the phase of the C/A replica code generator in coarse phase increments. This results in a very low resolution code measurement (large quantization noise) and a noisy pseudorange measurement in comparison to the code NCO technique. The algorithm for the code accumulator output to the C/A code setter is (note that the equals sign in the algorithm means "is replaced with"):

$$G = \text{remainder of } \left[\left\{\text{whole part of } \left[(X1/10)\right]\right\}/1023\right] \tag{8.99}$$

where G = future scheduled C/A code time value sent to the code setter and X1 = future scheduled GPS time of week in P-chips ($0 \cdot X1 \cdot 15{,}344{,}999$).

An alternative design to the code setter technique timing technique that would be more suitable for an SDR replica code generator design is to precompute and store the actual 1,023 10-bit C/A code sequences for every SV PRN number and store these as 1,023-bit entries in a C/A code table. Then use the PRN number as the index to these tables. When the SDR channel is activated and initialized, the 1,023-bit sequence of the selected PRN number would be transferred into a 1,023-bit holding register. A 1,023-bit circular shift register in the activated SDR channel becomes the replica C/A code generator designed so that every time a reset command is received from the code setter, the 1,023-bit holding register synchronously transfers its contents into the circular shift register. At the instance of the future code clock that corresponds to the computation of G, the holding register contents would be parallel transferred into the circular shift register C/A code generator. This instantly aligns the replica code generator to the C/A code portion of the code accumulator.

8.10.1.6 Resolving Ambiguity in Code Transmit Time

Every open GNSS signal has an ambiguous transmit time due to its PRN code length (i.e., none are 1 week long). This results in an initial ambiguity in the code accumulator that is resolved by various means but always with information provided in the navigation message data as described in the relevant interface specification. In the case of the C/A code, the ambiguity is resolved by reading the navigation message handover word and then placing this into the code accumulator in the correct format and at the correct epoch. This step is preceded by bit and frame synchronization after the signal is first acquired. The technique described here for resolving the C/A code ambiguity can be used as a model for other GNSS codes beginning with the preparation of a timing diagram.

Figure 8.79 [81] illustrates a timing diagram for the C/A code that is used to determine the true GPS transmit time. The C/A code repeats every 1 ms and is therefore ambiguous every 1 ms of GPS time (about every 300 km of range). There is a handover word (HOW) at the beginning of every one of the five subframes of the satellite navigation message. The HOW contains the Z-count of the first data bit transition boundary at the beginning of the next subframe. This is the first data bit of the telemetry message (TLM) that precedes every HOW. The beginning of this 20-ms data bit is synchronized with the beginning of one of the satellite's C/A code 1-ms epochs, but there are 20 C/A code epochs in every data bit period. At this subframe epoch, the X1 register has just produced a carry to the Z-count, so the X1-count is zero. The C/A code ambiguity is resolved by setting the Z-count to the HOW value and the X1-count to zero at the beginning of the next subframe. In practice, the actual values of the Z-count and X1-count are computed for a near term C/A code epoch without waiting for the next subframe.

The Z-count and X1-count will be correct if the GNSS receiver has determined its bit synchronization to within 1 ms or better accuracy. This level of accuracy will

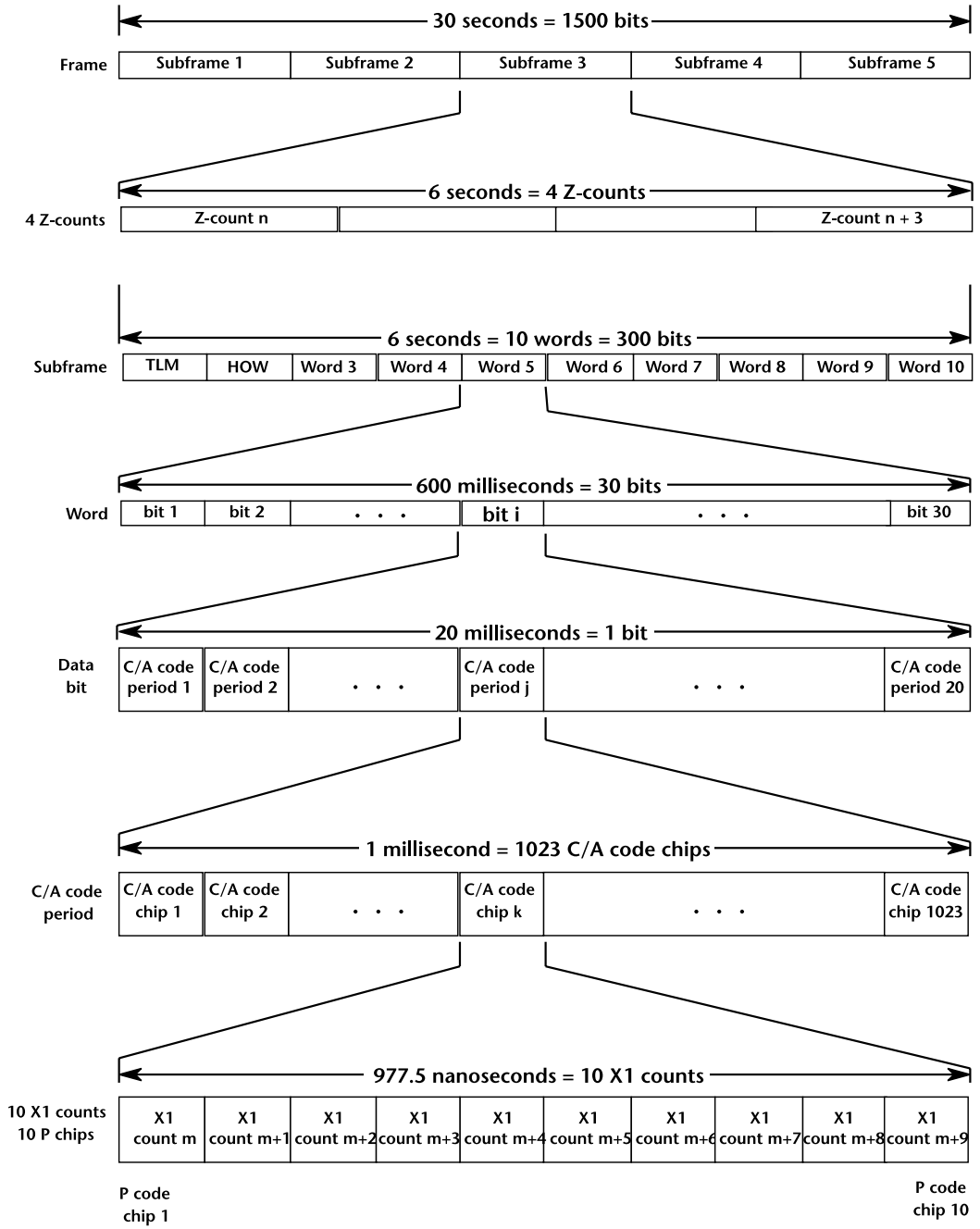


Figure 8.79 GPS C/A code timing relationships.

perfectly align the 1-ms C/A code epoch with the 20-ms data bit transition point of the first bit in the following TLM word. Therefore, the C/A code transmit time will be unambiguous and correct. If the bit synchronization process makes an error in the alignment of the 1-ms replica C/A code epoch with the 20-ms data bit epoch, then the X1-count will be off by some integer multiple of 1 ms.

The original intention of this design was that the receiver would attempt to handover to P(Y) code, and if this fails, the receiver would try the handover again

with 1-ms changes in the value used for X1 and then 2 ms changes, before performing bit synchronization again. This is because this can take 6 seconds or longer and prevents processing of GPS measurements until it is successful. So successful handover to P(Y) code verifies the bit synchronization process. However, for commercial GNSS receivers attempting to resolve the ambiguity in C/A code, the verification for correct bit synchronization is more difficult. This verification task must be performed by the navigation process. Since 1 ms of GNSS time error is equivalent to about 300 km of pseudorange error, the navigation error will be quite serious. In the unlikely case that every channel makes the identical bit sync error, the navigation position error washes out of the position solution into the time bias solution and the GNSS time is in error by 1 ms. The typical bit sync error manifestations in the navigation solution are unrealistic local-level velocity and elevation computations. The latitude and longitude computations are also unrealistic, but there is usually no boundary condition for comparison. However, the velocity and elevation computations can be compared to acceptable boundary conditions.

The bit synchronization process is a statistical process that is dependent on C/N_0. It will occasionally be incorrect. It will be incorrect almost every time the C/N_0 drops below the bit synchronization design threshold. This situation causes serious navigation integrity problems for C/A code receivers under conditions of signal attenuation or RF interference. This problem is compounded if there is no design provision to adapt the bit synchronization process for poor C/N_0 conditions and/or for the navigation process to check for bit synchronization errors.

Modernized GNSS PRN codes with overlays increase signal acquisition time to resolve their ambiguity with respect to the underlying PRN code but this signal innovation removes the necessity for bit and frame sync and represents a significant improvement in reliability of the PRN code ambiguity resolution. The remaining part of removing the code accumulator ambiguity is simply to match the navigation message equivalent to the C/A code handover word to the correct epoch in the code accumulator.

8.10.2 Delta Pseudorange

The technical definition of delta pseudorange to SV_i is as follows:

$$\Delta\rho_i(n) = \rho_i(n+J) - \rho_i(n-K) \quad \left(\text{m per}\left(\mathbf{J}+\mathbf{K}\right)\times\mathbf{FTF}(n)\right)\text{time interval} \qquad (8.100)$$

where $\rho_i(n + J)$ = pseudorange at J FTF epochs later than FTF(n)(m), $\rho_i(n - K)$ = pseudorange at K FTF epochs earlier than FTF(n)(m), and $J = 0$ or K depending on design preferences (dimensionless).

Even though (8.100) implies that delta pseudorange is derived from the code tracking loop, the result would be a very noisy measurement. Instead, the most precise delta pseudorange is derived from the carrier-tracking loop when operating as a PLL. If the carrier loop is operated as a FLL, the measurement is a delta pseudorange rate.

Assuming PLL operation, the precise delta pseudorange measurement is obtained as a slow function using the output of the carrier tracking loop that is producing smoothed carrier Doppler phase errors that are sent back to the carrier NCO every carrier predetection integration interval T_i for SV_i. The loop filter

carrier Doppler phase error, $\Delta\Phi_{CAi}$, is added to a constant that accounts for any fixed carrier frequency in the incoming baseband signal. This constant is zero for the baseband input shown in Figure 8.13 and is the IF for the baseband input shown in Figure 8.14. It is the unbiased portion of the carrier Doppler phase error, $\Delta\Phi_{CAi}$, as it is output to the carrier NCO from the carrier loop filter that is used for this measurement. It is similar to (but less complex than) using the code accumulator to extract transmit time measurements from the code tracking loop. Assuming SV_i is being tracked in PLL mode, the carrier accumulator maintains an integer cycle count, N_{CAi}, and a fractional cycle count, Φ_{CAi}, of the carrier Doppler phase component that is being sent to the carrier NCO. The carrier accumulator is updated after each carrier loop output to the carrier NCO using the following algorithm (note that the equals sign in the algorithm means "is replaced with"):

$$\begin{aligned}
\Phi_{temp} &= \Phi_{CAi} + f_S \Delta\Phi_{CAi} T_i \\
\Phi_{CAi} &= \text{fractional part of } \Phi_{temp} \\
N_{CAi} &= N_{CAi}\,(\text{last value}) + \text{integer part of } \Phi_{temp}
\end{aligned} \tag{8.101}$$

where

Φ_{temp} = temporary Φ_{CAi} register;
f_S = carrier NCO sample frequency (ADC sample rate) (Hz);
$\Delta\Phi_{CAi}$ = last value of the Doppler component sent to the carrier NCO output
= carrier Doppler phase increment per sample
T_i = carrier loop predetection integration time (s)
= time between carrier NCO updates (s)
N_{CAi} = integer cycles count of carrier Doppler phase since some starting point.

The fractional part of the carrier accumulator, Φ_{CAi} , is initialized to the same state as the carrier NCO at the beginning of the search process, which is typically zero. The integer number of carrier Doppler phase cycles, N_{CAi}, is ambiguous. Since only differential measurements are used, the ambiguity does not matter because the common mode ambiguous count is canceled. For this reason, if the bias term sent to the NCO were included in Φ_{CAi} it would also cancel out but would significantly increase the count capacity of the registers for the IF case. The counter rolls over when the Doppler cycle count exceeds the count capacity or underflows if the Doppler count is in the reverse direction and drops below the zero count. The differential measurement comes out correct if the counter capacity is large enough to ensure that this happens no more than once between any set of differential measurements extracted from the carrier accumulator.

To extract a carrier Doppler phase measurement, N_{CAi}, Φ_{CAi}, for SV_i corresponding to the carrier accumulator, the carrier accumulator contents must not be disturbed, so the measurements must be stored in separate registers after the contents are propagated forward to the nearest FTF(n) by the skew time for SV_i,

T_{Si}, similar to the technique used in the code tracking loop as follows (note that the equals sign in the algorithm means "is replaced with"):

$$\begin{aligned}
&\Phi_{temp} = \Phi_{CAi} + f_S \Delta\Phi_{CAi} T_{si} \\
&\Phi_{CAi}(n) = \text{fractional part of } \Phi_{temp} \quad \text{(cycles)} \\
&N_{CAi}(n) = \text{integer part of } \Phi_{temp}
\end{aligned} \tag{8.102}$$

There is no error due to the measurement propagation process for the carrier Doppler phase measurement because the carrier NCO is running at a constant Doppler rate, $\Delta\Phi_{CAi}$ per sample, during the propagation interval. The precise delta pseudorange is simply the change in phase in the carrier accumulator during a specified time. The equation for converting the carrier accumulator measurements into a precise delta pseudorange measurement is:

$$\Delta\rho_i(n) = \left\{ \begin{matrix} \left[N_{CAi}(n+J) - N_{CAi}(n-K) \right] \\ + \left[\Phi_{CAi}(n+J) - \Phi_{CAi}(n-K) \right] \end{matrix} \right\} \lambda_L \quad \text{(m)} \tag{8.103}$$

where λ_L = wavelength of the L-band carrier frequency (m).

As a design example, suppose the navigation measurement incorporation rate is 1 Hz. The delta pseudorange measurement over that time interval would begin after the previous transmit time measurement and end with the current measurement. To make this the precise change in range over the previous 1-second interval for an FTF period of 10 ms, set $J = 0$ and $K = -100$. Alternatively, if the navigation throughput permits, precise delta pseudorange measurements could be made at a 100-Hz rate, each one representing the precise delta pseudorange in the previous 0.01-second FTF interval using $J = 0$ and $K = -1$. In either case, delta pseudorange should be modeled by the navigation process as a (corrected) change in range over the previous interval, not as an average velocity over the interval.

8.10.3 Integrated Doppler

The measurement of integrated Doppler uses the carrier Doppler phase measurement obtained by (8.102). The integrated carrier Doppler phase for SV_i at FTF(n) can be converted to units of meters as follows.

$$ID_i(n) = \left[N_{CAi}(n) + \Phi_{CAi}(n) \right] \lambda_L \quad \text{(m)} \tag{8.104}$$

Recall that the integer cycle count portion of this measurement is ambiguous. The measurement, when derived from a PLL, is used for ultraprecise differential interferometric GNSS applications such as static and kinematic interferometry applications. When the integer cycle count ambiguity is resolved by the interferometric process, this measurement is equivalent to a pseudorange measurement with more than two orders of magnitude less noise than the transmit time (pseudorange)

measurements obtained from the code loop. The integrated Doppler noise for a high quality GNSS receiver designed for interferometric applications typically is about 1 mm (1-sigma) under good signal conditions. A transmit time (pseudorange) measurement under the same signal conditions will be in the vicinity of 3 orders of magnitude (1m) more noise using C/A code and 2 orders of magnitude (0.1m) using a modernized signal such as L5 or a BOC modulated signal. However, the code noise can be significantly reduced by the carrier smoothed pseudorange technique described in Section 8.10.4. Once the integer cycle ambiguity is resolved, so long as the PLL does not slip cycles, the ambiguity remains resolved thereafter. (Further information on differential interferometric processing and ambiguity resolution is provided in Section 12.3.)

Two GNSS receivers that are making transmit time and carrier Doppler phase measurements on their respective receiver epochs will in general be time skewed with respect to one another. For ultraprecise differential applications, it is possible to remove virtually all of the effects of time variable bias by eliminating this time skew between GNSS receivers (i.e., spatially separated GNSS receivers can make synchronous measurements based on common GNSS time). This is accomplished by precisely aligning the measurements to GNSS time epochs instead of to (asynchronous) receiver FTF epochs. Initially, the measurements must be obtained with respect to the receiver FTF epochs. After the navigation process determines the time bias between its FTF epochs and true GNSS time, each navigation request for a set of receiver measurements should include the current estimate of the time bias with respect to the FTF (a very slowly changing value if the reference oscillator is stable). The receiver measurement process then propagates the measurements to the FTF plus the time bias as a nearly perfect (within nanoseconds) of true GNSS time. These measurements are typically on the GNSS 1-second time-of-week epoch. This synchronization is important for precision differential operation since as little as 1 second of time skew between receivers corresponds to MEO satellite position changes of approximately 4,000m. The differential measurements can be propagated to align to the same time epoch if the GNSS receiver's measurements are time skewed, but not with the accuracy that can be obtained if they are aligned to a common GNSS time epoch within each GNSS receiver during the original measurement process. The carrier Doppler measurement must be corrected for the frequency error in the satellite's atomic standard (i.e., reference oscillator) before measurement incorporation. This correction is broadcast in the satellite's navigation message. The measurement also includes the receiver's reference oscillator frequency error. This error is determined as a common-mode time bias rate correction by the navigation solution. For some applications, it is also corrected for the differential ionospheric delay, but this is usually a negligible error for short baselines.

8.10.4 Carrier Smoothing of Pseudorange

The concept of carrier smoothing the code measurements was first presented in [82] as a method for using the ambiguous but low noise carrier range (integrated carrier Doppler phase) measurements to smooth the noisy but unambiguous code range (pseudorange) measurements. The technique is called the Hatch filter, named in recognition of the author. The Hatch filter is an averaging filter implemented as a recursive filter that produces smoothed pseudoranges. This code loop noise filtering

process is important for static and kinematic interferometry applications because it lowers the position uncertainty thereby reducing the processing time to resolve the ambiguity in the differential integrated carrier Doppler phase measurements. The Hatch filter is corrupted by any cycle slip that occurs and must be reinitialized when that happens. It is only possible to detect the likelihood that a cycle has slipped in the PLL tracking loop with a single frequency receiver, but almost certain and identified detection with a two frequency receiver because of the size of the step change observed in the ionospheric delay. In this case, the cycle slip change can be corrected, but the Hatch filter associated with the cycle slip cannot be corrected in real time. The following Hatch equation is expressed in units of meters based on code and carrier loop measurements from SVi as defined in (8.98) and (8.104):

$$\hat{\rho}_i(n) = \frac{1}{k}\rho_i(n) + \frac{k-1}{k}\left[\hat{\rho}_i(n-1) + \left(ID_i(n) - ID_i(n-1)\right)\right] \quad \text{(m)} \tag{8.105}$$

for $k = n$ when $n \le N$ and $k = N$ when $n > N$ after initializing $\hat{\rho}_i(0) = \rho_i(0)$ where

$\rho_i(n) = SV_i$ raw pseudorange measurement at epoch(n) (m);
$\rho_i(n) = [T_R(n) - T_{Ti}(n)]c$ as defined in (8.98) (m);
$ID_i(n)$ = integrated carrier Doppler phase as defined in (8.104) at epoch(n) (m);
$\hat{\rho}_i(n)$ = smoothed pseudorange at epoch(n) (m);
$\hat{\rho}_i(n-1)$= smoothed pseudorange at previous epoch (m);
n = smoothing interval of Hatch filter;
N = maximum value of n.

Note that the above definition of the pseudorange measurement for SV_i uses its unique transmit time, $T_{Ti}(n)$, at epoch(n) [called the set time defined as $FTF(n)$ in the previous sections] and converts this natural measurement into pseudorange using the common receive time, $T_R(n)$, that is in common with all pseudorange measurements at epoch(n). That common receive time is typically maintained by the navigation function and at the same floating point precision as the transmit time measurements but updated in FTF increments. The value of N is typically 100 for navigation measurement incorporation intervals of 1 second, but there are ionospheric situations where this is too long. There are papers that have proposed optimal or adaptive variations of the Hatch filter to overcome the code versus carrier divergence problem caused by ionospheric delay [83], but the basic design is easy to implement and has proven effective using an appropriate value for N. Although this filter could be implemented in the slow functions (code and carrier loop) area of the GNSS receiver, it plays no role in the receiver channel operation and should be implemented in the navigation filter area of the receiver prior to measurement incorporation where it is also recommended that the common receive time, $T_R(n)$, be initialized and maintained and the pseudorange measurements be computed using the natural measurements provided by the slow functions of the receiver baseband process.

8.11 Sequence of Initial Receiver Operations

The sequence of initial receiver operations begins with either a cold-start or warm-start power-up condition. The cold-start condition is recognized by the receiver as an initialization mode including built-in test (BIT) and, for high-end receivers, calibration of critical components or signal paths to ensure that the receiver integrity is sufficient to operate reliably. Unless some intelligent external source removes the uncertainties, the cold-start condition requires a sky search for visible satellites that eventually bootstraps the receiver from a total uncertainty condition to an almost total certainty condition with respect to PVT, up to date almanac for all SVs and ephemeris for the SVs being tracked. Warm start typically does not perform BIT, although most designs do run a lowest priority background, but noninterfering, BIT (i.e., testing everything that does not require stopping the normal operation of the receiver acquisition and tracking processes).

Warm start has much less uncertainty about PVT than cold start. The warm-start condition will typically have the benefit of approximate time (from a low power time-keeper) that was accurate before the previous power-off and it has the previous position and velocity information from the navigation state and other information such as the reference oscillator frequency offset, almanac and the ephemeris with age of data time tags for all previous SVs tracked. So it is reasonable for warm start to use this information to determine which satellites should be visible and then proceed to acquire them. The search engine is used in either case to rapidly acquire the first four SVs that also rapidly removes virtually all of the PVT uncertainty in the receiver. Subsequent SV acquisitions are performed with the search resources that are available in every GNSS receiver channel for all low uncertainty acquisition or reacquisition conditions.

The intended operational environment of the receiver plays a key role in the efficiency of its initial acquisition process and the designed robustness of its tracking loops. Other refinements in the design of the GNSS receiver also depend on the intended operational environment. For example, if the operational environment is stationary (e.g., in a building with a roof-top antenna) and that assurance is a user-provided option in the receiver design, then substantial robustness and accuracy improvements can be achieved using precise carrier aiding that would not otherwise be available. If the operational environment is high precision under low dynamics, for example, precision farming, then substantial robustness is achieved without the necessity for velocity aiding while centimeter-level accuracy is achieved by the built-in provision of a special receiver that shares a common wideband antenna and receives corrections from a geostationary SV. These corrections are obtained by privately operated worldwide ground-based reference stations with precisely located antenna phase centers that continuously observe the errors in the same SVs used by their clients. If the operational environment involves potential high dynamic stress and the need for precise attitude control (e.g., kinematic-based aerial mapping), then precise velocity aiding and attitude determination are provided by the synergism between an IMU and the GNSS receiver.

However, there is substantial commonality in the sequence of initial receiver operations of all GNSS receivers. The ultimate goal of all initial receiver operations is to get the receiver channels into the steady-state tracking condition with four

or more SVs. The BIT, initialization, and signal acquisition performed after cold start and the better-informed warm-start signal acquisition are other examples of this commonality. There is usually a built-in almanac backup in every autonomous GNSS receiver to speed up time to first fix in case the almanac obtained from the space segment is not available from a previous receiver operation. That operation would have to be of sufficient duration to permit this data to be received from the space segment. For example, the legacy GPS C/A code almanac takes about 12.5 minutes to acquire, assuming no failures in the error detecting/correcting navigation message data process.

The major differences are not in the initial receiver operations, but are in the parameters that define the worst-case operating conditions during signal acquisition, such as signal conditions, user dynamics, Doppler range due to user velocity, and reference oscillator frequency offset. There are usually different levels of tolerance of how fast the receiver will acquire the first four satellites (i.e., user patience level on time to first fix under cold-start and warm-start conditions) and that is dependent on the search engine design and the uncertainties in the receiver at the time. The search engine operation is described in Section 8.4.3.1, ultrafast FFT-based search techniques are described in Section 8.5.5, and Vernier Doppler and peak code search, required before attempting carrier and code loop closure are described in Section 8.5.7. The acquisition process is much faster if the receiver knows which SVs are visible. This requires the following information: (1) an almanac for all SVs of interest; (2) a rough estimate of user position; and (3) a rough estimate of time. If any of these parameters are missing, SV visibility cannot be determined for the benefit of speeding up the search process. Useful SV visibility does not require high precision in any of the three critical pieces of information. If all three are available, as is sometimes the case for warm start and always the case when first-fix has been achieved assuming that the almanac data is assured, then using the user position, the GNSS time estimate and the almanac data, a first-pass estimate of the SV positions and the most suitable set of four SVs chosen. This selection might be on the basis of position dilution of precision (PDOP) estimation or it might be based on some other criteria. For example, if mountains or buildings might block low-elevation SVs, then PDOP would be limited to higher elevation SVs (because PDOP tends to select the three lowest elevation SVs as close to 120° apart as possible and one highest SV). When the constellation has been selected, the search process begins. Using the SV line-of-sight Doppler and the user velocity (if known or the maximum user velocity specification), the total line-of-sight Doppler can be determined. This information is used in the Doppler search pattern for the SV. The range search pattern may be all possible combinations of the PRN code or the actual range uncertainty if that is smaller. If the approximate time and position are known and the ephemeris data has been obtained during a recent operation, the first fix will be more accurate and faster since there is no delay waiting on the ephemeris data transfer from the SV to the receiver channel. For example, it can require up to 30 seconds just to read the ephemeris data from a GPS C/A code signal for the SV following signal acquisition. If the ephemeris is not available for the first fix, the almanac data is ordinarily used until the more precise ephemeris data become available. Reading the GPS C/A code navigation data message to obtain almanac data following signal acquisition takes 12.5 minutes, so this is the reason

for the built-in almanac. The almanac data received from the GPS C/A code signal is intended for SV selection and acquisition (not navigation measurement incorporation). It is valid for several days, whereas the ephemeris data, used for navigation measurement incorporation, begins to deteriorate after about 3 hours. For the best navigation accuracy, the ephemeris data should be updated any time newer data is available from the space segment. The modernized GNSS signals, including GPS modernized signals, have improved the time efficiency of obtaining the ephemeris data (as well as the SV clock correction data) and in some cases there are additional parameters that help to extend the accuracy of the ephemeris data.

As was discussed in Section 8.5.6 (direct acquisition of GPS military signals), a critical piece of information for any receiver is GNSS time. Most modern GNSS receivers have a built-in timepiece that continues to run even when the set is powered down. They also have nonvolatile memory that stores the last user position, velocity and time when the set is powered down, plus all ephemeris data (and its age) for all SVs recently tracked, the most recent almanac, reference oscillator frequency offset, and so forth. These nonvolatile memory features support fast initial acquisition the next time the GNSS receiver is powered up, assuming that the receiver has not been transported hundreds of miles to a new location while powered down (but timepiece running) or that several days have elapsed between operations. The stored ephemeris (if it matches the SV acquired) can be used to compute the first fix if the age of data has not exceeded a specified time limit since the receiver was last powered down.

The sky search is actually a bootstrap mode of operation to get the GNSS receiver into operation when one or more of the almanac, position/velocity, and time parameters are missing or corrupt. The FFT sky search is a remarkable feature made possible by the increased speed of DSPs for any of the GNSS PRN codes that permits the receiver to enter into the navigation mode without any a priori knowledge or any external help from the operator (it can be faster than an operator could key in the most primitive of useful information). Bootstrapping is virtually impossible for encrypted signals such as the P(Y) code or M code without help from open signals (e.g., C/A code) unless the authorized receiver has a precise estimate of system time (see Section 8.5.6).

The sky search mode theoretically requires the receiver to have the capability of searching the sky for all possible PRN codes, in all possible Doppler bins and for all possible code states of each PRN code until at least four SVs are acquired. In practice, only the most favorable GNSS signals will be supported for this bootstrap operation. Using FFT-based searches, the cold start sky search process will typically require a few seconds for the receiver to find four visible SVs from a total uncertainty condition in the navigation process. At first thought, the GPS L1 C/A code would be the best of the favorable choices because of its short length (1,023 chip) code and simple BPSK modulation properties, but the fixed overhead of bit and frame synchronization (that the FFT-based process cannot speed up), plus reading the handover word and the ephemeris data should be compared with other GNSS signal choices with overlay codes that do not require bit and frame synchronization. This is because the FFT-based search does speed up the acquisition of these codes while eliminating the time-consuming bit and frame synchronization process required by the C/A code.

The first four SVs found by sky search are unlikely to provide the best geometric performance. After the first four SVs acquired have provided any missing almanac, position/velocity, and time information (i.e., have reduced uncertainty in the navigation process), the navigation process can then determine which SVs are visible and what is the best subset for navigation. The remaining visible SVs can be quickly acquired if their view is not blocked because the navigation process uncertainty is very small. After any SV is acquired, the major part of the delay before measurement incorporation can take place is the time taken to read their ephemeris data and clock corrections in the SV navigation message data. For all-in-view GNSS receivers that track numerous SVs from multiple constellations simultaneously, good geometry is assured if all SVs in view have been acquired and their measurements incorporated into the navigation solution. This all-in-view feature fully utilizes the significantly improved signal availability provided by multiple GNSS constellations. This provides robustness in the GNSS receiver when multiple signals are temporarily blocked and later reappear so long as four or more remain unblocked (a serious problem driving through urban canyon conditions or an area surrounded by mountains). Reacquisition is almost instant when the signals become unblocked. There is no need to determine the best geometry since, by definition, all-in-view signals that are available are being tracked. The only need is to continually determine which SVs should be visible.

After a receiver channel enters into steady state operation, typically in PLL mode, data demodulation immediately begins (described next in Section 8.12) and special (slow) baseband functions are activated to measure signal quality as well as integrity monitors of the tracking loops (described in Section 8.13). The signal quality measurement is used for making decisions about the tracking loops to hold onto the signal while the integrity monitors assist in the decision-making process and provide information needed when corrections are made.

8.12 Data Demodulation

Data demodulation uses the (prompt) carrier signal in the receiver when it is in stable closed PLL operation. Demodulation can also be performed with stable FLL operation, but is suboptimal in both bit error rate performance as well as in recovery and reinitialization efficiency. The legacy GPS C/A code and P(Y) code signals transmit the same 50-bps navigation message data using binary modulation that is synchronously combined with the PRN spreading codes using an exclusive-or logic gate.

Modernized GNSS signals have unique navigation messages that are sometimes interleaved. For many modern GNSS signals, the binary data bits are encoded into a higher-rate binary symbol stream through a forward error correction (FEC) algorithm. One popular FEC scheme is rate ½ convolutional encoding with constraint length 7. Figure 8.80 depicts this relatively simple encoding technique. This encoding results in two symbols being generated for every 1 bit of data input into the convolution encoder. Figure 8.80 depicts a 50-bps data input that has a 100-sps output. The encoded message data are synchronously combined with the data channel PRN spreading code using exclusive-or logic before being transmitted in a separate data channel. The data channel and the pilot channel have different PRN

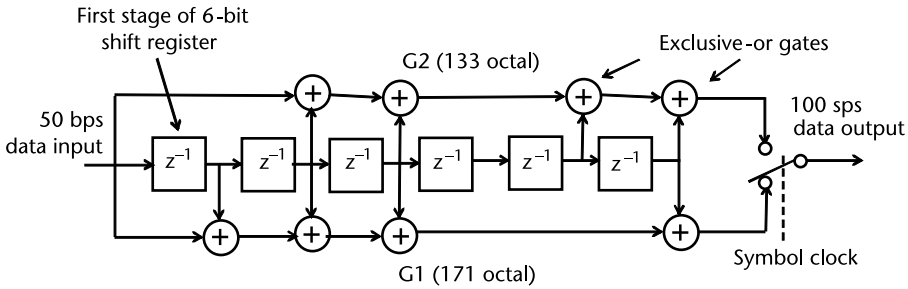


Figure 8.80 Constraint length 7 rate ½ convolutional encoder.

spreading codes and the same carrier frequency. If short synchronization codes are used, these are also different in the data and pilot channels.

Figure 8.20 illustrates how the predetection integration time is eventually phased to approximately align with the data or symbol transition boundaries in the prompt data channel. How accurately this transition boundary can be aligned is determined by the integrate-and-dump period used in Figures 8.13 and 8.14. This time increment (TINC) should be small compared to the data or symbol period to obtain maximum energy for the detection of each data bit or symbol (e.g., for a 10-ms symbol period, a TINC alignment accuracy of 0.01/32 = 312.5 μs will integrate about 97% of the available symbol energy, but the remaining 3% integration time on the split boundary may be counterproductive). The TINC alignment to the incoming SV signal data transition boundaries is maintained by the set time sync signal that occasionally advances or retards the dump phase, that, in turn, changes the value of T by ±1 TINC for one period. The next code and carrier tracking loop update must include ±1 TINC in the effective value of T used for in each loop filter. The alignment is managed by and maintained in the code accumulator of each channel based on its X1 phase, so the set time sync command originates there. Legacy and modernized data demodulation are described next in the context of the source and management features of the receiver data demodulation signal have been described.

8.12.1 Legacy GPS Signal Data Demodulation

Since there is data modulation present in the legacy GPS signals, a Costas PLL is used to track the carrier Doppler phase and to detect the data bits in the SV data message stream after the transition boundaries are determined by a process called bit synchronization, hereafter called bit sync. Keep in mind that the legacy C/A code has a 1-ms period and the 50-Hz navigation data message has a 20-ms period that is aligned with a C/A code transition boundary. If the receiver time uncertainty is greater than 1 ms, then the data transition boundary is ambiguous, so bit sync must be performed before the 50-Hz data can be successfully demodulated.

8.12.1.1 Bit Sync

When the data transition boundaries are not yet known, there is a higher likelihood that there will be cycle slips when the carrier loop is in PLL, so forced FLL operation can be more reliable during bit sync. The following C/A code technique can be

performed in FLL or PLL and is easily adapted for other GNSS signals that require bit sync [84]:

1. Initialize 20 cells (the bit sync accumulator) indexed with the bit sync counter, $K = 0$ to 19, using an arbitrary starting phase, $K = 0$, with respect to the unknown data transition boundary with all cells initialized to zero.
2. Referring to Figure 8.18, collect I_P, Q_P samples every C/A code epoch (1 ms) and then associate the first sample with cell $K = 0$, the second sample with cell $K = 1$, and so forth, up to $K = 19$, and add 1 to the associated cell every time a sign change is sensed at this phase; otherwise, proceed, modulo 20.
   a. If in FLL, sign changes are detected by phase discrimination of I_{Pi}, Q_{Pi} in the current 20 ms dwell (i) and for I_{Pi-1}, Q_{Pi-1} in the previous 20 ms dwell ($i - 1$), for example,

$$\delta_{fi} = \tan^{-1}\frac{Q_{Pi}}{I_{Pi}} - \tan^{-1}\frac{Q_{Pi-1}}{I_{Pi-1}}$$

   b. If in PLL, sign changes are detected by comparing the sign of I_{Pi} in the current 20-ms dwell (i) with the sign of I_{Pi-1} in the previous 20 ms dwell ($i - 1$).
3. Figure 8.81 depicts the successful outcome of the 20-point bit sync histogram after several iterations where the count in one cell (the cell where the data transition boundary is located) has reached or exceeded the upper (pass) bit sync threshold, N_{BSp}. The bit sync counter K is reset to zero at this pass index so that $K = 0$ corresponds to the data transition boundary. It continues to be incremented modulo 20 by the C/A code epoch until no longer needed (i.e., after the code accumulator has been initialized).
4. Two possible failure modes result in early termination of this bit sync process:

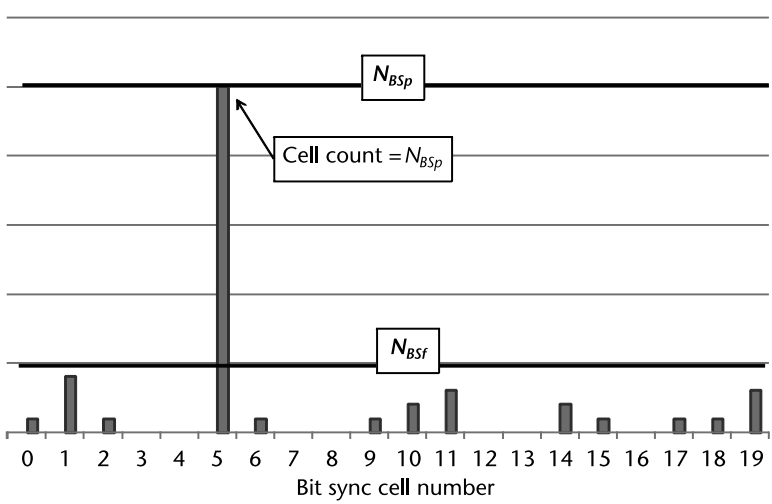


Figure 8.81 Bit sync histogram after successful location of data transition boundary.

a. Carrier lock is lost, so bit sync is abandoned until carrier lock has been achieved.
b. Two or more cell counts exceed the lower (fail) bit sync threshold, N_{BSf}. This failure is an indication of low C/N_0, so bit sync is reinitialized, possibly with an increase in this lower threshold.

The upper (pass) threshold, N_{BSp} is the expected number of data bit transitions in the total bit sync time interval, T_{BS}. Although the number of 1s and 0s in a typical SV data message data stream is not equal, this is a reasonable assumption for a rough approximation. So assume that on average in 50 data bit intervals per second, there will be approximately 25 data edge transitions. The upper (pass) threshold is therefore set at $N_{BSp} = 25T_{BS}$, recognizing that it will likely take longer than T_{BS} seconds to complete the bit sync process. Four to six seconds is a typical range for the bit sync time period, T_{BS}, but the successful bit sync process terminates when the upper (pass) threshold is reached by one cell without more than one cell passing the lower (fail) threshold (i.e., T_{BS} is a variable).

The lower (fail) threshold, N_{BSf}, is set to a value that is greater than or equal to $50T_{BS}P_{esc}$ where P_{esc} is the probability of making an error in determining sign change. This probability is determined using the following equations:

$$P_{esc} = 2P_e\left(1 - P_e\right) \tag{8.106}$$

where P_e is the probability of a data bit error for a given C/N_0 and predetection integration time, T.

For FLL operation [85]:

$$P_e = \frac{1}{2}e^{-(C/N_0)T} \tag{8.107}$$

For PLL operation [86]:

$$P_e = \text{erfc}'\left(\sqrt{2C/N_0T}\right) \tag{8.108}$$

where

$$\text{erfc}'(x) = \frac{1}{2\pi}\int_x^\infty e^{-y^2/2}dy \tag{8.109}$$

The number of entries in any cell, N_{BS}, has a binomial distribution, so that in the correct cell over T_{BS} seconds the average number of data bit transitions is $25T_{BS} = N_{BSp}$ (the pass threshold), and in any other cell over T_{BS} seconds, the average number of data bit transitions is $50T_{BS}P_{esc}$. The standard deviation of N_{BS} in any cell is [84]:

$$\sigma_{N_{BS}} = \sqrt{50T_{BS}P_{esc}\left(1 - P_{esc}\right)} \tag{8.110}$$

The thresholds and T_{BS} are selected to provide a safe 3-sigma spread between pass and fail thresholds at the desired C/N_0 using an estimate of T_{BS} as follows [84]:

$$25T_{BS} - 3\sqrt{50T_{BS}P_{esc}\left(1-P_{esc}\right)} \geq NBS_f \geq 50T_{BS}P_{esc} \tag{8.111}$$

Achieving reliable bit sync with the shortest T_{BS} requires optimizing the bit sync thresholds through extensive testing with real navigation message data for a range of C/N_0 conditions, but even an optimized bit sync time represents a substantial portion of time to first fix.

8.12.1.2 Detecting Data Bits in PLL and Frame Sync

The data demodulation process starts with the typical sequence of bit sync, bit detection, frame sync, and then message data processing. After successful bit sync, the data transition boundaries are known, so the Costas carrier tracking loop can now be operated at the maximum predetection integration time, T, equal to the data bit period (20 ms for C/A code). This usually provides robust carrier tracking in PLL at typical C/N_0 levels, so the data bits can also be reliably detected in PLL. Referring to Figure 8.18, the $\overline{I}_P$ samples are accumulated for one 20-ms data bit interval (between transition boundaries) and compared to a threshold for detection. The sign of the result is the detected data bit.

After bit detection, the frame sync process begins. Refer to Figure 8.79 for the organization of each subframe and the location of the telemetry (TLM) word at the beginning of each subframe of the navigation message data. In the frame sync design shown in Figure 8.82, a 32-bit data register is activated with register bits 0 to 31 as designated by the register index shown in the lower part of the figure. Then the demodulated bits (shown in the upper part of the figure) are shifted from right to left in this register to form words. Since the C/A code navigation message words are 30 bits in length, they are held in bits 0 to 29 of the data register. Bits 30 and 31 of the register hold bits 29 and 30 of the previous word. These two bits are required by the parity algorithm and are referred to as D29* and D30* in [67].

Initially, a bit-by-bit pattern test is performed that eventually finds the preamble at the start of every subframe. The preamble is an eight-bit binary pattern, 10001011 ($8B_{16}$ in hexadecimal notation), that is in the first eight bits of every subframe [i.e., the first 8 bits in the 30-bit telemetry (TLM) word of every subframe]. The bit-by-bit preamble pattern search begins with the current bit and the previous 7 bits as shown in Figure 8.82. The pattern match is in data register bits 29 to 22 corresponding to the TLM preamble bits 1 to 8 using two search patterns: $8B_{16}$

previous word		8-bit TLM preamble ($8B_{16}$ = upright, 74_{16} = inverted)									SV data message bits			
D29*	D30*	1	2	3	4	5	6	7	8	...	27	28	29	30
31	30	29	28	27	26	25	24	23	22	...	3	2	1	0
32-bit data register bits														

Figure 8.82 Frame sync register searching for preamble in telemetry (TLM) word.

to determine if the Costas loop is demodulating data bits upright or its inverted counterpart, called one's compliment, 74_{16} if inverted. This possible inversion is because of the 180° phase ambiguity in any Costas PLL after it closes resulting in the detected data bits being normal (upright) or inverted. A match probably means that the TLM message follows in the data register bits 21 to 0 (corresponding to data word bits 9 to 30). However, there is a low probability that a match could be found within the subframe, so the first match does not complete the frame sync process. When a match is found, the next word is considered to be the HOW. The bit stream is processed upright if the preamble was matched with $8B_{16}$. It is processed inverted if the preamble was matched with 74_{16}. If the assumed HOW has good parity, it is examined for a valid time-of-week (TOW), subframe ID, and agreement between the TOW and the subframe ID. If these checks are successful, the TOW is used to calculate the current Z-count (GPS transmit time with a 1.5-second least significant bit). When the Z-count in the code accumulator is correctly set, precise transmit time measurements can then be made for this receiver channel. If any of these sequences fail, the frame sync process continues to search for another match. Note that a successful frame sync not only finds the 300-bit subframe boundaries, but it also provides word sync (i.e., it finds the 30-bit word boundaries).

After frame sync successfully initializes the Z-counter in the code accumulator frame sync continues to examine the data message truncated Z-count in the HOW every 6 seconds in each subframe and compares this with the Z-count in the code accumulator. Data demodulation follows frame sync in accordance with the data format specified in [66]. Typically, the receiver channel performs data demodulation at the bit and word levels including parity checking, then sends those words to a data block processing function in the receiver control function that extracts, formats and separates parameters required by the receiver control and the navigation processes. A cycle slip in the Costas PLL is actually a half-cycle slip that causes a reversal of this polarity that requires corrective action before data demodulation can proceed. Typically, the frame sync logic runs continuously as insurance that a cycle slip has not occurred.

The probability of bit error for the C/A code (as well as P(Y) code) signals is:

$$P_b = \frac{1}{2} erfc\left(\sqrt{\frac{E_b}{N_0} \cdot \cos\phi}\right) \tag{8.112}$$

where

E_b = energy per bit (J);

N_0 = noise power in 1-Hz bandwidth (W/Hz);

ϕ = phase error (assume zero in PLL)

$$E_b = \frac{C}{R_b}$$

where R_b = data bit rate (in bps).

Assuming PLL operation and replacing E_b in (8.112), the probability of a data bit error is:

$$P_b = \frac{1}{2} erfc\left(\sqrt{\frac{C/N_0}{R_b}}\right) \tag{8.113}$$

where

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_{t=x}^{\infty} e^{-t^2} dt$$

is the complementary error function.

8.12.2 Other GNSS Signal Data Demodulation

As shown in Figure 8.19, the typical modernized GNSS receiver will use the more robust pilot channel for acquiring and tracking the incoming signal, then slave the data channel with the pilot channel. The pilot and data channels share the same carrier signal. Only the prompt signal is used in the data channel replica code generator (synchronized by the pilot replica code generator including overlay codes, if applicable). GNSS signals with overlay codes have the advantage that the level of ambiguity resolution achieved as a by-product of signal acquisition in the pilot channel automatically synchronizes the phase of the data channel overlay code that also resolves the symbol transition boundaries. In the case of the L2 CL pilot channel that has no overlay code, its PRN code length is X1 = 1.5 seconds that resolves ambiguity sufficiently to align the symbol transition boundaries in the L2 CM data channel.

8.12.2.1 Detecting Data Bits in PLL

Each symbol is detected by continuing the integration of the in-phase data channel signal, $\overline{I}_{Pd}$, shown in Figure 8.19, beginning at the current sample transition boundary and ending at the next transition boundary, so that one sample period has been integrated. As will be described later, the symbol values are retained as a soft decision instead of being detected (hard decision) as a 1 or 0 at this point. As shown in Figure 8.80, for many GNSS signals the original data bits are convolutionally encoded into symbols in the SV before they are transmitted. Other FEC schemes may also be encountered (e.g., LDPC encoding for GPS L1C). When FEC is used, an inverse decoding process must be used to read the data. The decoding process using soft decisions significantly improves the bit error rate in comparison to performance with hard decision techniques.

8.12.2.2 Viterbi Decoder

For GNSS signals that use convolutional encoding of the navigation data, one of the most efficient decoding techniques is the Viterbi algorithm (VA) [87] usually called the Viterbi decoder, named in honor of the inventor. The rate ½ constraint length 7 Viterbi decoder takes in two sequential symbols per encoded data bit and there is a delay of six symbols before the next (original and error corrected) data bit is

produced. There is also a residual of the previous six soft decision symbols remaining in the decoder. These pairs of input symbols must be synchronized to the replica X1 boundary to ensure that the Viterbi decoder is making decisions on the original data bit using the correct symbol sequence starting point. If the GNSS signal has a pilot component, a pure PLL may be used and there is no ambiguity in the signs of the symbols, that is, they are always upright. (If it is elected to independently track the data channel, then the Costas PLL discriminator must be used so the uncertainty of the symbols being upright or inverted must be resolved.)

The convolutional encoder is much easier to implement than is the extremely complex Viterbi decoder. Because the rate ½ constraint length 7 Viterbi decoder is used for so many communications applications, the maturity level of this specific VA technology is such that numerable design resources are available. Reference [88] provided basic theory and design insight into the VA. Reference [89] was an application note on implementing the VA in a commercial DSP. Reference [90] provided insight into the constructs of simulating and testing the VA design and [91] provided HDL code generation support for checking, generating, and verifying the Viterbi decoder HDL code that the designer generates using a fixed-point model. It also discusses the settings that can be used by the designer to alter the generated HDL code. However, the designer must have an HDL Coder (trademark of MathWorks) license.

The Viterbi decoder design is more simple using hard decisions but the resulting data bit error performance is much better if soft decisions are used (i.e., the value for each symbol is used rather than making a 1 or 0 decision on each symbol prior to VA decoding). This is because a hard decision makes a premature decision rather than taking full advantage of the power of convolutional decoding in the decision trellis. For FEC ½ rate constraint length 7 convolutional code using soft decision Viterbi decoding, the bit error rate is upper-bounded for values of interest using [92]:

$$P_b \le \frac{1}{2} \cdot \left(36 \cdot D^{10} + 211 \cdot D^{12} + 1404 \cdot D^{14} + 11633 \cdot D^{16}\right) \tag{8.114}$$

where

$$D = \exp\left(-\frac{1}{2} \cdot \frac{E_b}{N_0} \cdot \cos^2 \phi\right) \tag{8.115}$$

N_0 = noise power in 1-Hz bandwidth
ϕ = phase error (assume zero in PLL)
$E_b = \dfrac{C}{R_b}$

where R_b = data bit rate in bps.

Assuming PLL operation and replacing E_b in (8.115):

$$D = \exp\left(-\frac{1}{2 \cdot R_b} \cdot \frac{C}{N_0}\right) \tag{8.116}$$

It should be noted that the structure of (8.114) depends only on the coder and decoder characteristics (i.e., the rate ½ constraint length 7 Viterbi decoder using soft decision on each symbol, while the parameter D is separately dependent on the type of jamming and the detection metric). In (8.115) white noise is assumed as the type of jamming and PLL operation is assumed as the detection metric. The parameter D is sometimes called the Hamming distance.

8.12.3 Data Bit Error Rate Comparison

To illustrate the benefits of FEC, Figure 8.83 compares the probability of error for 50-bps data bit rates of legacy GPS signals such as C/A and P(Y) code that use binary modulation detected by hard decision using (8.113) and modernized signals such as L2, L5 or M code with rate ½ constraint length 7 convolution coding that is decoded by a compatible soft decision Viterbi decoder using (8.114). It should be noted that both equations presume that the PLL is perfectly tracking carrier phase without any slips. At low C/N_0, phase tracking errors degrade the bit error rate. If the C/N_0 is too low or if the signal dynamics are too severe, then as discussed in Section 8.9.7, the PLL is unable to track carrier phase and data demodulation can no longer be performed. Therefore, the receiver channel should cease data demodulation when C/N_0 is observed to be too low for an acceptable probability of data bit error rate, assumed in Figure 8.83 to be 1 in 10^{-6}.

Inspection of Figure 8.83 shows that at the acceptable probability of data bit error rate, the modernized signals have more than a 5-dB margin over the legacy signals. This more than compensates for the 3-dB loss of $(C/N_0)_{dB}$ if the data channel in L2 M and L5 I5 signals. The net 3-dB gain in carrier tracking threshold more

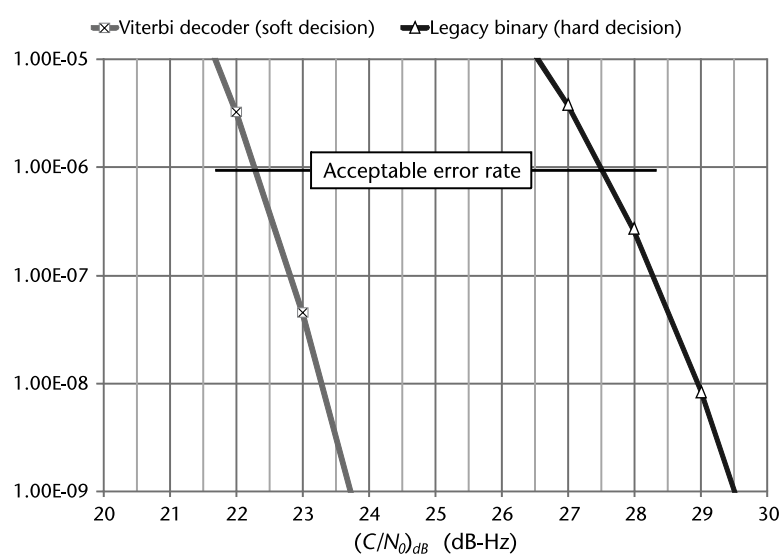


Figure 8.83 Probability of error comparisons for 50-bps data bit error rates.

than compensates for the 3-dB loss of $(C/N_0)_{dB}$ in the pilot channels of L2 CL and L5 Q5 signals. More powerful LDPC FEC encoding provides lower data bit error probabilities with sufficient margin to justify providing only one-fourth of the power (a 6-dB loss in $(C/N_0)_{dB}$) in the interoperable L1C data channel so that the pilot channel has three-fourths of the carrier power (a 1.25-dB loss in $(C/N_0)_{dB}$) for even more robust carrier tracking performance.

8.13 Special Baseband Functions

There are numerous special baseband functions that must be implemented in a GNSS receiver design, but the following design examples are among the most important.

8.13.1 Signal-to-Noise Power Ratio Estimation

An accurate measure of signal-to-noise power ratio (S/N) in each receiver tracking channel is the most important special baseband function in the receiver baseband area because it provides receiver channel tracking status. Therefore, it is the basis for many receiver channel state transitions into appropriate operating states.

The basic *S/N* meter design in Figure 8.84 is based on subtracting the mean of the noise power estimate in the lower path from the mean of the signal plus noise power in the upper path and then forming the *S/N* power ratio using the same noise estimate for the denominator. Referring to the figure, the prompt $\bar{I}_P$ and $\bar{Q}_P$ signal samples and $\bar{I}_N$ and $\bar{Q}_N$ noise samples in the same receiver channel comprise the upper path and lower path inputs, respectively. These input signals have been integrated (and normalized) in the fast functions part of the receiver channel over the time period T seconds. They are further integrated and normalized using K samples to reach the current carrier tracking loop predetection integration time,

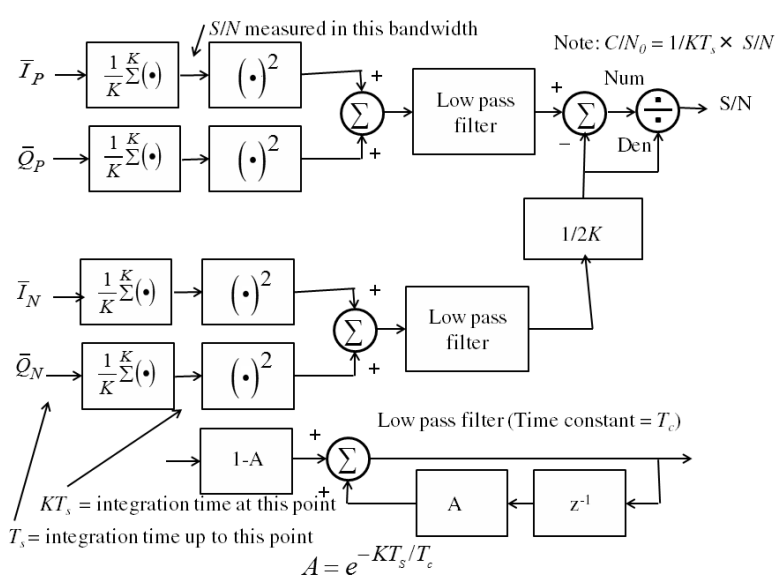


Figure 8.84 Basic signal-to-noise power ratio meter.

KT_s. As noted in the figure, the *S/N* is measured in the bandwidth established by this integration time (i.e., $1/KT_s$ Hz). The *I* and *Q* signals in both signal paths are formed into two power envelopes that are passed through identical lowpass filters. Note that the signal processing functions prior to the lowpass filters in Figure 8.84 are virtually identical to their counterparts in Figure 8.18. The lowpass filters provide the estimate of the mean power in each signal and they must be initialized with the best estimate of the signal-to-noise ratio by the peak search process. Each filter is a simple recursive low pass filter design as shown at the bottom of Figure 8.84 based on the parameter value $A = e^{-KT_S/T_c}$ where T_c is the desired time constant of the filter (not to be confused with the reciprocal of the spreading code rate, R_c). Also note that the squaring process makes the design equally effective for Costas signals so long as the predetection integration time is compatible. It is also equally effective for both PLL and FLL operations. The mean noise power from the lower path, scaled by 1/2*K*, is subtracted from the mean signal plus noise in the upper path. This process yields a good estimate of the mean signal power for the numerator (Num) of the last stage divider. The scaled mean noise power estimate is also used for the denominator (Den) in the last stage divider. The divider output is the estimated signal-to-noise power ratio, *S/N*. As shown in the top of the figure, it becomes a C/N_0 (ratio-Hz) meter when the *S/N* estimate is divided by the current predetection integration time, KT_s.

A higher accuracy and wider range C/N_0 meter design is shown in Figure 8.85. It has similarities to the basic design in that the *S/N* ratio is still measured in the same bandwidth set by the first stage *K* normalized integrations, but in this case the design is based on the variance of the noise power estimate, *N*, from the lower path that has been scaled to equal the variance of the noise in the signal plus noise power, *S* + *N*, in the upper path. In the lower path, the scaled and squared noise terms are added and the result multiplied by *K* before being passed to the divisor

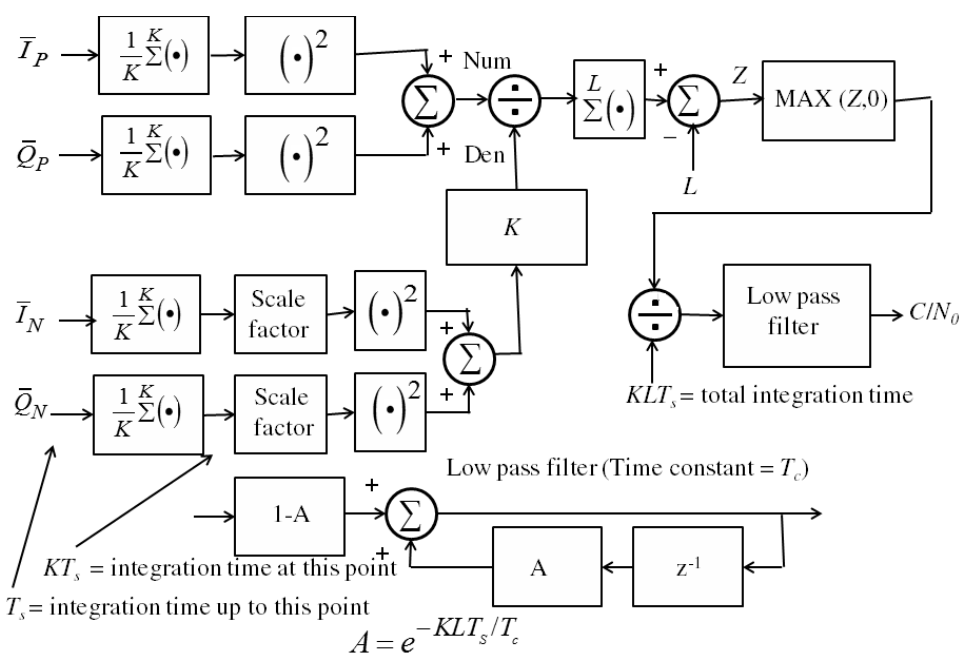


Figure 8.85 Accurate wide range C/N_0 meter.

as shown in the figure. The division using $S + N$ in the numerator (Num) and the scaled noise term, N, in the denominator (Den) is performed first to form $S/N + 1$ followed by L integrations. Then L is subtracted from the result to remove the constant to produce $Z = S/N$. The next stage places a lower bound of zero on Z when this ratio goes negative. Then the bounded value of Z is divided by the total integration time KLT_s to convert it to a C/N_0 estimate. This is fed to the low pass filter to produce the mean value of C/N_0 using the filter parameter $A = e^{-KLT_S/T_c}$ where T_c is the desired filter time constant. Note that only one lowpass filter is required in this design and must be initialized by the best estimate of that ratio by the peak code search process that is used by all forms of signal acquisition or reacquisition.

The scaling of the noise terms in the lower path of Figure 8.85 is a key part of this design. As observed in the figure, both noise terms are scaled appropriately prior to being squared. The scale factor is based on the integration time T that has already been performed on the signals at the input plus other factors such as the spreading code rate of the replica code generator and the type of interference that may be present. Wideband noise is assumed unless CW has been detected by the receiver situational awareness function. Table 8.25 shows typical noise meter scale factors for a prior version of this design in which only $\bar{Q}_N$ was used in the noise estimate so the denominator scale factor was multiplied by $2K$ instead of K as shown in Figure 8.85. These scale factors assume that $\bar{Q}_N = \bar{Q}_P$ when there is no correlation with the incoming signal. Assuming in the present design that $\bar{I}_N = \bar{I}_P$ and $\bar{Q}_N = \bar{Q}_P$ when there is no correlation with the incoming signal, these scale factors remain the same.

Based on these scale factors, Table 8.26 shows typical values of the remaining parameters of the C/N_0 meter design in Figure 8.85. When properly tuned, this design provides estimates that are accurate to ±0.5 dB 99% of the time over the C/N_0 range of 30 to 50 dB-Hz and 50% of the time down to 20 dB-Hz. It diverges to about +1.5 and –2.3 dB 99% of the time at $C/N_0 = 20$ dB-Hz.

Initialization of the memory in the lowpass filter in the C/N_0 meter is very important when transitioning from peak search (shown in Figure 8.43) into the tracking mode using the value of the accumulated envelopes and the noise standard deviation from the peak search algorithm. The lowpass filter memory is initialized with $\frac{A^2}{2\sigma^2 T_S}$, where $A^2 = (E_{\max}/N)^2 - 2\sigma^2$. E_{max} is obtained from the maximum

Table 8.25 Typical Noise Meter Scale Factors for Accurate C/N_0 Meter

Interference Type	*Code Type*	*Prior Integration Time T (ms)*	*Scale Factor*
Noise	C/A	5	0.001630722
		10	0.001153095
	P(Y)	5	0.001527985
		10	0.001080449
CW	C/A	5	0.001146000
		10	0.000810344
	P(Y)	5	0.000692690
		10	0.000489806

Table 8.26 Typical Design Parameters for Accurate C/N_0 Meter

Data Wipe-Off	*Data Bit Edge Known*	*Code Loop Integration Time (ms)*	*Input Sample Time T (ms)*	*K*	*L*
No	No	5	5	1	4
No	Yes	20	10	2	1
Yes (proxy for pilot channel)	N/A	20	10	2	1
Yes (proxy for pilot channel)	N/A	320	10	32	15

envelope value found during peak search using a likelihood ratio test of each $E_j = \sum_{k=1}^{N} \sqrt{I_{j,k}^2 + Q_{j,k}^2}$ where N is the number of envelopes used to compute each E_j. The peak search noise standard deviation, σ, is the value calculated for use in the likelihood ratio test.

8.13.2 Lock Detectors

There are many receiver control decisions that are made based on the status of the carrier and code tracking loops. The weaker of the two is the carrier-tracking loop in the desired carrier phase lock condition, so the carrier phase lock detector is described first. Illustrative implementations are described for the GPS signals, but the techniques can be readily adapted to other GNSS signals.

8.13.2.1 Phase Lock Detector

The phase lock detection concept is simple: if the loop is in phase lock, then in Figure 8.18, I_P will be maximum and Q_P will be minimum. The phase of the envelope tends to stay near the I-axis when in phase lock for the pilot channel PLL. For a data (Costas) channel, it transitions 180° between the positive I-axis to the negative I-axis with every data bit sign transition in a data channel. Because there is always noise present, the appearance on an oscilloscope looks like a fuzz ball. The phase noise is often called jitter. As the jitter increases in the PLL tracking loop due to noise, dynamic stress, and so forth, it eventually reaches a level where a cycle will be slipped or complete loss of phase lock occurs. A cycle slip would be observed on an oscilloscope as a rotation followed by resumed phase lock. In the case of a data (Costas) loop, the rotation would be 180° and in a pilot channel, the rotation would be 360° for each cycle slip. It is easier to observe on a pilot channel because the jitter stays on the positive I-axis until it slips or totally loses phase lock. The phase lock detector measures this jitter, using the absolute value of the data jitter and the actual value of the pilot jitter to verify they remain in the phase lock vicinity of the I-axis.

Figure 8.86 is an example of a basic phase lock detector for either a data (Costas) channel or a pilot channel. The difference is simply the need for an absolute value function in the phase lock detector input to remove the 180° inversions caused by the presence of data modulation to remove the limitation on predetection integration time T for the data (Costas) inputs. No absolute value function is needed in a

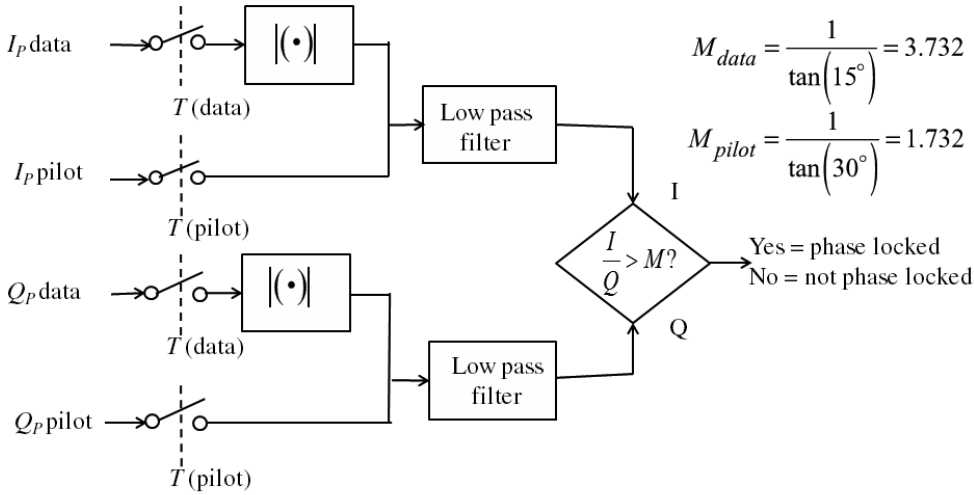


Figure 8.86 Basic phase lock detector design for either a data (Costas) or a pilot PLL channel.

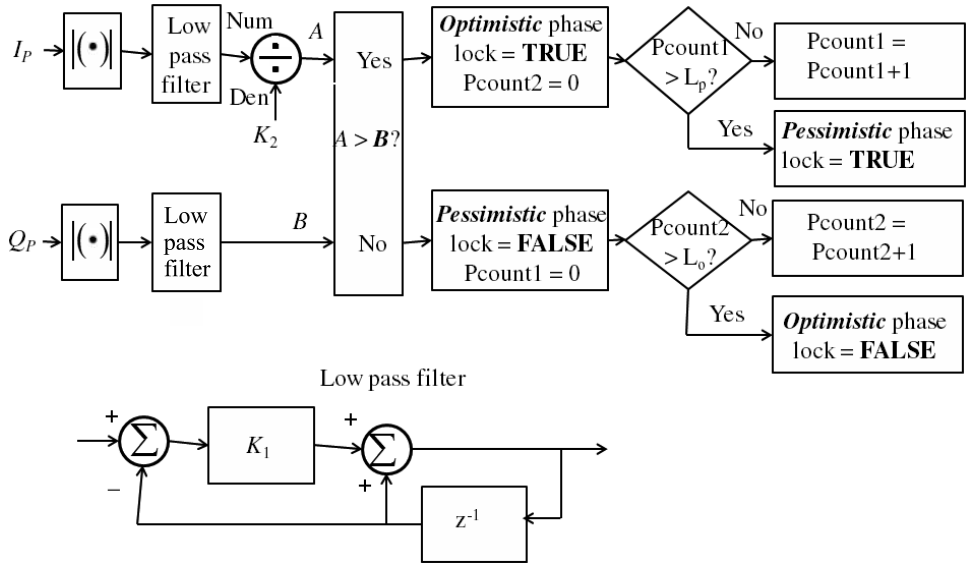
pilot channel phase lock detector. In either case, the I_P and Q_P signals that are used at the input of the carrier tracking loop discriminator (see Figure 8.18) are also sent to the phase lock detector. As indicated in Figure 8.86, these have been integrated for T seconds, so the phase lock detector update period is T. The ratio (I/Q) of the outputs of the low pass filters is tested based on the PLL rule-of-thumb loss of lock threshold using M as the criteria. As shown in the figure, $M_{data} = 1/\tan(15°) = 3.732$ and $M_{pilot} = 1/\tan(30°) = 1.732$ as their respective smallest value of this ratio that phase lock is sustained. However, the rule-of-thumb criteria are not optimum.

The precaution to check for $\tan(0°) = 0$ is necessary prior to avoid a divide-by-zero error when computing this ratio. It is also important to initialize the lowpass filters to zero on first use of the detector after it has been previously disengaged.

Some baseband functions require a higher degree of certainty of phase lock than others since loss of phase lock in the carrier tracking loop will cause erroneous results if the function believes that there is phase lock when in fact there has either been a cycle slip or total loss of phase lock. Figure 8.87 depicts an advanced phase lock detector design that provides an optimistic phase lock indicator that decides quickly and changes its mind slowly, but is not as reliable as the pessimistic phase lock indicator that decides slowly and changes its mind quickly.

The basic design of Figure 8.86 illustrated the need for an absolute value function for data channels and had a provision to bypass this function for pilot channels. This design can be used for a data channel or a pilot channel since it always takes the absolute values of the inputs before passing these to their respective lowpass filters. However, if the receiver channel knows that it will always be working with pilot channels, the absolute value function should be eliminated.

Referring to Figure 8.87 the divide by zero problem is avoided by dividing the lowpass filtered output of I_P with an optimized scale factor, K_2. This result is compared to the filtered quadraphase result, Q_P. The decision is made that phase lock has been achieved if the scaled mean value of the absolute amplitude of I_P divided



Note: The pilot PLL I_P and Q_P signals do not require the absolute value functions.

Figure 8.87 Advanced phase lock detector design providing optimistic and pessimistic indicators.

by a constant K_2 is greater than the mean value of the absolute amplitude of Q_P. The lowpass filters provide the mean value. It is very important to initialize the accumulators to zero prior to first use any time it was previously disengaged. The selection of K_2 is based on optimizing the decision threshold so as to find a balance between the probability that the detector reports that the carrier tracking loop is in phase lock when it is actually out of phase lock (Type-1 error) and the probability that the detector reports that the tracking loop is not in phase lock when it is actually in phase lock (Type-II error). The pessimistic feature in this design permits making this threshold tradeoff slightly favoring the Type-I error to reduce the Type-II error. The key feature of this design is that the pessimistic feature provides a means for a simple likelihood ratio test.

Observe in Figure 8.87 that the optimistic phase lock indicator is set TRUE and Pcount1 is incremented by 1 after the first positive single trial phase lock decision. The pessimistic phase lock indicator remains FALSE until Pcount1 is greater than L_p (i.e., the pessimistic phase lock indicator count based on the number of optimistic decisions in a row). After the pessimistic phase lock indicator has been set TRUE, the first single trial phase lock decision that is negative sets the pessimistic phase lock indicator FALSE and Pcount1 is set to zero. However, after the optimistic phase lock indicator is set TRUE, its Pcount2 is also set to zero by every single trial phase lock detector positive decision. When the single trial phase lock detector makes a negative decision, the optimistic phase lock indicator remains TRUE and Pcount2 begins to increment for every negative single trial phase lock detector decision in a row until it exceeds the optimistic hold on threshold count, L_o. When either the L_p or the L_o counters have exceeded their threshold values, their count is suspended until they are reset in order to avoid rollover problems.

The data demodulation function requires its own data channel phase lock detector shown in Figure 8.88. This uses a wider-band lowpass filter and works in

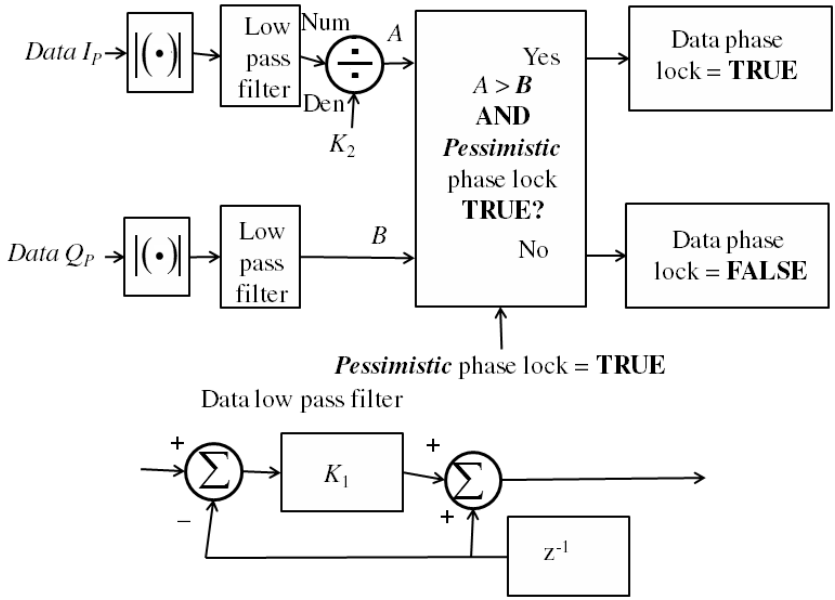


Figure 8.88 Data channel phase lock detector using pilot pessimistic indicator.

combination with the pessimistic phase lock indicator from the phase lock detector shown in Figure 8.87.

Data demodulation proceeds if the data channel phase lock indicator is TRUE and the pessimistic phase lock indicator is TRUE. This provides the desired higher probability of phase lock for data demodulation for operation with legacy data channels. It is optimum for modernized pilot and data channels since the data channel phase lock detector is monitoring the proper functioning of the data channel phase lock condition of the data channel while also communicating with the advanced phase lock detector in the pilot channel. Note that there is no data channel phase lock loop because this is closed by the pilot channel PLL, but the second detector verifies the expected phase lock condition after data channel code wipe-off is performed. Also, note that in Figure 8.19 the $\tilde{Q}_n$ leg of the modernized data channel would have to be implemented the same as the in-phase signal so that there will be a quadraphase component signal for the data phase lock detector. Typical values for the design parameters of both phase lock detectors are shown in Table 8.27.

8.13.2.2 False Frequency Lock Detector

False frequency lock can occur in FLL. This can be detected when the DLL velocity state does not match the FLL velocity state. Only a comparison check is necessary in FLL to correct the FLL velocity state. The DLL and FLL velocity states can be compared at their respective loop filter outputs as shown in Figure 8.18 when the carrier loop is operating in FLL by using the appropriate carrier loop scale factor from Table 8.13.

Table 8.27 Typical Design Parameters for Advanced Phase Lock Detector

Channel Type	*Predetection Integration Time (ms)*	*Lowpass Filter K_1*	*Threshold Denominator K_2*	L_p	L_o
P(Y) code in data wipe-off (proxy for pilot channel)	20	0.0198	0.36	5	240
C/A or P(Y)	20	0.0247	1.5	50	240
Second detector[1]	20	0.0952[1]	1.5	N/A[1]	N/A[1]

Note 1: The preferred data demodulation phase lock detector is a second detector like the primary detector design, except it has no optimistic/pessimistic logic of its own. It uses a wider bandwidth lowpass filter (K_1) plus the phase lock decision is based on $A > B$ and pessimistic phase lock = TRUE.

8.13.2.3 False Phase Lock Detector

False phase lock can occur after loop closure in PLL. This can be observed when the phase lock indicator declares phase lock but the PLL replica frequency state is incorrect. The incorrect frequency is typically some multiple of 25 Hz owing to typical sampling rates to match the data demodulation process. The FLL-assisted-PLL loop design ordinarily prevents false phase lock if the FLL is allowed enough time to pull in the frequency before transition into PLL and the transition is not into a narrowband PLL with a very small pull in range, but it is prudent to implement a false phase lock indicator to detect this possible false carrier loop condition. The false phase lock indicator is used only when the phase lock indicator declares that a phase lock condition exists.

Figure 8.89 is a design example of a false phase lock indicator. It performs a frequency discriminator function on a pair of prompt in-phase and quadraphase samples, I_{Pi-1}, Q_{Pi-1}, I_{Pi}, and Q_{Pi} formed into the cross product, *Cross*, and dot product, *Dot*, functions shown as inputs into the detector. Legacy GPS signals with 50-bps data modulation typically collect in-phase and quadraphase samples every 10 ms (2 samples per data bit) and the *Cross* and *Dot* products are formed every T = 20 ms then applied to the input. The pilot channel of a modernized signal is not vulnerable to data transitions but can provide these samples the same way since its data channel will typically demodulate 10-ms samples. Since both detectors are

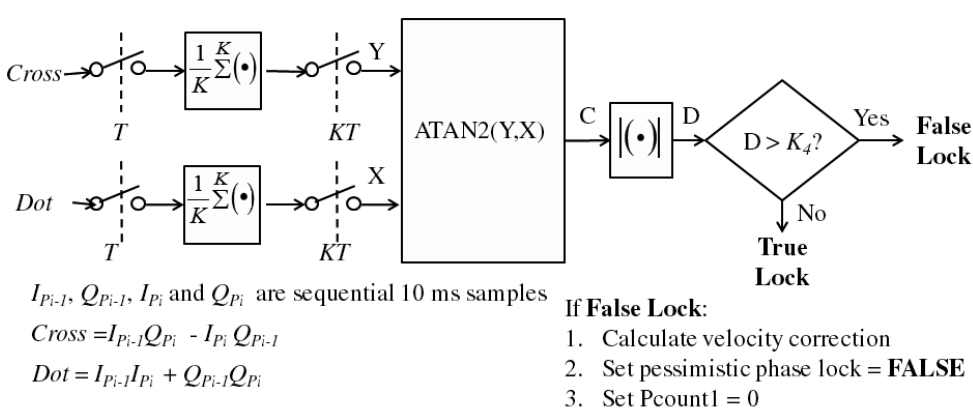


Figure 8.89 False phase lock detector.

associated with the same carrier accumulator, the false phase lock detector must be synchronized with the phase lock detector so that both are detecting the same data.

As shown in Figure 8.89, the inputs are integrated and dumped for K samples. At KT second intervals, the four-quadrant arctangent is computed with output C. The absolute value of $C = D$ represents the change in phase in KT seconds in units of hertz, and this is compared to the threshold parameter, K_4. If D exceeds K_4, then the detector declares false lock. As shown in the figure, the three actions taken when false phase lock is determined are: (1) calculate the velocity correction; (2) set the pessimistic phase lock to FALSE; and (3) reset the pessimistic counter, Pcount1, to zero. The velocity correction is applied to the carrier accumulator and the calculation is: Sign $C(2\pi K_5)$, where K_5 is typically 25 and C is the output of the four-quadrant arctangent function. No action is required if the detector declares true lock. Typical false phase lock parameters are shown in Table 8.28.

8.13.2.4 Code Lock Detector

One of the most difficult detectors to design is the code lock detector. The code-tracking loop is far more robust than the carrier-tracking loop, but that is academic if the carrier-tracking loop loses track in an unaided GNSS receiver. The code loop soon loses lock for lack of accurate carrier wipe-off and that cannot be provided with sufficient accuracy by the code-tracking loop even under moderate dynamic stress. The C/N_0 meter is normally the most sensitive and reliable detector that provides loss of code track information for an unaided GNSS receiver. However, stationary operation (if known and utilized as aiding) or inertial aiding of the carrier tracking loop can sustain carrier wipe-off under weak signal hold-on conditions for long periods of time if the velocity aiding is properly implemented. In this case, a code lock detector is essential because the code-tracking loop can be sustained down to the region of $(C/N_0)_{dB} = 5$ dB-Hz with high quality of velocity aiding and an open carrier tracking loop. Figure 8.90 is representative of a code lock detector design intended for such receiver applications.

The code lock detector design parameters are shown in Table 8.29 where it is noted that the design parameters are for legacy C/A and P(Y) code, but one version utilized data wipe-off so these parameters can be a proxy for a modernized pilot channel with the same spreading code rate such as L5. Referring to Figure 8.90, the input signals are the prompt (I_P and Q_P) signals and the envelope of the noise meter (I_N and Q_N) signals, both at 5-ms sample periods. The code-tracking loop uses the error in the early minus late signals to maintain lock but it also keeps the prompt signal close to being centered if it is in code lock. This is normally the strongest

Table 8.28 False Phase Lock Detector Parameters

Channel Type	I_{Pi-1}, Q_{Pi-1}, I_{Pi}, Q_{Pi} *Samples (ms)*	K	K_4
P(Y) in data wipe-off (proxy for pilot channel)	10	N/A	15.5
C/A or P(Y)	10	50	15.5

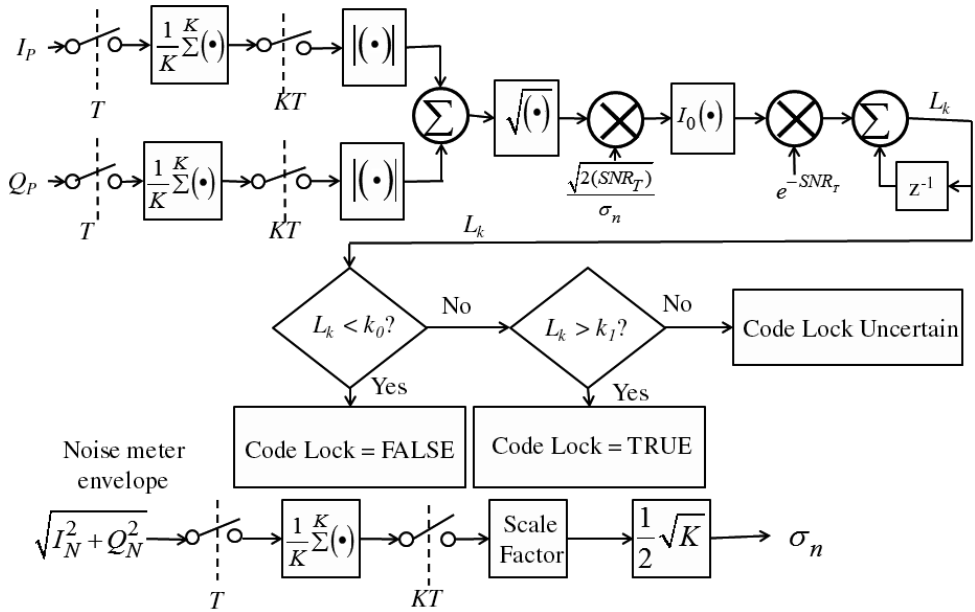


Figure 8.90 Code lock detector.

Table 8.29 Code Lock Detector Parameters

Code Type	*T (ms)*	*K*	SNR_T	k_0	k_1
C/A[1]	5	1[1]	1.990	0.01	99.0
C/A	10	2	7.962	0.01	99.0
P(Y)	10	2	0.200	0.01	99.0
P(Y)[2]	10	32[2]	0.507	0.01	99.0

Note 1: When $K = 1$ the code lock detector uses only one of every four 5-ms input samples (data bit edge unknown).
Note 2: When $K = 32$ the data wipe-off feature is enabled on P(Y) code and is a proxy for a pilot channel design.

signal of the three. After some normalized integration defined by K, the absolute values are added and the square root is taken. That result is multiplied by the ratio of SNR_T over the noise meter standard deviation. SNR_T is precomputed as the signal-to-noise ratio threshold parameter. The process shown in the lower part of the figure determines the noise meter standard deviation. That result is used as the argument for the Bessel function of the first kind computation, then multiplied by $\exp(SNR_T)$. The final step is a boxcar integration that produces the estimated code lock level for the kth iteration, L_k. The code lock is FALSE if L_k is smaller than k_0 and TRUE if greater than k_1. It is uncertain when it is between these two limits. The appropriate values of K, SNR_T, k_0, and k_1 are provided in Table 8.29 and the scale factor used in noise meter standard deviation estimation is from Table 8.25. It is very important to zero the boxcar integrator memory on first use of this detector any time it has been previously disabled.

8.13.3 Cycle Slip Editing

Cycle slip editing is essential for GNSS receiver applications involving precision interferometry measurements. For examples, precision surveying and real-time kinematic (RTK) GNSS receivers utilize the integrated carrier Doppler phase measurements described in Section 8.10.3 to obtain centimeter (and with the modernized civil signals even millimeter) level differential precision. The common objective of these applications is to resolve the integer ambiguity in these measurements by a variety of techniques. The common problem is the correction of carrier Doppler phase cycle slips that alter this integer by half-cycle increments in data channel (Costas) PLL tracking and full-cycle slips in pilot channel PLL tracking. The cycle slip problem for civil signals has been compounded by limitations of the legacy L1 C/A code that was designed to be a stepping-stone into P(Y) code and even greater limitations of semicodeless L2 P code receivers (described in Section 8.7.4). The C/A code slips in half-cycle integers and the most effective semicodeless L2 P code receivers slip in full-cycle integers. As a result of these commercial signal shortcomings there have been an overwhelming number of papers published under different titles that have the common pursuit of detecting and correcting carrier cycle slips for both single-frequency C/A code and dual-frequency L1 C/A and L2 semicodeless receivers. These cycle slip editing techniques are almost universally based on sophisticated detection and correction algorithms that are performed by the real-time navigation process or the non-real-time post-mission process. These are not addressed herein. However, a cycle slip detection and correction method that can be performed by the receiver control process within a modernized dual-frequency (or triple-frequency) GNSS receiver will be described. Modernized GNSS signals with robust multifrequency pilot channels will greatly diminish the seriousness of the cycle slip problem. However, cycle slips will still happen under certain circumstances of excessive dynamic stress or natural interference (ionospheric noise). Excessive dynamic stress will tend to make different frequency PLLs for a given SV slip in the same direction while natural interference (random noise) may cause them to slip in opposite directions.

The basic phase lock detector shown in Figure 8.86 could be used to provide an alert that a cycle slip *may* have occurred, but a cycle slip *cannot be detected with assurance* in a single-frequency GNSS receiver because it is a statistical occurrence during dynamic stress or in the presence of natural interference (that lowers the C/N_0). However, cycle slip detection and correction (editing) can be performed by the receiver control function of a dual- (or triple-) frequency GNSS receiver as follows.

The receiver-based cycle slip editing technique described herein depends on short-term constancy of the ionospheric delay over the time period between integrated carrier Doppler phase measurements from two receiver channels tracking the same SV at two different frequencies, typically in 1-second intervals. The design example presented is based on using the L5 Q pilot channel and the L1 C/A (Costas) channel. The observable is the double difference between a *new* single difference between L5 Q and L1 C/A integrated carrier Doppler phase measurements and the preceding *old* single difference. The *new* single difference for SV_i is:

$$\Delta\lambda_{new} = \Delta ID_{iL5-iL1C/A}(new) = ID_{iL5}(new) - ID_{iL1C/A}(new) \quad (\text{m}) \tag{8.117}$$

where

ID_{iL5} (new) = Current SV_i L5 Q pilot channel integrated carrier Doppler phase measurement

$ID_{iL1C/A}(new)$ = Current SV_i L1 C/A Costas integrated carrier Doppler phase measurement

The *old* single difference is

$$\Delta\lambda_{old} = \Delta ID_{iL5-iL1C/A}(old) = ID_{iL5}(old) - ID_{iL1C/A}(old) \quad (\text{m}) \tag{8.118}$$

where

$ID_{L5}(old)$ = Initial or previous SV_i L5 Q pilot channel measurement

$ID_{L1CA}(old)$ = Initial or previous SV_i L1 C/A Costas measurement

The assumption is that there are no cycle slips in the old L5 Q PLL or L1 C/A PLL single-difference measurements because the initial single-difference measurement was made during highly favorable PLL conditions and thereafter any cycle slips detected were corrected during the previous editing operation. The editor uses the double-difference observable:

$$X = \Delta ID_{iL5-iL1C/A}(new) - \Delta ID_{iL5-iL1C/A}(old) = \Delta\lambda_{new} - \Delta\lambda_{old} \quad (\text{m}) \tag{8.119}$$

Note that

$$X = \left[ID_{iL5}(new) - ID_{iL5}(old)\right] - \left[ID_{iL1C/A}(new) - ID_{iL1C/A}(old)\right] + (\Delta_{L5Iono} - \Delta_{L1Iono}) \quad (\text{m})$$

where Δ_{L5Iono} is the change in the L5 ionospheric delay between new and old L5 Q measurements and Δ_{L1Iono} is the change in the L1 ionospheric delay between new and old L1 C/A measurements, with a typical time interval of 1 s between new and old measurements. Since the ionospheric delay change in this short time interval is typically a small order value then the double difference is approximately zero. Also note that X is the delta pseudorange on L5 (plus its delta ionospheric delay) minus the delta pseudorange on L1 C/A (plus its delta ionospheric delay). Therefore, if there are small-order noise errors in the double differences due to the measurement noise and ionospheric delay difference noise, then X will be approximately zero if there are no cycle slips in either PLL (i.e., the delta pseudoranges are effectively canceled by the double-difference measurement because they would be equal if there were no noise present). If there is a cycle slip in $ID_{iL5}(new)$, then the first difference will be approximately the delta pseudorange plus or minus one L5 wavelength, depending on which way it slipped. Likewise, if a cycle slip occurs in $ID_{iL1C/A}(new)$, then the first difference will be approximately the delta pseudorange plus or minus one L1 half-wavelength, depending on which way it slipped. In general, the double difference yields the algebraic value of the difference between the number of L5 wavelengths slipped minus the number of L1 half-wavelengths slipped, so cycle slip editing can be performed based on the value of the *jump* in this observable as

well as the sign of the *jump*. Multiple cycle slips can be observed and the sources corrected so long as the net *jump* in this observable is much larger than the noise associated with each double difference measurement. Table 8.30 depicts the L5Q and L1 C/A parameters computed for use in the cycle slip editor design. As observed in the table, when various combinations of L5 PLL cycle slips and L1 C/A Costas PLL half-cycle slips occur, there are some surprising results in the double-difference observables. The wavelength of the L5 carrier is 0.254828049 m (Error Rank 9 in the table) and the half wavelength of the L1 carrier frequency is 0.095146836 m (Error Rank 3 in the table). Even though the L1 C/A signal is much more likely to slip than L2 CL, the former signal was chosen to compare with the very robust L5 Q signal

Table 8.30 Cycle Slip Editor Ranked Error Values for L5 and L1 C/A Cycle Slips

Error Rank	*Slip Error Combinations*	*Error Values (m) Z = ABS(X)*	*Locations 12* Edits*	*2L5,3L1C/A 19 Edits*	*1L5,3L1C/A 12* Edits*	*1L5, 2L1C/A 8 Edits*
0	0L5+0L1	0	E0	Edit	Edit 0	Edit
1	–1L5+3L1	0.03061246	E1	Edit	Edit 1	LPEC
2	1L5-2L1	0.064534376	E2	Edit	Edit 2	Edit
3	0L5+1L1	0.095146836	E3	Edit	Edit 3	Edit
4	–1L5+4L1	0.125759297		LPEC	LPEC	LPEC
5	2L5-4L1	0.129068752		LPEC	LPEC	LPEC
6	1L5-1L1	0.159681212	E4	Edit	Edit 4	Edit
7	0L5+2L1	0.190293673	E5	Edit	Edit 5	Edit
8	2L5-3L1	0.224215588		Edit	LPEC	LPEC
9	1L5+0L1	0.254828049	E6	Edit	Edit 6	Edit
10	0L5+3L1	0.285440509	E7	Edit	Edit 7	LPEC
11	2L5-2L1	0.319362425		Edit	LPEC	LPEC
12	1L5+1L1	0.349974885	E8	Edit	Edit 8	Edit
13	0L5+4L1	0.380587346		LPEC	LPEC	LPEC
14	3L5-4L1	0.383896801		LPEC	LPEC	LPEC
15	2L5-1L1	0.414509261		Edit	LPEC	LPEC
16	1L5+2L1	0.445121722	E9	Edit	Edit 9	Edit
17	3L5-3L1	0.479043637		LPEC	LPEC	FAIL
18	2L5+0L1	0.509656098		Edit	LPEC	
19	1L5+3L1	0.540268558	E10	Edit	Edit 10	
20	3L5-2L1	0.574190474		LPEC	LPEC	
21	2L5+1L1	0.604802934	E11	Edit	FAIL	
22	1L5+4L1	0.635415394		LPEC		
23	4L5-4L1	0.63872485		LPEC		
24	3L5-1L1	0.66933731		LPEC		
25	2L5+2L1	0.69994977		Edit		
26	1L5+5L1	0.730562231		LPEC		
27	4L5-3L1	0.733871686		LPEC		
28	3L5+0L1	0.764484146		LPEC		
29	2L5+3L1	0.795096607		Edit		
30	4L5-2L1	0.829018522		FAIL		

LPEC = low probability error condition.

because there is a greater absolute difference of 0.159681212 m (Error Rank 6 in the table) than would be the absolute difference (0.010617835 m) between the L5 carrier wavelength and the L2 CL carrier wavelength of 0.244210213 m.

Table 8.30 was created after numerous slip error combinations of L5 and L1 C/A cycle slips were computed, recognizing that cycle slips can occur in the same direction or in the opposite direction in each carrier tracking loop. Then the Error Values $Z = ABS(X)$ (assuming no noise on the double-difference measurements) were sorted in Error Rank order as shown in the first three columns of Table 8.30. The combinations in the second column are signed so that a positive value is produced by their sum in the third column, but in reality each combination can have a positive or negative outcome, so the absolute value of the double difference, $Z = ABS(X)$, is used in the actual design and the sign of the double difference, $Y = SIGN(X)$, is retained to determine the proper direction for each slip correction. These are shown in rank order from Rank 0 to Rank 30 that includes one more slip rank combination above the maximum double-difference absolute value of 2 L5 cycle slips plus 3 L1 C/A half-cycle slips.

The last three columns of Table 8.30 correspond to the editing limits of three different cycle slip editor designs: 2 L5 and 3 L1 C/A slips requiring 19 Edits, 1 L5 and 3 L1 C/A slips requiring 12 edits, and 1 L5 and 2 L1 C/A slips requiring 8 edits, respectively. Each edit count includes the typical edit of zero errors and the FAIL conditions. The 12 edits design that detects and corrects up to 1 L5 and 3 L1 C/A cycle slips was chosen (denoted with an asterisk in Table 8.30), so 11 reference designators (E0 through E11) are used as labels corresponding to the absolute value of the errors that are used to calculate the thresholds. Some table entries are labeled as low probability error combination (LPEC), because these values contain slips that are not checked in the test range of the cycle slip editor design, but their values fall within the test range. For higher probability error condition examples, two combinations that are very close to the chosen slip detection range, such as Rank 4 (–1L5+4L1) and Rank 5 (2L5-4L1), that fall within the test range are included in the table but marked LPEC. For lower probability error condition examples, two combinations that are far from the chosen slip detection range, such as Rank 14 (3L5-4L1) and Rank 17 (3L5-3L1), also fall within the test range and are also marked LPEC. Any LPEC condition that actually occurs will not be detected, but the subsequent incorrect editing will compound the error and typically be detected as a failure during the next cycle. The cycle slip editor must be reinitialized under more stable PLL conditions when a failure occurs, so a failure flag should also be a warning that the previous edit could have been incorrect.

Table 8.31 completes the design of the 12 edit cycle slip editor and Figure 8.91 illustrates the logic of the design based on this table.

The cycle slip editing is performed by receiver control (RC) on each measurement before the corrected measurement is sent to the navigation process and the integer slip corrections are sent to their respective L5 and L1 C/A receiver channels for the same SV being tracked in PLL. After the L1 C/A and L5 Q integrated carrier Doppler phase signals have been corrected, the corrected L1 C/A signal can then be used to edit L2 CL cycle slips if that signal is also being tracked on the same SV by another receiver channel. But in the second case it can be assumed that the L1 C/A integrated Doppler measurement has been corrected. The number of combinations to be detected are fewer since there will only be L2 C cycle slips to detect and

Table 8.31 Cycle Slip Editor Design to Detect and Correct 1 L5 and 3 L1 C/A Cycle Slips

Error Rank	*Actions*	*Error Symbols and Threshold Locations*	*Values (m)*	*Slips L5, L1*	*Y = +*		*Y = −*		*Comments*
					CL5	*CL1*	*CL5*	*CL1*	
0	Edit 0	E0	0	0, 0	0	0	0	0	0 slips
	T0	T(0)=E0+(E1-E0)/2	0.01530623						
1	Edit 1	E1	0.03061246	–1, 3	1	3	–1	–3	1, 3 slips same direction
	T1	T(1)=E1+(E2-E1)/2	0.047573418						
2	Edit 2	E2	0.064534376	1, –2	–1	–2	1	2	1, 2 slips same direction
	T2	T(2)=E2+(E3-E2)/2	0.079840606						
3	Edit 3	E3	0.095146836	0, 1	0	1	0	–1	1 L1 slip
4	LPEC		0.125759297	–1, 4					–1, 4 slips not detected
	T3	T(3)=E3+(E4-E3)/2	0.127414024						
5	LPEC		0.129068752	2, –4					2, –4 slips not detected
6	Edit 4	E4	0.159681212	1, –1	–1	–1	1	1	1, 1 slips same direction
	T4	T(4)=E4+(E5-E4)/2	0.174987443						
7	Edit 5	E5	0.190293673	0, 2	0	2	0	-2	2 L1 slips
	T5	T(5)=E5+(E6-E5)/2	0.222560861						
8	LPEC		0.224215588	2, –3					2, –3 slips not detected
9	Edit 6	E6	0.254828049	1, 0	–1	0	1	0	1 L5 slip
	T6	T(6)=E6+(E7-E6)/2	0.270134279						
10	Edit 7	E7	0.285440509	0, 3	0	3	0	–3	3 L1 slips
	T7	T(7)=E7+(E8-E7)/2	0.317707697						
11	LPEC		0.319362425	2, –2					2, –2 slips not detected
12	Edit 8	E8	0.349974885	1, 1	–1	1	1	–1	1, 1 slips opposite
13	LPEC		0.380587346	0, 4					4 L1 slips not detected
14	LPEC		0.383896801	3, –4					3, –4 slips not detected
	T8	T(8)=E8+(E9-E8)/2	0.397548303						
15	LPEC		0.414509261	2, –1					2, –1 slips not detected
16	Edit 9	E9	0.445121722	1, 2	–1	2	1	–2	1, 2 slips opposite
17	LPEC		0.479043637	3, –3					3, –3 slips not detected
	T9	T(9)=E9+(E10-E9)/2	0.49269514						
18	LPEC		0.509656098	2, 0					2 L5 slips not detected
19	Edit 10	E10	0.540268558	1, 3	–1	3	1	–3	1, 3 slips opposite
	T10	T(10)=E10+(E11-E10)/2	0.557229516						
20	FAIL	E11	0.574190474	3, –2					Set FAIL = 1

LPEC = low probability error condition.

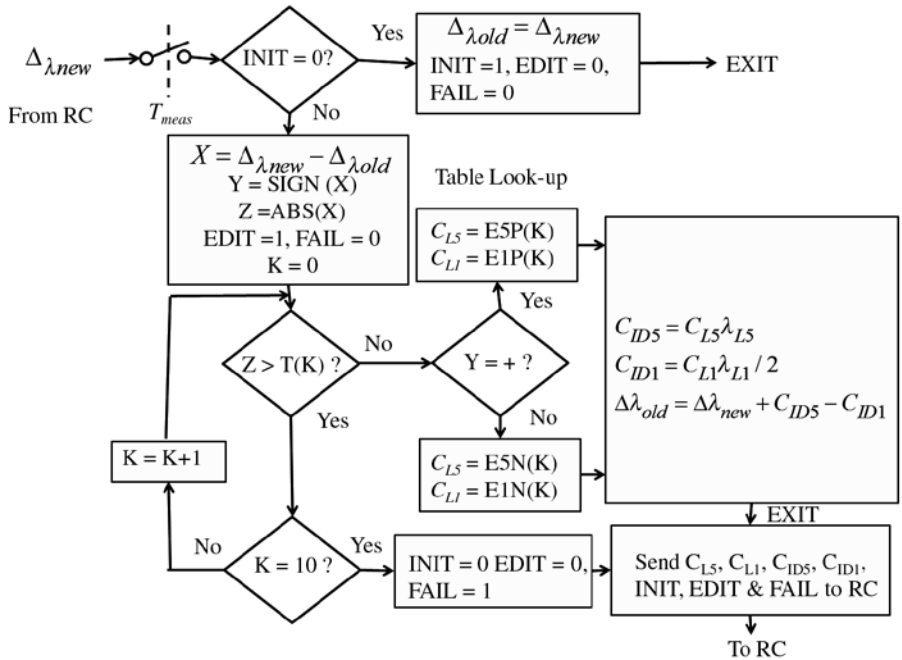


Figure 8.91 Simple receiver-based L5 and L1 C/A cycle slip editor.

correct. When L1 C_p becomes available, it can be used by this technique as a more reliable and much more accurate PLL mate with L5 Q5 for cycle slip editing (and then with L2 CL), but the fact that it slips in full cycles means that there will be smaller double-difference values than with L1 C/A half-cycle slips.

Referring to Figure 8.91, RC receives the integrated Doppler measurements from the L5 Q and L1 C/A PLL measurements in their respective receiver channels every measurement interval, T_{meas}, then performs the first difference, $\Delta_{\lambda new}$, in accordance with (8.117), and sends this to the cycle slip editor. If this is the first time the editor is used then RC must ensure that PLL conditions are reliable in both receiver channels, such as verifying that both channels are operating in *pessimistic PLL mode* (described in Section 8.13.2.1). The editor uses the first input to initialize $\Delta_{\lambda old}$, then sets INIT =1, EDIT =0 and FAIL = 0 before exiting to indicate to RC that it has been initialized, not in failure condition, and is providing no edited cycle slips during this feedback operation.

Subsequent inputs are used to perform the double difference with the result in *X*, the sign in *Y*, and the absolute value in *Z*, followed by a test loop iterated by the index K from K = 0 to 10, that compares *Z* to the (table lookup) threshold T(K). When any comparison does not exceed T(K) then the index K plus the sign in *Y* are used to obtain the integer cycle slip corrections, C_{L5} and C_{L1}, from (table lookup) E5P(K) and E1P(K) if the sign is positive, or from E5N(K) and E1N(K) if the sign is negative, respectively. The tables contain the values shown in Table 8.31 for each edit condition, E(K) = E0 to E10, under the two column headings, *Y* = + or *Y* = –, with each column heading containing two columns for the C_{L5} and C_{L1} corrections, respectively. The editor computes the integrated Doppler corrections, $C_{ID5} = C_{L5}\lambda_{L5}$ and $C_{ID1} = C_{L1}\lambda_{L1}/2$. These are used to provide corrections to the NAV measurements and to replace the *old* single difference as shown in the figure:

$\Delta_{\lambda old} = \Delta_{\lambda new} + C_{ID5} - C_{ID1}$. The editor exits successfully when INIT = 1, EDIT = 1, and FAIL = 0. Under this condition, RC adds the integrated Doppler corrections, C_{ID5} and C_{ID1}, to their corresponding measurements sent to NAV and sends the integer cycle slip corrections, C_{L5} and C_{L1}, back to their respective L5 and L1 C/A channels to retain and add to the integer parts of future carrier Doppler phase measurements. The editor fails when INIT = 0, EDIT = 0 and FAIL = 1 because the threshold was exceeded when K = 10; that is, the highest threshold value tested for this case example. If FAIL = 1, then RC must thereafter wait until both receiver channels are operating in stable PLL condition, for example when both receiver channels are operating in *pessimistic PLL mode*.

This is a simple but powerful example of a carrier tracking loop cycle slip editing because it is performed in real time on every carrier observable before each measurement is sent to NAV. It becomes very complex when multiple combinations are considered, so it is a judgment call as to how many cycle slip combinations to attempt to detect every T_{meas}. For example, it should be recognized that all PLLs are on the verge of losing lock if there is more than one cycle slip in one second and the T_{meas} sample time shown in the figure is rarely longer than one second. This technique also becomes increasingly unreliable as combinations are increased because threshold margins are reduced. For this case example, the first and worst-case threshold margin from Table 8.31 is T(0) – 0 = 0.01531 m and the next two worst-case margins are T(1) – T(0) = T(2) – T(1) = 0.03227 m. The remaining margins are 0.04757 m and higher. If the 8 edit example of 1 L5 and 2 L1C/A cycle slips from Table 8.30 had been chosen, then the first and worst-case threshold margin would have been 0.03227 m, a significant reliability improvement.

To provide additional insight into the reliability of detecting cycle slips by this method, Figure 8.92 depicts the third-order PLL errors (in units of m 1-sigma) as a function of $(C/N_0)_{dB}$ for L1 C/A, L2 CL and L5 Q5 carrier tracking loops. This figure also includes the worst-case cycle slip editor threshold margin. The L1 C/A

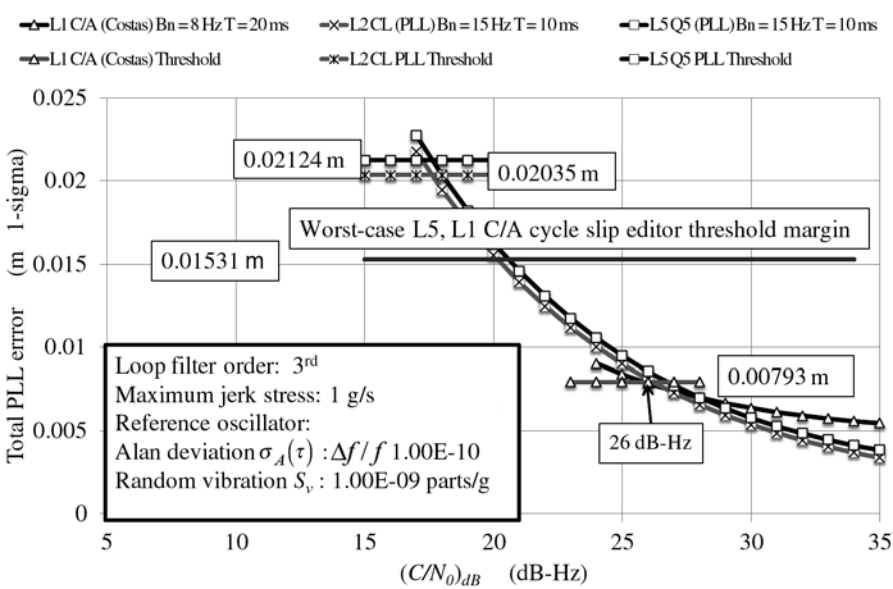


Figure 8.92 Total PLL error plots for L1 C/A, L2 CL, and L5 Q5 .

Table 8.32 Assumed External PLL Error Values Used in Figure 8.92

	L1 C/A PLL 1-Sigma Error (m)	*L2 CL PLL 1-Sigma Error (m)*	*L5 Q5 PLL 1-Sigma Error (m)*
Random vibration error for $S_v = 1.00E\text{-}09$ *(parts/g)*	0.00075	0.00075	0.00075
Alan deviation error for $\sigma_A(\tau) = 1.00E - 10(\Delta f/f)$	(B_n = 8 Hz) 0.00167	(B_n = 15 Hz) 0.00089	(B_n = 15 Hz) 0.00089
Maximum dynamic stress error for 1 g/s jerk	(B_n = 8 Hz) 0.00308	(B_n = 15 Hz) 0.00047	(B_n = 15 Hz) 0.00047

(Costas) predetection integration time is optimized at $T = 0.020$ s to accommodate the presence of 50 Hz message data. The third-order PLLs require $B_nT = 0.146$ for 30 deg phase margin assuming T computation delay (from Table 8.24). This would require $B_n = 7.3$ Hz for L1 C/A, but this has been increased to 8 Hz that provides adequate margin. The two pilot channels are optimized for $T = 0.010$ s so that their slaved data channels have their transition boundaries aligned with their 100-Hz symbol rates. This results in $B_n = 14.6$ Hz for 30-deg phase margin, but this has been increased to 15 Hz that provides adequate margin. Table 8.32 provides the values (in units of m 1-sigma) assumed for the remaining PLL error contributions. The random vibration specification is the same as used in the case example for this topic in Section 8.93.

Inspection of Figure 8.92 shows that the L1 C/A Costas loop is clearly the most likely to slip of the three PLLs. The L1 C/A loop loses lock if the jerk dynamic stress increases to 2 g/s even for $(C/N_0)_{dB} = 35$ dB-Hz. The other two carrier tracking loops remain in phase lock when the dynamic stress increases beyond 15 g/s when $(C/N_0)_{dB} \geq 25$ dB-Hz because their noise bandwidths and tracking thresholds are much larger. However, for the L1 C/A carrier tracking loop operating under a maximum jerk dynamic stress of 1 g/s or less and with $(C/N_0)_{dB} \geq 30$ dB-Hz, this cycle clip editor should perform reliably. This means that the dynamic stress toward the SV being tracked must remain at or below 1 g/s for all three carrier tracking loops. The natural interference (ionospheric noise) on the L1 signal must not lower its $(C/N_0)_{dB}$ to below about 30 dB-Hz to ensure reliable PLL tracking and the L5 signal must not be reduced to below about 25 dB-Hz to ensure that the PLL tracking errors remain well below the cycle slip editor threshold margin. Although this design has not been tested under simulated or actual field conditions, the theoretical design indicates that up to 1 L5 cycle slip and 3 L1 C/A cycle slips of all combinations can be detected and corrected reliably typically every second in real time using this technique under the assumed operating conditions.

References

[1] Hegarty, C. J., et al., "An Overview of the Effects of Out-of-Band Interference on GNSS Receivers," *Proc. of the 24th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2011)*, Portland, OR, September 2011, pp. 1941–1956.

[2] Milligan, T. A., *Modern Antenna Design*, 2nd ed., New York: Wiley-IEEE Press, 2005.

[3] Elliot, P. G., E. N. Rosario, and R. J. Davis, "Novel Quadrifilar Helix Antenna Combining GNSS, Iridium, and a UHF Communications Monopole," *Proc. of the IEEE Military Communications (MILCOM) Conference*, Orlando, FL, November 2012.

[4] Taga, T., *Analysis of Planar Inverted-F Antennas and Antenna Design for Portable Radio Equipment*, Norwood, MA: Artech House, 1992.

[5] Dilssner, F., et al., "Impact of Near-Field Effects on the GNSS Position Solution," *Proc. of the 21st International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2008)*, Savannah, GA, September 2008, pp. 612–624.

[6] Bucher, J., (ed.), *The Metrology Handbook*, Milwaukee, WI: ASQ Quality Press, 2004.

[7] Antenna Temperature, http://www.antenna-theory.com/basics/temperature.php

[8] Spirent Application Note, "Simulation of Realistic Antenna Noise," Spirent Communications, http://www.spirent.cn/~/media/Application%20notes/Positioning/Simulation_of_Realistic_Antenna_Noise_AppNote.pdf.

[9] Van Diggelen, F., *A-GPS: Assisted GPS, GNSS, and SBAS*, Norwood, MA: Artech House, 2009.

[10] Ward, P. W., "RFI Situational Awareness in GNSS Receivers: Design Techniques and Advantages," *Proc. of the 63rd Annual Meeting of The Institute of Navigation,* Cambridge, MA, April 2007, pp. 189–197.

[11] Ward, P. W., "What's Going On? RFI Situational Awareness in GNSS Receivers," *Inside GNSS*, September-October 2007, pp. 34–42.

[12] Ward, P. W., "Simple Techniques for RFI Situational Awareness and Characterization in GNSS Receivers," *Proc. of the 2008 National Technical Meeting of The Institute of Navigation*, San Diego, CA, January 2008, pp. 154–163.

[13] Ward, P. W., "Interference Heads-Up – Receiver Techniques for Detecting and Characterizing RFI," *GPS World*, June 2008, pp. 64–73.

[14] Analog Devices AD9265, 16-bit, 125 MSPS/105 MSPS/80 MSPS, 1.8 V Analog-to-Digital Converter, http://www.analog.com/media/en/technical-documentation/data-sheets/AD9265.pdf

[15] Ould, P. C., and R. J. Van Wechel, "All-Digital GPS Receiver Mechanization," *NAVIGATION: Journal of The Institute of Navigation,* Vol. 28, No. 3, Fall 1981, pp. 178–188.

[16] Spilker, J. J., and F. D. Natali, "Interference Effects and Mitigation Techniques." *Global Positioning System: Theory and Applications*, Vol. 1, AIAA Volume 163, 1996, p. 726.

[17] Hegarty, C. J., "Analytical Model for GNSS Receiver Implementation Losses," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 58, No. 1, Spring 2011, pp. 29–44.

[18] Analog Devices, Inc., *Data Conversion Handbook*, December 16, 2004, http://www.analog.com/library/analogDialogue/archives/39-06/data_conversion_handbook.html.

[19] Kester, W., "ADC Architectures I: The Flash Converter," Analog Devices MT-020, http://www.analog.com/media/en/training-seminars/tutorials/MT-020.pdf.

[20] Scott, L., "The New Military and Civil Signals: Structure, Processing & Tracking," Navtech Seminars Course 413, 2004.

[21] Susskind, A. K., (ed.), *Notes on Analog-Digital Conversion Techniques*, The M.I.T. Press, Cambridge, MA: MIT Press, 1963, Ch. 2, p. 6.

[22] Kester, W., "What the Nyquist Criterion Means to Your Sampled Data System Design," Analog Devices MT-002, http://www.analog.com/media/en/training-seminars/tutorials/MT-002.pdf.

[23] Kester, W., "The Good, the Bad, and the Ugly Aspects of ADC Input Noise—Is No Noise Good Noise?" Analog Device MT-004, http://www.analog.com/media/en/training-seminars/tutorials/MT-004.pdf.

[24] Kester, W., "ADC Noise Figure—An Often Misunderstood and Misinterpreted Specification," Analog Devices MT-006 Tutorial, http://www.analog.com/media/en/training-seminars/tutorials/MT-006.pdf.

[25] Smith, S. W., "The Scientist and Engineer's Guide to Digital Signal Processing," Ch. 14 in *Introduction to Digital Filters, High-Pass, Low-Pass and Band Reject Filters*, http://www.dspguide.com/ch14.htm.

[26] Kester, W., "Mixed Signal and DSP Design Techniques," Section 6: Digital Filters, Finite Impulse Response (FIR) Filters, http://www.analog.com/media/en/training-seminars/design-handbooks/MixedSignal_Sect6.pdf.

[27] Hegarty, C., "GPS/GNSS Operations for Engineers and Technical Professionals," Navtech Seminars Course 346, 2012.

[28] Ward, P. W., "Design Technique for Precise GNSS Receiver Post-Correlation Noise Floor Measurements with Usage Design Examples by the Search and Tracking Processes," *Proc. 2010 International Technical Meeting of The Institute of Navigation*, San Diego, CA, January 2010, pp. 607–617.

[29] Ward, P., "An Inside View of Pseudorange and Delta Pseudorange Measurements in a Digital NAVSTAR GPS Receiver," *International Telemetering Conference, GPS-Military and Civil Applications*, San Diego, CA, October 14, 1981.

[30] Ward, P., "The Natural Measurements of a GPS Receiver," *Proc. 51st Annual Meeting of The Institute of Navigation*, Colorado Springs, CO, June 1995, pp. 67–85.

[31] Jovancevic, A., et al., "Reconfigurable Dual Frequency Software GPS Receiver and Applications," *Proc. 14th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 2001)*, Salt Lake City, UT, September 2001, pp. 2888–2899.

[32] Gunawardena, S., "A Universal GNSS Software Receiver Toolbox," *InsideGNSS*, July/August 2014, http://www.insidegnss.com/auto/IGM_julaug14-Gunawardena.pdf

[33] USAF, Interface Specification, IS-GPS-800D, "Navstar GPS Space Segment/User Segment L1C Interface," September 24, 2013, http://www.gps.gov/technical/icwg/IS-GPS-800D.pdf

[34] Ledvina, B. M., et al., "Bit-Wise Parallel Algorithms for Efficient Software Correlation Applied to a GPS Software Receiver," *IEEE Trans. on Wireless Communications*, Vol. 3, No. 5, September 2004, pp. 1469–1473.

[35] Humphreys, T. E., et al., "GNSS Receiver Implementation on a DSP: Status, Challenges, and Prospects," *Proc. ION GNSS 2006*, The Institute of Navigation, Fort Worth, TX, 2006.

[36] Crowley, G., et al., "CASES: A Novel Low-Cost Ground-based Dual-Frequency GPS Software Receiver and Space Weather Monitor," *Proc. 24th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2011)*, Portland, OR, September 2011, pp. 1437–1446.

[37] Hein, G. W., et al., "Platforms for a Future GNSS Receiver - A Discussion of ASIC, FPGA, and DSP Technologies," *Inside GNSS*, March 2006, http://www.insidegnss.com/auto/0306_Working_Papers_IGM.pdf

[38] Hegarty, C., "Evaluation of the Proposed Signal Structure for the New Civil GPS Signal at 1176.45 MHz," Working Note WN 99W0000034, The MITRE Corporation, https://www.mitrecaasd.org/library/documents/gps_l5_signal.pdf

[39] Betz, J. W., *Engineering Satellite-Based Navigation and Timing: Global Navigation Satellite Systems, Signals, and Receivers*, New York: Wiley-IEEE Press, 2016.

[40] Ward, P. W., "Performance Comparisons Between FLL, PLL and a Novel FLL-Assisted-PLL Carrier Tracking Loop Under RF Interference Conditions," *Proc. 11th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1998)*, Nashville, TN, September 15–18, 1998.

[41] Hegarty, C., M. Tran, and A. J. Van Dierendonck, "Acquisition Algorithms for the GPS L5 Signal," *Proc. 16th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS/GNSS 2003)*, Portland, OR, September 2003, pp. 165–177.

[42] Smith, S. W., "The Scientist and Engineer's Guide to Digital Signal Processing," http://www.dspguide.com/ch12/4.htm

[43] McKeown, M., "FFT Implementation on the TMS320VC5505, TMS320C5505, and TMS320C5515 DSPs," Application Report SPRABB6B, Texas Instruments Incorporated, http://www.ti.com/lit/an/sprabb6b/sprabb6b.pdf

[44] Tsui, J. B. -Y., *Fundamentals of Global Positioning System Receivers: A Software Approach*, New York: John Wiley & Sons, 2000.

[45] Van Nee, D. J. R., and J. R. M. Coenen, "New Fast GPS Code-Acquisition Technique Using FFT," *Electronics Letters*, Vol. 27, No. 2, January 17, 1991, updated August 6, 2002, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=83178

[46] Ward, P., "GPS Receiver Search Techniques," *Proc. IEEE PLANS '96*, Atlanta, GA, April 1996.

[47] Scott, L., A. Jovancevic, and S. Ganguly, "Rapid Signal Acquisition Techniques for Civilian & Military User Equipments Using DSP Based FFT Processing," *Proc. 14th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 2001)*, Salt Lake City, UT, September 2001, pp. 2418–2427.

[48] Fortin, M. -A., F. Bourdeau, and R. L. Francis, Jr., "Implementation Strategies for a Software-Compensated FFT-Based Generic Acquisition Architecture with Minimal FPGA Resources," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 62, No. 3, Fall 2015, pp. 171–188.

[49] Holmes, J. K., *Spread Spectrum Systems for GNSS and Wireless Communications*, Norwood, MA: Artech House, 2007.

[50] Scott, L., "Envelope Statistics," Internal Memorandum GPS-3001, Texas Instruments Incorporated, Dallas, Texas, March 31, 1980.

[51] Tong, P. S., "A Suboptimum Synchronization Procedure for Pseudo Noise Communication Systems," *National Telemetry Conference*, November 26–28, 1973.

[52] Scott, L., "PRS Acquisition and Aiding," Internal Memorandum GPS-2924, Texas Instruments Incorporated, Dallas, TX, November 20, 1979.

[53] Scott, L., "GPS Principles and Practices," The George Washington University, Course 1081, Vol. 1, March 1994.

[54] Barron, K. S., "M of N Search Detector," Personal Correspondence, Texas Instruments Incorporated, Dallas, Texas, May 25, 1995.

[55] Przyjemski, J., E. Balboni, and J. Dowdle, "GPS Anti-Jam Enhancement Techniques," *Proc. ION 49th Annual Meeting*, June 1993, pp. 41–50.

[56] Fishman, P., and J. W. Betz, "Predicting Performance of Direct Acquisition for the M Code Signal," *Proc. ION 2000 National Technical Meeting*, January 2000.

[57] Betz, J. W., J. D. Fite, and P. T. Capozza, "DirAc: An Integrated Circuit for Direct Acquisition of the M-Code Signal," *Proc. ION GNSS 2004*, Institute of Navigation, Long Beach, CA, September 2004.

[58] Shnidman, D. A., "The Calculation of the Probability of Detection and the Generalized Marcum Q-Function," *IEEE Trans. on Information Theory*, Vol. 35, No. 2, March 1989.

[59] Martin, N., and C. Blandine, "Method and Device to Compute the Discriminant Function of Signals Modulated with One or More Subcarriers," U.S. Patent No. 2003/0231580 A1, December 18, 2003, Assignee: Bourg Les Valence (France).

[60] Lillo, W. E., P. W. Ward, and A. S. Abbott, "Binary Offset Carrier M-Code Envelope Detector," U.S. Patent No. 2005/0281325 A1, December 22, 2005, Assignee: The Aerospace Corporation, El Segundo, CA.

[61] Ward, P. W., and W. E. Lillo, "Ambiguity Removal Method for Any GNSS Binary Offset Carrier (BOC) Modulation," *Proc. 2009 International Technical Meeting of The Institute of Navigation*, Anaheim, CA, January 2009, pp. 406–419.

[62] Ward, P. W., "A Design Technique to Remove the Correlation Ambiguity in Binary Offset Carrier (BOC) Spread Spectrum Signals," *Proc. 2004 National Technical Meeting of The Institute of Navigation*, San Diego, CA, January 2004, pp. 886–896.

[63] Ward, P. W., "An Advanced NAVSTAR GPS Multiplex Receiver," *Proc. IEEE PLANS '80, Position Location and Navigation Symposium*, December 1980.

[64] Johnson, C. R., et al., "Global Positioning System (GPS) Multiplexed Receiver," U.S. Patent No. 4,468,793, August 28, 1984, Assignee: Texas Instruments Incorporated, Dallas, TX.

[65] Przyjemski, J., E. Balboni, and J. Dowdle, "GPS Anti-Jam Enhancement Techniques," *Proc. ION 49th Annual Meeting*, June 1993, pp. 41–50.

[66] IS-GPS-200, Navstar GPS Space Segment/Navigation User Interfaces, Revision H, September 24, 2013, http://www.gps.gov/technical/icwg/IS-GPS-200D.pdf

[67] Woo, K. T., "Optimum Semicodeless Processing of GPS L2," *NAVIGATION: The Journal of The Institute of Navigation*, Vol. 47, No. 2, Summer 2000, pp. 82–99.

[68] Holmes, J. D., Originally developed these analog and digital loop filter architectures and filter parameters. His first-, second-, third-, and fourth-order digital loop filter designs were used in the first commercial GPS receiver design, the TI 4100 NAVSTAR Navigator, Texas Instruments Incorporated, 1982.

[69] Stephens, S. A., and J. B. Thomas, "Controlled Root Formulation Digital Phase Locked Loops," *IEEE Trans. on Aerospace and Electronic Systems*, January 1995.

[70] Thomas, J. B., "An Analysis of Digital Phase-Locked Loops," JPL Publication 89-2, 1989.

[71] Ward, P. W., and T. D. Fuchser, "Stability Criteria for GNSS Receiver Tracking Loops," *NAVIGATION, Journal of The Institute of Navigation*, Vol. 61, No. 4, Winter 2014, pp. 293–309.

[72] Gardner, F. M., *Phaselock Techniques*, 3rd ed., Section 4.6.1, "Basis of Bode Plots," New York: John Wiley & Sons, 2005.

[73] Ward, P. W., "Using a GPS Receiver Monte Carlo Simulator to Predict RF Interference Performance," *Proc. 10th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1997)*, Kansas City, MO, September 1997, pp. 1473–1482.

[74] Ward, P. W., and K. S. Barron, "Design and Monte Carlo Simulations of a WAAS GPS Receiver Channel with Decision Feedback," *Proc. 9th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1996)*, Kansas City, MO, September 1996, pp. 1735–1743.

[75] Fuchser, T. D., "Oscillator Stability for Carrier Phase Lock," *Internal Memorandum G(S) 60233, Texas Instruments Incorporated*, February 6, 1976.

[76] Betz, J. W., and K. R. Kolodziejski, "Generalized Theory of GPS Code-Tracking Accuracy with an Early-Late Discriminator," *IEEE Transactions on Aerospace and Electronic Systems*, October 2009.

[77] Betz, J. W., and K. R. Kolodziejski, "Extended Theory of Early-Late Code Tracking for a Bandlimited GPS Receiver," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 47, No. 3. Fall 2000, pp. 211–226.

[78] Betz, J. W., "Binary Offset Carrier Modulations for Radio Navigation," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 48, No. 4. Winter 2001-2002, pp. 227–246.

[79] Betz, J. W., "Design and Performance of Code Tracking for the GPS M Code Signal," *Proc. 13th International Technical Meeting of The Satellite Division of The Institute of Navigation*, Salt Lake City, UT, September 2000, pp. 2140–2150.

[80] Wendel, J., F. M. Schubert, and S. Hager, "A Robust Technique for Unambiguous BOC Tracking," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 61, No. 3, Fall 2014, pp. 179–190.

[81] Holmes, J. D., Originally developed a similar GPS timing chart for use by the GPS Systems Engineering staff at Texas Instruments Incorporated during the GPS Phase I development program, circa 1976.

[82] Hatch, R. R., "The Synergism of GPS Code and Carrier Measurements," *Proc. 3rd International Geodetic Symposium on Satellite Doppler Positioning*, New Mexico, 1982, pp. 1213–1232.

[83] Park, B., and C. Kee, "Optimal Hatch Filter with a Flexible Smoothing Window Width," *Proc. 18th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2005)*, Long Beach, CA, September 2005, pp. 592–602.

[84] Van Dierendonck, A. J., "GPS Receivers," Ch. 8 in Parkinson, B. W., et al, (eds.), *Global Positioning System, Theory and Applications, Vol. 1*, American Institute of Aeronautics and Astronautics, Inc., Volume 163, 1996, p. 395.

[85] Natali, F. D., "Noise Performance of a Cross-Product AFC with Decision Feedback for BPSK Signals," *IEEE Trans. on Communications*, Vol. COM-34, No. 3, March 1986, pp. 303–307.

[86] Spilker, J. J., Jr., *Digital Communications by Satellite*, Ch. 12, Englewood Cliffs, NJ: Prentice Hall, 1977, pp. 347–357.

[87] Viterbi, A. J., "Convolutional Codes and Their Performance in Communication Systems," *IEEE Trans. on Information Theory*, Vol. IT-19, No. 5, October 1971, pp. 751–772.

[88] "MIT Lecture 9 on Viterbi Decoding of Convolutional Codes," MIT 6.02 DRAFT Lecture Notes, Fall 2010, http://web.mit.edu/6.02/www/f2010/handouts/lectures/L9.pdf.

[89] Hendrix, H., "Viterbi Decoding Techniques for the TMS320C55x DSP Generation," Texas Instruments Incorporated, Application Report SPRA776A, April 2009, http://www.ti.com/lit/an/spra776a/spra776a.pdf.

[90] *Viterbi Decoder: Convolutional Sublibrary of Error Detection and Correction*, MathWorks, https://www.mathworks.com/help/comm/ref/viterbidecoder.html.

[91] *HDL Code Generation for Viterbi Decoder*, MathWorks, http://www.mathworks.com/help/comm/examples/hdl-code-generation-for-viterbi-decoder.html?searchHighlight=constraint%20length%207%201%2F2%20rate%20viterbi%20decoder&s_tid=doc_srchtitle.

[92] Simon, M. K., et al., *Spread Spectrum Communications Handbook*, rev. ed., New York: McGraw-Hill, 1994, pp. 545–548.

CHAPTER 9

GNSS Disruptions

Phillp W. Ward, John W. Betz, and Christopher Hegarty

9.1 Overview

This chapter 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*, which is the focus of Section 9.2. Interference is caused by RF signals from any undesired source that is not rejected by a GNSS receiver. This disruption is commonly referred to as *radio frequency interference*. RF interference can be unintentional, for example, out-of-band emissions from other licensed RF transmitters located nearby (sometimes even co-located with) the GNSS receiver antenna that overpower the receiver's front-end bandpass filters. The interference may also be intentional and is therefore in-band. This disruption is commonly referred to as *jamming*. The types and sources and the effects of interference, as well as interference mitigation are described in Section 9.2

Section 9.3 discusses the second class of GNSS disruptions, called *ionospheric scintillation*. Ionospheric scintillation 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* that is discussed in Section 9.4. Signal blockage is manifested when the line-of-sight paths of GNSS RF signals are attenuated excessively by, for example, heavy vegetation, 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. These echoes are referred to as multipath, a term that originated from the fact that each transmitted signal is transiting over multiple paths to the receiver: the single direct path, and a number of unwanted indirect (reflected) paths. Multipath characteristics and models, effects of multipath on receiver performance and multipath mitigation are described in Section 9.5.

9.2 Interference

Because GNSS receivers rely on external RF signals they are vulnerable to RF interference (unintentional interference or jamming). RF interference can result in degraded navigation accuracy or complete loss of receiver tracking. This section first describes the types and sources of interference in Section 9.2.1. Next, the effects of interference on receiver performance are discussed in Section 9.2.2. Finally, Section 9.2.3 discusses interference mitigation techniques.

9.2.1 Types and Sources

Table 9.1 summarizes various types and potential sources of RF interference. Interference is normally classified as either *wideband* or *narrowband*, depending on whether its bandwidth is large or small relative to the bandwidth of the desired GNSS signal. Note that what might be considered wideband interference to a GNSS signal with a smaller null-to-null bandwidth (such as L1 C/A, L1C, E1 OS, or L2C) might be narrowband to a GNSS signal with a larger null-to-null bandwidth (such as L5 or E5). The ultimate limit in narrowband interference is a signal consisting of a single tone, which is referred to as a *continuous wave* (CW). (In the literature, the term continuous wave is sometimes defined differently to mean continuously transmitting, as opposed to pulsed.) The RF interference may be unintentional or intentional (jamming).

There is a certain level of interference among GNSS signals using the same carrier frequency. Such interference from signals on the same satellite is referred to as *self-interference*. Interference from signals from different satellites in the same constellation is referred to as *intrasystem interference*. Interference between two satellite constellation systems such as between GPS and Galileo signals is referred to as *intersystem interference*.

If pseudolites are used, operation at close range to these ground transmitters will almost certainly result in interference to the same satellite signals and possibly

Table 9.1 Types of RF Interference and Potential Sources

Class: Type	*Potential Sources*
Wideband: band-limited Gaussian	Intentional matched bandwidth noise jammers
Wideband: phase/frequency modulation	Television transmitter's harmonics or near-band microwave link transmitters overcoming the front-end filters of a GNSS receiver
Wideband: matched spectrum	Intentional matched-spectrum jammers, spoofers, or nearby pseudolites
Wideband: pulse	Any type of burst transmitters such as radar or ultrawideband (UWB)
Narrowband: phase/frequency modulation	Intentional chirp jammers or harmonics from an amplitude modulation (AM) radio station, Citizens Band (CB) radio, or amateur radio transmitter
Narrowband: swept continuous wave	Intentional swept continuous-wave (CW) jammers or frequency modulation (FM) stations transmitter's harmonics
Narrowband: continuous wave	Intentional CW jammers or near-band unmodulated transmitter's carriers

to other satellite signals on the same carrier frequency, although the effects of such interference can be reduced through use of burst (pulse) techniques by the pseudolites to reduce the duty cycle. In fact, an efficient wideband jamming technique uses a waveform based on the same modulation, at the same carrier frequency to form matched spectrum interference. If the intent of the source transmission is to not just disrupt GNSS operation, but rather to produce a false position within the victim receiver through the broadcast of false GNSS signals, the transmission is referred to as *spoofing*. As a benign example of spoofing, when a GNSS receiver is connected to a GNSS satellite signal simulator for testing, that receiver under test is being spoofed.

9.2.1.1 Jamming and Spoofing

Intentional jamming and spoofing must be anticipated in the design of military receivers, and are growing concerns for civilian applications as well [1]. Hence, all classes of in-band jammers, including multiple access jammers (i.e., jammers from a strategic array of multiple locations), may be considered in the design of GNSS receivers. *Smart spoofers* track the location of the target GNSS receiver and use this information along with a quasi-real-time GNSS signal generator to create strong GNSS signals that initially match the actual weaker signals in time of arrival until loop capture is assured, then lead the target receiver astray. The smart spoofer must be able to synthesize (duplicate) the target signal's received characteristics in terms of carrier frequency, spreading code, spreading modulation, and data message symbols. *Repeat-back spoofers* utilize an array of steered very high-gain antennas to track all satellites in view, then rebroadcast an amplified version toward the target receiver. The end effect on the target receiver navigation solution (if captured by these signals) is the location and velocity of the repeat-back spoofer antenna array phase center with a time bias solution that includes the common mode range between the spoofer and the victim receiver. The major weakness of both spoofing techniques is that all the spoofing signals arrive from the same direction (unless spatial diversity is also used), so that directional null-steering antenna techniques can be used to defeat spoofers. The GPS encrypted antispoofing (AS) Y code is used to replace the public P code for military applications to minimize the potential for spoofing military GPS receivers. The GPS encrypted M code is even more secure. Other constellation signals that are encrypted include Galileo E1 PRS and E6 PRS and BeiDou B1-A, B3, B3-A, and B2 and future GLONASS L3OC, L1SC, and L2SC.

9.2.1.2 Unintentional Interference

Unintentional RF interference can be expected at low levels for a GNSS receiver operating practically anywhere in the world. There are a large number of other essential systems that rely on the transmission of RF energy within L-band, so these are also potential sources of unintentional RF interference. Table 9.2 shows the International and U.S. Tables of Frequency Allocations near those used by GNSS signals. The services shown in all capital letters are *primary*, and the ones shown with initial capitals are *secondary*. Secondary services are permitted to operate in their designated bands but are not generally provided protection from the primary

services and further are not allowed to provide harmful levels of interference to the primary services.

As shown in Table 9.2, the 1,559–1,610-MHz band is designated for use by only satellite navigation signals in most regions of the world. The L1 signals of GPS, GLONASS, Galileo, BeiDou, QZSS, and SBAS are within this protected band.

The GLONASS L2, Galileo E6, BeiDou B3, and QZSS L6 signals are in the 1,240–1,300-MHz band. The L2 signals of GPS and QZSS are in the 1,215–1,240-MHz band. A number of countries permit fixed and mobile services to operate in this band. It has a coprimary radiolocation allocation worldwide and is shared by Radiolocation services that operate in the band that include a large number of radars that are used for air traffic control, military surveillance, and drug interdiction. Some of these radars operate with very high transmit power (kilowatts to megawatts). They are pulsed systems and fortunately, as will be discussed in Section 9.2.3, GNSS receiver front-end designs can be made very robust against pulsed interference by various means of pulse amplitude suppression (blanking) during the low-duty cycle pulse intervals.

The GPS L5, Galileo E5A and E5B, BeiDou B2, NAVIC L5, QZSS L5, and SBAS L5 signals are in the 1,164–1,215-MHz band. The 960–1,215-MHz band is also used worldwide for electronic aids to air navigation. Distance Measuring Equipment (DME) and Tactical Air Navigation (TACAN) ground beacons transmit at power levels up to 10 kW on frequencies that fall within the passband of a receiver processing 1,164–1,215-MHz GNSS signals. Some nations also permit the use of Link 16, a tactical military communications system with radios that nominally transmit 200W over 51 frequencies throughout the 960–1,215-MHz band. Fortunately, DME/TACAN and Link 16 are pulsed.

It is inevitable that some out-of-band energy from the signals in adjacent bands will at times fall within the range of frequencies processed by GNSS receivers. This energy can originate from *adjacent band interference*, from *harmonics*, or from *intermodulation products*. Adjacent band interference can occur from the spillover of energy from bands immediately above or below one of the GNSS carrier frequencies (i.e., the adjacent band transmitters have not adequately suppressed their energy outside of their allocated frequency band). There is also the contemporary threat of high-density, ground-based, high-powered transmitters operating in adjacent bands to a GNSS band but keeping its high power within acceptable limits outside that adjacent band. There is a clear and present danger for this to happen to the GNSS frequency bands [2]. It can happen if the adjacent band is reassigned for high-density ground transmitter use when that band had been historically reserved and authorized for space-based satellite transmitters with power levels on or near the Earth's surface that are below the thermal noise level [3]. This creates operational GNSS receiver adjacent band power levels that are more than a billion times powerful than historical levels. This, in turn, obsoletes many existing GNSS receivers operating in that band and creates difficult bandpass filtering problems for next-generation GNSS receivers operating adjacent to that band.

Harmonics are signals at integer multiples of the carrier frequency of a transmitter that are caused by nonlinearities (e.g., saturation of an amplifier that leads to clipping) upon transmission. Intermodulation products occur when two or more signals at different frequencies are passed through a nonlinearity.

Table 9.2 Frequency Allocations Near GNSS

International Table	*U.S. Table*	*Notes*
960–1,164 MHz AERONAUTICAL MOBILE (R), AERONAUTICAL RADIONAVIGATION	AERONAUTICAL MOBILE (R), AERONAUTICAL RADIONAVIGATION	Used worldwide for electronic aids to air navigation including Distance Measuring Equipment (DME), Tactical Air Navigation (TACAN), Secondary Surveillance Radar (SSR), and Automatic Dependence Surveillance (ADS). Some nations permit Link 16 (a military communication system) usage on a noninterference basis.
1,164–1,215 MHz AERONAUTICAL RADIONAVIGATION, RADIONAVIGATION-SATELLITE	AERONAUTICAL RADIONAVIGATION, RADIONAVIGATION-SATELLITE	Lower L-band signals of GPS, Galileo, BeiDou, NAVIC, QZSS, SBAS and future GLONASS CDMA signals are in the 1164-1215 MHz band.
1,215–1,240 MHz EARTH EXPLORATION-SATELLITE (active), RADIOLOCATION, RADIONAVIGATION-SATELLITE, SPACE RESEARCH (active)	EARTH EXPLORATION-SATELLITE (active), RADIOLOCATION, RADIONAVIGATION-SATELLITE, SPACE RESEARCH (active)	Used worldwide for primary radars for purposes including air traffic control, military surveillance, and drug interdiction. Many countries have coprimary allocations for fixed and mobile services. Band also used for active spaceborne sensors for, for example, ocean surface measurements.
1,240–1,300 MHz EARTH EXPLORATION-SATELLITE (active) RADIOLOCATION RADIONAVIGATION-SATELLITE SPACE RESEARCH (active) Amateur	AERONAUTICAL RADIONAVIGATION, EARTH EXPLORATION-SATELLITE (active), RADIOLOCATION, SPACE RESEARCH (active), Amateur	GPS L2 and QZSS are at 1,227.6 MHz, the upper part of the GPS L2 band extends to 1,227.6 + 15.345 = 1,242.945 MHz. GLONASS L2, Galileo E6, BeiDou B3 and QZSS L6 frequencies are within the 1,240–1,300 MHz band.
1,300–1,350 MHz AERONAUTICAL RADIONAVIGATION, RADIOLOCATION, RADIONAVIGATION-SATELLITE	AERONAUTICAL RADIONAVIGATION, Radiolocation	
1,350–1,400 MHz FIXED*, MOBILE*, RADIOLOCATION	FIXED*, LAND MOBILE* MOBILE*, RADIOLOCATION*	Varied band usage worldwide among fixed services, land mobile, mobile, and radiolocation.
1,525–1,559 MHz MOBILE-SATELLITE (space-to-Earth) SPACE OPERATION (space-to-Earth) Earth exploration-satellite Fixed* Mobile*	MOBILE-SATELLITE (space-to-Earth)	Downlink frequencies for satellite communications services (e.g., INMARSAT)
1,559–1,610 MHz AERONAUTICAL RADIONAVIGATION RADIONAVIGATION-SATELLITE	AERONAUTICAL RADIONAVIGATION RADIONAVIGATION-SATELLITE	The upper L-band signals of GPS, GLONASS, Galileo, BeiDou, QZSS and SBAS signals are within this band.
1,610–1,626.5 MHz MOBILE-SATELLITE (Earth-to-space) AERONAUTICAL RADIONAVIGATION RADIO ASTRONOMY* RADIODETERMINATION-SATELLITE* Mobile-satellite (space-to-Earth)*	MOBILE-SATELLITE (Earth-to-space) AERONAUTICAL RADIONAVIGATION RADIO ASTRONOMY* RADIODETERMINATION-SATELLITE Mobile-satellite (space-to-Earth)*	Uplink frequencies for commercial satellite communications services. Portions of the band are protected for radio astronomy sensors.

*Only in some nations or portions of the band.

Even if interfering signals are out of the nominal band processed by a GNSS receiver, strong RF signals can still deteriorate GNSS receiver performance (e.g., by saturating the low-noise amplifiers used in the receiver front end). Although regulations are in place within the United States and internationally to protect GNSS spectrum, there are occasionally instances of equipment malfunctions or equipment misuse that can lead to intolerable levels of interference. Nonlinear effects (e.g., amplifier saturation) may accidentally occur in high-powered transmitters causing lower power harmonics that become in-band RF interference to GNSS receivers. The offending transmitter source has to be located and corrected before normal GNSS operation in that vicinity can resume. In some regions of the world there are more frequent problems with interference to GNSS than in others. For example, in the Mediterranean, there were a number of reports of GPS L1 C/A code receivers failing to operate properly because of strong in-band harmonics from television (TV) transmitters in the region, but this now appears to have been largely corrected primarily because these harmonics also seriously deteriorate the video quality on the TV receivers.

9.2.2 Effects

The performance of signal acquisition, carrier tracking, and data demodulation all depend on the signal-to-noise-plus-interference ratio (SNIR) at the output of each correlator in a receiver. Consequently, evaluating the effect of RF interference on correlator output SNIR provides the basis for assessing the effect of this interference on these three receiver functions. This section describes the underlying theory behind this effect and then presents approximation techniques for such analysis.

When the aggregate interference can be modeled as statistically stationary, and when the spectra of either the interference or the desired signal (or both) are well approximated by a straight line over a bandwidth that is the reciprocal of the integration time used in the correlation, the prompt correlator output SNIR is as follows [4]

$$\rho_c = \frac{2TC_s/N_0\left[\int_{-\beta_r/2}^{\beta_r/2} S_s(f)e^{j2\pi f\tau}df\right]^2}{\int_{-\beta_r/2}^{\beta_r/2}\left|H_R(f)\right|^2 S_s(f)df + C_t/N_0\int_{-\beta_r/2}^{\beta_r/2}\left|H_R(f)\right|^2 S_t(f)S_s(f)df} \tag{9.1}$$

where T is the integration time of the correlator (in seconds), C_S is the received power of the desired received signal (in watts) over infinite bandwidth, N_0 is the power spectral density of the white noise (in W/Hz), $H_R(f)$ is the transfer function of the receiver (ratio), $S_S(f)$ is the power spectral density (in W/Hz) of the transmitted signal, normalized to unit area over a specified bandwidth, the transmit filter on the satellite is ideal, the effect of all filtering in the receive chain is modeled as bandlimited to $-\beta_r/2 \le f \le \beta_r/2$, C_t, is the power of the received interference signal (in watts) and $S_t(f)$ is the power spectral density (in W/Hz) of the aggregate received interference, normalized to unit area over infinite bandwidth. (The transmitted

signal bandwidth needs to be defined. Different conventions may use an infinite bandwidth, a defined transmit bandwidth, or the precorrelation bandwidth of the receiver. As long as the defined bandwidth is greater than the null-to-null bandwidth of the signal, the numerical differences are usually less than 1 dB.)

The quality of a received GNSS signal is commonly described in terms of its carrier-power-to-noise-density ratio implying that the noise is white and thus can be described by a scalar noise density. Yet (9.1) shows that any nonwhite interference must be accounted for as well, and must be described by its power spectral density, including its power. Thus, analyzing the correlator output SNIR in interference is extremely cumbersome. However, if a fictitious white noise density is formulated that produces the same output SNIR as the combination of the actual white noise and interference, then the resulting effective carrier-power-to-noise-density ratio is both correct and straightforward to analyze using the fiction of effective white noise.

To derive an *effective* C_S/N_0, or $(C_S/N_0)_{eff}$, observe that (9.1) with no interference and infinite receive bandwidth, the signal-to-noise ratio (SNR) at the prompt correlator tap is

$$\rho_c = 2TC_s / N_0 \tag{9.2}$$

Equivalently, C_S/N_0 can be found from the SNIR at the prompt correlator tap as:

$$C_s/N_0 = \frac{\rho_c}{2T} \tag{9.3}$$

When there is both interference and white noise, $(C_S/N_0)_{eff}$ is defined in a way analogous to (9.3), but using the output SNR from (9.1) as follows

$$\begin{aligned}
\left(C_s/N_0\right)_{eff} &= \frac{\rho_c}{2T} \\
&= \left(C_s/N_0\right)\frac{\left[\int_{-\beta_r/2}^{\beta_r/2} S_s(f)df\right]^2}{\int_{-\beta_r/2}^{\beta_r/2} S_s(f)df + \frac{C_t}{N_0}\int_{-\beta_r/2}^{\beta_r/2} S_t(f)S_s(f)df} \\
&= \frac{\int_{-\beta_r/2}^{\beta_r/2} S_s(f)df}{\frac{1}{\left(C_s/N_0\right)} + \frac{C_t/C_s}{\int_{-\beta_r/2}^{\beta_r/2} S_s(f)df \,/\, \int_{-\beta_r/2}^{\beta_r/2} S_t(f)S_s(f)df}}
\end{aligned} \tag{9.4}$$

Observe that (9.4) can be expressed as [5]

$$\left(C_s/N_0\right)_{eff} = \frac{\int_{-\beta_r/2}^{\beta_r/2} S_s(f)df}{\frac{1}{\left(C_s/N_0\right)} + \frac{C_t/C_s}{QR_c}} \tag{9.5}$$

where C_S/N_0 is the unjammed carrier-to-noise-power ratio of the received signal before receiver filtering, C_t/C_S is the jamming-to-received-signal power ratio before receiver filtering, Q is a dimensionless jamming resistance quality factor to be determined for various types of jammers and signal modulators, and R_c is the spreading code rate of the code generator in chips/s. Note that increasing the value of Q in (9.5) improves $(C_S/N_0)_{eff}$. Therefore, higher jamming resistance quality factor, Q, results in increased jamming effectiveness. Comparing (9.4) and (9.5) yields

$$Q = \frac{\int_{-\beta_r/2}^{\beta_r/2} S_s(f)df}{R_c \int_{-\beta_r/2}^{\beta_r/2} S_t(f)S_s(f)df} \tag{9.6}$$

Then (9.6) can be expressed succinctly as

$$Q = \frac{\int_{-\infty}^{\infty} \left|H_R(f)\right|^2 S_s(f)df}{R_c\kappa_{ts}} \tag{9.7}$$

where κ_{ts} is called the spectral separation coefficient (SSC) [5], which is defined as

$$\kappa_{ts} = \int_{-\beta_r/2}^{\beta_r/2} S_t(f)S_s(f)df \tag{9.8}$$

Equation (9.8) has units of seconds or reciprocal hertz. Observe that the SSC depends on the spectrum of the desired signal as well as the spectrum of the interference. Different interferers may have the same SSC with a given desired signal, and when they do, the different interferers affect $(C_S/N_0)_{eff}$ the same way if the interfering signals are received with equal powers. It follows then that different interferers may have a different SSC with a given desired signal. For example, if interferer A has x dB smaller SSC with the desired signal than interferer B, then interferer A has the same effect on $(C_S/N_0)_{eff}$ as interferer B when the power in A is increased by x dB.

9.2.2.1 Computing Jamming Resistance Quality Factor Q

To consider interference effects further for some nominal situations, suppose that the receive filter is very wide, so that $H_R(f)$ can be treated as approximately unity in (9.7) at frequencies where the desired signal has appreciable power, and the limits on the integrals can be approximated by infinity, so that (9.7) is approximated by

$$Q \cong \frac{1}{R_c \int_{-\infty}^{\infty} S_I(f) S_s(f) df} \tag{9.9}$$

Examples of different types of interference can now be evaluated.

- *Case 1—Narrowband Interference.* For narrowband interference centered at f_I, the power spectrum can be modeled as $S_I(f) = \delta(f - f_I)$, where $\delta(\cdot)$is the Dirac delta function having infinite amplitude, vanishing width, and unit area. Substituting for this interference power spectral density in (9.9) yields

$$Q = \frac{1}{R_c S_s(f_I)} \tag{9.10}$$

In general, narrowband interference affects $(C_S/N_0)_{eff}$ more when the interference frequency is at or near the maximum of the desired GNSS signal power spectrum. Moreover, when the normalized power spectrum of the desired signal has a smaller maximum, the desired signal is degraded less by narrowband interference at the worst-case frequency.

The baseband power spectral density functions for BPSK-R(n) and BOC$_s$(m,n) signals are given in Section 2.4.4. If the narrowband interference is placed at the spectral maximum of a BPSK-R(n) signal (f = 0 for the baseband power spectral density), $S_S(f_I = 0) = 1/R_c$, and (9.10) becomes $Q = \frac{1}{R_c / R_c} = 1$. If instead the interference is placed at a frequency other than the signal's spectral peak, Q is greater than unity, meaning that the interference has less effect. For a BOC(m,n) modulation, if the interferer is located at one or both of the spectral peaks, Q takes on values in the range $1.9 \leq Q \leq 2.5$, depending upon the subcarrier frequency, the spreading code rate, and whether cosine-phasing or sine-phasing is used. If the narrowband interferer is at any other frequency, Q again takes on larger values, indicating that interference of fixed power has less effect on $(C_S/N_0)_{eff}$.

For signals having AltBOC spreading modulations like Galileo E5 and BeiDou B2, if there is a single narrowband interferer, the minimum value of Q is approximately 0.5 when the interference is located at the spectral peak, when considering the total power in the AltBOC waveform. If the signal is considered to be a BPSK-R signal on only one subcarrier, the minimum value of Q is approximately unity. For signals having an MBOC spectrum like L1C, E1 OS and B1-C, when the narrowband interference is on the spectral

peak, the value of Q is 2.1. As the interference moves from the spectral peak in any of these cases, the value of Q increases.

- *Case 2—Matched Spectrum Interference.* Consider now when the interference has the same power spectral density as the desired signal. This situation could arise from multiple access interference, or from a jamming waveform whose spectrum is matched to that of the desired signal.

$$Q = \frac{1}{R_c \int_{-\infty}^{\infty} \left[S_s(f) \right]^2 df} \tag{9.11}$$

When the signal is BPSK-R(n), substituting (2.25) into (9.11) yields

$$Q = \frac{1}{R_c \int_{-\infty}^{\infty} T_c^2 \, \mathrm{sinc}^4 \left(\pi f T_c \right) df} = 1.5 \tag{9.12}$$

For a BOC(m,n) modulation, Q takes on values in the range $3 \le Q \le 4.5$, depending upon the subcarrier frequency, the spreading code rate, and whether cosine phasing or sine phasing is used. For matched spectrum interference to full bandwidth signals having AltBOC spreading modulation, Q is approximately 4.4. For matched spectrum interference to signals having MBOC spreading modulation, Q is approximately 3.6.

- *Case 3—Bandlimited White Noise Interference.* When the interference has flat spectrum centered at f_t and extending from $f_t - \beta_t/2 \le f \le f_t + \beta_t/2$, its spectrum is expressed as

$$S_t(f) = \begin{cases} \dfrac{1}{\beta_t}, & f_t - \beta_t / 2 \le f \le f_t + \beta_t / 2 \\ 0, & \text{elsewhere.} \end{cases} \tag{9.13}$$

Substituting (9.13) into (9.9) yields

$$Q = \frac{1}{\dfrac{R_c}{\beta_t} \int_{f_t - \beta_t/2}^{f_t + \beta_t/2} S_s(f) df} \tag{9.14}$$

If β_t becomes small, (9.14) approaches (9.10), the result for narrowband interference.

If β_t is large enough so that almost all of the signal power is included within $f_t - \beta_t/2 \le f \le f_t + \beta_t/2$, then (9.14) becomes

$$Q = \frac{\beta_t}{R_c} \tag{9.15}$$

Rearranging (9.15), $QR_c = \beta_t$ which shows that modulation design, and in particular higher spreading code rates, provide no benefit to $(C_S/N_0)_{eff}$ when the noise spectrum is flat over the frequency range occupied by the signal. Moreover, for fixed interference power, the wider the interference bandwidth, the larger the value of Q, and hence the smaller influence of the interference on $(C_S/N_0)_{eff}$.

When the signal is BPSK-R(n) and the interference spectrum is centered on the signal spectrum so that $f_t = 0$, substituting (2.25) into (9.14) yields

$$Q = \frac{1}{\frac{1}{\beta_t}\int_{\beta_t/2}^{\beta_t/2} \operatorname{sinc}^2(\pi f T_c)\,df} \tag{9.16}$$

When in addition $\beta_t = 2R_c$ so that the interference covers the null-to-null main lobe of the signal spectrum, (9.16) becomes

$$Q = \frac{2}{\int_{-R_c}^{R_c} \operatorname{sinc}^2(\pi f T_c)\,df} \cong 2.2 \tag{9.17}$$

When the modulation is BOC(m.n) and the interference spectrum extends from $-(m+n) \times 1.023$ MHz to $(m+n) \times 1.023$ Q can be very large when the subcarrier frequency is much greater than the chip rate. In this case, the center of the frequency band has little signal power, making the jamming inefficiently matched to the signal spectrum. The value of Q is 4.7 for BOC(m,m), and increasingly larger for BOC(m,n) with $m > n$.

Table 9.3 summarizes the above Q's for C/A, L2C, P(Y), L5, M, E1B, E1C and E5 signals, along with their associated modulation types and spreading code rates for the three classes of jammer types analyzed above.

9.2.2.2 Computing J/S and Tolerable Jamming Power

Substituting the approximation that $\int_{-\beta_r/2}^{\beta_r/2} S_s(f)\,df = 1$ in (9.5) assumes that a negligible amount of the available signal power is lost in the finite bandwidth of the receiver [6] and the simplified equation becomes

Table 9.3 Examples of Jamming Resistance Quality Factors (Q)

	Q for Different Interference Spectra		
Modulation/Signal	*Narrowband at spectral peak(s)*	*Matched spectrum*	*Bandlimited white noise null-to-null spectrum*
BPSK	1	1.5	2.2
GPS: C/A, L2C, P(Y), L5			
SBAS: L1, L5			
GLONASS: L1OF, L1SF, L2OF, L2SF, L3OC, L1OC and L2OC			
Galileo: E6 CS, E5a, E5b			
BeiDou: B1I, B1Q, B2I, B2Q, B2a, B2b, B3			
QZSS: C/A, L1S, L2C, L5, L5S			
NAVIC: S SPS, L5 SPS			
BOC(10,5)	2.3	4.0	7.2
GPS: M			
GLONASS: L1SC, L2SC			
MBOC	2.1	3.6	5.1
GPS:L1C			
Galileo: E1 OS			
BeiDou: B1-C			
QZSS: L1C			
GLONASS BOC(1,1)	1.9	3.0	4.7
GLONASS: L1OC and L2OC			
$BOC_c(15,2.5)$	2.5	4.5	18.7
Galileo: E1 PRS			
$BOC_c(10,5)$	2.4	4.4	8.2
Galileo: E6 PRS			
AltBOC(15,10)	2.5	4.4	11.3
Galileo: E5			
BeiDou: B2			
BOC(14,2)	2.5	4.5	20.0
BeiDou: B1-A			
BOC(15,2.5)	2.4	4.4	17.4
BeiDou: B3-A			
BOC(5,2)	2.4	4.2	8.5
NAVIC: S RS, L5 RS			

$$
\begin{aligned}
(C_s/N_0)_{eff,dB} &\triangleq 10\log_{10}\left(C_s/N_0\right)_{eff} \\
&= -10\log_{10}\left[10^{-\frac{(C_s/N_0)_{dB}}{10}} + \frac{10^{\frac{(C_t/C_s)_{dB}}{10}}}{QR_c}\right] \quad \text{(dB-Hz)}
\end{aligned}
\tag{9.18}
$$

where

$(C_S/N_0)_{dB} = 10\log_{10}(C_S/N_0)$ (dB-Hz)

$(C_I/C_S)_{dB} = 10 \log_{10} (C_I/C_S)$ (dB)

Q = jamming resistance quality factor (dimensionless)

R_c = spreading code chipping rate (chips/s)

Equation (9.18) shows that the effect of jamming is to reduce the unjammed $(C_S/N_0)_{dB}$ to a lower value, $(C_S/N_0)_{eff,dB}$. As discussed in Chapter 8, the signal acquisition, carrier tracking, and data demodulation functions deteriorate as $(C_S/N_0)_{eff,dB}$ is reduced. There is a region of deterioration for each function in which these functions are likely to fail, but a 1-sigma threshold is typically used to describe when that limit has been reached. Typically, data demodulation and signal acquisition are the first to be lost as $(C_S/N_0)_{eff,dB}$ is reduced (i.e., have higher thresholds than for carrier tracking). Chapter 8 shows that interference affects code tracking differently from carrier tracking, and that, in general, code tracking is more robust against the effect of interference than carrier tracking, so separate assessment must be performed to evaluate the effect on code tracking and loss of lock, but code-tracking threshold is meaningful only if there is precise external velocity aiding that can estimate the carrier Doppler in an open loop fashion after the carrier loop has lost lock. If there is no external velocity aiding available, then the code-tracking loop loses lock very shortly after the carrier-tracking loop loses frequency lock owing to severe signal roll-off due to rapid deterioration of the carrier wipe-off process. Note that the code-tracking loop cannot provide the carrier Doppler estimate reliably, but a known stationary antenna condition can provide an ideal Doppler estimate. When $(C_S/N_0)_{eff,dB}$ has been reduced by jamming to the code tracking loop threshold, the aided receiver loses lock.

Equation (9.18) can be rearranged to solve for $(C_I/C_S)_{dB}$ (the interference to signal ratio at the antenna input in decibels) as follows

$$\left(C_I/C_s\right)_{dB} = 10\log_{10}\left[QR_c\left(10^{-\frac{(C_s/N_0)_{eff,dB}}{10}} - 10^{-\frac{(C_s/N_0)_{dB}}{10}}\right)\right] \tag{9.19}$$

Computing the unjammed $(C_S/N_0)_{dB}$ in (9.18) and (9.19) in units of dB-Hz involves numerous parameters and is presented piecewise as follows

$$\begin{aligned}
\left(C_s/N_0\right)_{dB} &= \left(C_s\right)_{dB} - \left(N_0\right)_{\mathrm{dB}} \quad \text{(dB-Hz)} \\
\left(C_s\right)_{dB} &= \left(C_{Ri}\right)_{dB} + \left(G_{SVi}\right)_{dB} - L_{dB} \quad \text{(dBW)} \\
\left(N_0\right)_{dB} &= 10\log_{10}\left[k\left(T_{ant} + T_{receiver}\right)\right] \quad \text{(dBW)} \\
T_{receiver} &= 290\left(10^{\frac{(N_f)_{dB}}{10}} - 1\right) \quad \text{(K)}
\end{aligned} \tag{9.20}$$

where

$(C_S)_{dB}$ = recovered signal power received from SV_i (dBW);

$(N_0)_{dB}$ = thermal noise power component in a 1-Hz bandwidth (dBW/Hz);

$(C_{Ri})_{dB}$ = received signal power from SV_i at antenna input (dBW);

$(G_{SVi})_{dB}$ = antenna gain toward SV_i (dBic);

L_{dB} = receiver implementation loss including A/D converter loss (dB);

k = Boltzmann's constant = 1.38×10^{-23} (J/K);

T_{ant} = antenna noise temperature (K);

$T_{receiver}$ = receiver system temperature (K);

$(N_f)_{dB}$ = receiver noise figure at 290K (dB).

As a computation example of (9.20) for the GPS L1 C/A code signal, assume $(C_{Ri})_{dB}$ = –158.5 dBW (i.e., the IS-GPS-200 minimum specified received signal power level). Further assume a typical RHCP fixed reception pattern antenna (FRPA) is used with a gain rolloff to about –3 dBic at the elevation mask angle of 5° above the horizon. This is also the elevation angle where the minimum GPS received power specification is met. It is typical for a FRPA gain to increase to 1.5 dBic or more at zenith, where the GPS minimum received power specification is also met. In between these two elevation angles, the received signal power tends to increase slightly due to the satellite antenna array gain pattern. In other words, the received signal power and antenna gain combination tends to be lower by about –3 dB near the elevation mask angle of 5° and higher by about 1.5 dB at zenith with a fluctuation range of more than 4.5 dB in the approximately hemispherical gain coverage region of a typical FRPA. Antenna tilt can significantly increase this gain fluctuation range and it also can increase the noise temperature owing to the high temperature of the Earth. In this example, the antenna is assumed to have $(G_{SVi})_{dB}$ = 1.5 dB gain toward the SV to allow for the higher SV signal levels that exist most of the time counting the gains of both the receiver antenna and the SV antenna. The implementation loss (including A/D converter loss) is assumed to be 2 dB (L_{dB} = 2) for this high-quality receiver design example. Using these assumptions in (9.20), the total recovered signal power is $(C_s)_{dB}$ = –158.5 + 1.5 – 2 = –159.0 dBW.

Next assume that the antenna noise temperature, T_{ant} = 100K (see Section 8.2) and a front-end design that provides a low-noise figure, $(N_f)_{dB}$ = 2 dB at 290K, so $T_{receiver} = 290 \times (10^{0.2} - 1) = 169.6$K. Using these assumptions, the thermal noise can be computed as $N_0 = 10 \log [k \times (100 + 169.6)]$ = –204.3 dBW/Hz. Therefore, the unjammed $(C_S/N_0)_{dB}$= –159.0 + 204.29 = 45.3 dB-Hz.

Note that the unjammed $(C_S/N_0)_{dB}$ in (9.20) accounts for the antenna gain in the direction of the satellite as well as the implementation loss of the receiver. Similarly, if the antenna gain in the direction of the jammer, $(G_J)_{dB}$, is accounted for in (9.19), then

$$\begin{aligned}
\left(C_t / C_s\right)_{dB} &= \left(C_t\right)_{dB} - \left(C_s\right)_{dB} \\
\left(C_t\right)_{dB} &= J_{dB} + \left(G_J\right)_{dB} - L_{dB} \\
\left(C_s\right)_{dB} &= \left(C_{Ri}\right)_{dB} + \left(G_{SVi}\right)_{dB} - L_{dB} = S_{dB} + \left(G_{SVi}\right)_{dB} - L_{dB} \\
\left(C_t / C_s\right)_{dB} &= J_{dB} - S_{dB} + \left(G_J\right)_{dB} - \left(G_{SVi}\right)_{dB} = \left(J / S\right)_{dB} + \left(G_J\right)_{dB} - \left(G_{SVi}\right)_{dB}
\end{aligned} \tag{9.21}$$

where $(J/S)_{dB}$ is the jamming to signal power ratio at the antenna input in decibels. Substituting this into (9.19)

$$(J/S)_{dB} = (G_{SVi})_{dB} - (G_J)_{dB} + 10\log_{10}\left[QR_c\left(10^{-\frac{(C_s/N_0)_{eff,dB}}{10}} - 10^{-\frac{(C_s/N_0)_{dB}}{10}}\right)\right] \quad (9.22)$$

From (9.22), the receiver $(J/S)_{dB}$ performance can be computed for a given QR_c using the unjammed $(C_S/N_0)_{dB}$ from (9.20) and obtaining the value of $(C_S/N_0)_{eff,\ dB}$ by simply equating it to the receiver tracking threshold as determined from the approximation methods presented in Chapter 8. Recall that the carrier-tracking threshold $(C_S/N_0)_{eff,\ dB}$ is the weak link for an unaided GPS receiver.

As a computational example of (9.22) using the unjammed C/A code signal example where $(C_S/N_0)_{dB}$ = 45.3 dB-Hz, assume that the antenna gain toward the jammer, $(G_J)_{dB}$, is –3 dBi. Note that the jammer signal may or may not be right-hand circularly polarized (RHCP). If RHCP, then the antenna gain toward the jammer would be the same as its gain in that direction for an SV. But if the jammer is linearly polarized, then an additional 3-dB loss or so must be included in the gain toward the jammer depending on the polarization mismatch of the GPS antenna. If the jammer is ground-based, then recall from Section 8.2 that the typical GNSS antenna becomes almost linearly polarized as the elevation angle approaches the horizon.

Since the desired signal is C/A code signal with a BPSK-R(1) modulation, R_c = 1.023×10^6 chips/s, assume a band-limited white noise (BLWN) jamming waveform whose spectrum is rectangular, centered at the C/A center frequency, and approximately 2 MHz wide (null-to-null), so that Q = 2.2. Assume that the L1 C/A code PLL carrier-tracking threshold is $(C_S/N_0)_{eff,dB}$ = 27 dB-Hz. Substituting these and the unjammed $(C_S/N_0)_{dB}$ from the previous computational example into (9.22)

$$\begin{aligned}(J/S)_{dB} &= 1.5 + 3.0 + 10\log_{10}\left[2.2 \times 1.023 \times 10^6\left(10^{-2.7} - 10^{-4.529}\right)\right] \\ &= 41.0 \text{ dB}\end{aligned}$$

For L1 P(Y) code signal, there is BPSK-R(10) modulation with R_c = 10.23 Mchips/s and $(C_{Ri})_{dB}$ = –161.5 dBW. Assume the jamming waveform has a BLWN rectangular spectrum centered on L1 with width of 20.46 MHz (null-to-null) so that Q = 2.2 (and the remaining assumptions the same). The unjammed $(C_S/N_0)_{dB}$ = 42.3 dB-Hz, so assuming that the PLL tracking threshold is the same as for C/A code signal, then

$$\begin{aligned}(J/S)_{dB} &= 1.5 + 3.0 + 10\log_{10}\left[2.2 \times 10.23 \times 10^6\left(10^{-2.7} - 10^{-4.229}\right)\right] \\ &= 50.9 \text{ dB}\end{aligned}$$

Note that if the unjammed $(C_S/N_0)_{dB}$ for P(Y) code signal had been exactly the same as for C/A code signal above, the $(J/S)_{dB}$ would be exactly 10 dB greater, reflecting the factor of 10 increase in spreading code chip rate.

For the modernized M code signal with $BOC_s(10,5)$ modulation and R_c = 5.115 Mchips/s, the minimum specified received signal level for Block II satellites at normal power is $(C_{Ri})_{dB}$ = –158.0 dBW. However, this must be reduced by 3 dB

to –161 dBW since the pilot component (at 50% of the total power) is always used to achieve substantial improvement in tracking threshold. For the third order PLL, assume that the M code PLL achieves a $(C/N_0)_{eff,dB}$ = 17 dB-Hz (using the same antenna and receiver parameters in the previous two examples). The M code pilot component is extracted by using the data-less intervals of M code time division data modulation (TDDM) in the PLL carrier-tracking loop. Assume that the BLWN jamming spectrum consists of two (null-to-null) rectangles, centered ±10.23 MHz away from L1, each with a width of 10.23 MHz and that Q = 7.2. The unjammed C_S/N_0= 42.8 dB-Hz, so

$$\begin{aligned}(J/S)_{dB} &= 1.5 + 3.0 + 10\log_{10}\left[7.2 \times 5.115 \times 10^6 \left(10^{-1.7} - 10^{-4.28}\right)\right] \\ &= 63.2 \text{ dB}\end{aligned}$$

Table 9.4 shows the receiver $(J/S)_{dB}$ performance for GPS L1 C/A, L1 P(Y) and M (TDDM) signals used in the above BLWN wideband jammer examples plus GPS L2C and L5 signals, for three types of jamming (wideband, matched spectrum and narrowband). Table 9.5 shows the $(J/S)_{dB}$ performance for GPS L1C, Galileo E1 OS, BeiDou B1I, and GLONASS L1OF (FDMA) signals. The table parameters shown in the signal rows include the signal names, their chipping rates, R_c, spreading code modulation types, their specified minimum received signal powers, $(C_{Ri})_{dB}$, the unjammed carrier to noise power ratios, $(C_S/N_0)_{dB}$, and the unaided third-order PLL carrier-tracking thresholds, $(C_S/N_0)_{eff,dB}$, associated with each signal. The applicable Q is shown in parenthesis adjacent to each J/S entry for reference since this dimensionless factor makes a significant difference in signal robustness to jamming and is unique for each type of jammer.

Note that Table 9.4 produces fifteen different values of $(J/S)_{dB}$ performance for identical antenna performance, receiver reference oscillator quality, random vibration profile and dynamic stress environment assumptions. In other words, the basis

Table 9.4 *J/S* Performance Comparisons with Appropriate *Q* for GPS (except L1C)

Signal/Jammer Type	$(J/S)_{dB}$ *(dB), [Q (Dimensionless)]*				
Signal	L1 C/A	L1 P(Y)	L1 M (TDDM)[1]	L2 CL[1]	L5 Q5[1]
R_c *(chips/s)*	1.023×10^6	10.23×10^6	5.115×10^6	1.023×10^6	10.23×10^6
Modulation type	BPSK-R(1)	BPSK-R(10)	$BOC_s(10,5)$	BPSK-R(1)	BPSK-R(10)
$(C_{Ri})_{dB}$ *(dBW)*	–158.5	–161.5	–161.0	–163 (IIF)[2] –161.5 (III)	–157.9 (IIF)[3] –157.0 (III)
$(C_s/N_0)_{dB}$ *(dB-Hz)*	45.29	42.29	42.79	40.79	45.89
$(C_s/N_0)_{eff,dB}$ *(dB-Hz)*	27	27	17	17	17
Wideband null-to-null	41.0 [2.2]	50.9 [2.2]	63.2 [7.2]	51.0 [2.2]	61.1 [2.2]
Wideband matched spectrum	39.3 [1.5]	49.2 [1.5]	60.6 [4.0]	49.3 [1.5]	59.4 [2.2]
Narrowband at spectral peak(s)	37.5 [1.0]	47.5 [1.0]	58.2 [2.3]	47.6 [1.0]	57.6 [1.0]

Note 1: Minimum received power in pilot component assumed.

Note 2: –163 dBW from IIR and IIF SVs L2 civil long (CL) component assumed.

Note 3: –157.9 dBW from IIF SVs L5 quadrature (Q5) component assumed.

Table 9.5 *J/S* Performance Comparisons with Appropriate *Q* for GPS L1C and Selected Galileo, BeiDou, and GLONASS Signals

Signal/Jammer Type	$(J/S)_{dB}$ *(dB), [Q (Dimensionless)]*			
Constellation	GPS	Galileo	BeiDou	GLONASS
Signal	L1Cp[1]	E1 OS (CBOC-)[1]	B1I (MEO)	L1OF (FDMA)
R_c *(chips/s)*	1.023×10^6	1.023×10^6	1.023×10^6	0.511×10^6
Modulation type	BOC(1,1)	BOC(1,1)	BPSK-R(2)	BPSK-R(0.511)
$(C_{Ri})_{dB}$ *(dBW)*[1]	–158.25 (III)	–160.0	–166.0	–161.0
L1 carrier (MHz)	1,575.42	1,575.42	1,561.098	1,602.0
$(C_s/N_0)_{dB}$ *(dB-Hz)*	45.54	43.79	37.79	42.79
$(C_s/N_0)_{eff,dB}$ *(dB-Hz)*	17	17	27	27
Wideband null-to-null	54.7 [5.1]	54.7 [5.1]	40.6 [2.2]	37.9]2.2]
Wideband matched spectrum	53.2 [3.6]	53.2 [3.6]	39.0 [1.5]	36.2]1.5]
Narrowband at spectral peak(s)	50.8 [2.1]	50.8 [2.1]	37.2 [1.0]	34.5 [1.0]

Note 1: Minimum received power in component is used.

for comparison is identical. Table 9.5 has two matching entries for L1C and E1 OS because both have components with matching Q's and spreading code rates and the small difference in received power does not show up in their $(J/S)_{dB}$ until the third decimal place.

Table 9.6 summarizes the receiver design parameter assumptions used for all case examples. It is further assumed that the receiver is unaided, so the tracking thresholds are determined by the third-order PLL tracking threshold assumptions made for the three case examples presented earlier. These are 27 dB-Hz for Costas carrier tracking of the BPSK modulated GPS L1 C/A and P(Y), BeiDou B1I, and GLONASS L1OF signals and 17 dB-Hz for pilot component tracking of GPS M(TDDM), L2 CL, L5 Q5, $L1C_P$ and Galileo E1 OS (CBOC-). The minimum specified received power for all pilot component signals shown in the tables are lower than the total received power, for example, 50% or 3 dB lower for GPS M(TDDM), L2 CL, L5 Q5 and Galileo E1 OS (CBOC-), but only 25% or 1.25 dB lower for GPS $L1C_P$. Clearly the dataless (pilot) PLL tracking threshold improvement more than overcomes this signal power loss plus there is additional improvement at lower $(C/N_0)_{eff,dB}$ because there is no squaring loss for coherent PLL tracking.

The receiver tracking threshold improves with external velocity aiding of the PLL, especially when the external velocity aiding is sufficiently accurate to replace the closed carrier tracking loop by making open-loop carrier Doppler estimates

Table 9.6 Summary of Assumed Receiver Design Parameters

Symbol	*Parameter*	*Value*	*Units*
$(G_{SVi})_{dB}$	Antenna gain toward SV_i	1.5	dBic
$(G_J)_{dB}$	Antenna gain toward jammer	–3.0	dBic
T_{ant}	Antenna noise temperature	100	K
$(N_f)_{dB}$	Receiver noise figure at 290K	2	dB
$T_{receiver}$	Receiver system temperature based on $(N_f)_{dB}$	169.6	K
L_{dB}	Receiver implementation loss	2	dB

during severe jamming with sufficient accuracy to keep the code tracking delay lock loop (DLL) operable. This (temporary) tracking state makes the tracking threshold dependent on the much more robust code-tracking threshold, but data demodulation is impossible when carrier tracking is open in an aided receiver. Even if carrier tracking is sustained, the bit error rate may be excessive for the lower $(C/N_0)_{eff,dB}$ levels that can be sustained with tightly coupled external aiding. For modernized signals, the pilot component tracking will be at an even lower sustained tracking level than for the traditional Costas (data) components tracking owing to squaring loss and lower jitter tolerance of the Costas PLL. This tracking mode also gains added benefits from the use of a pilot component because the code DLL filter can also be coherent with no squaring loss so long as the carrier loop is closed in PLL. However, sustained receiver operation at levels where data demodulation cannot be achieved reliably will gradually deteriorate navigation accuracy unless current navigation data can be received by another means.

The most significant improvement against wideband jammers is obtained with the use of a controlled reception pattern antenna (CRPA). The CRPA can provide a small amount of additional gain toward the satellites plus a significant amount of attenuation (gain nulls) toward a finite number of jammers ($N - 1$ jammers if the CRPA contains N antenna elements).

The above strategies (external velocity aiding and a CRPA) significantly improve (lower the value of) the receiver tracking threshold, $(C/N_0)_{eff,dB}$. Figure 9.1 illustrates the corresponding improvement in $(J/S)_{dB}$ performance as a function of receiver tracking threshold, $(C/N_0)_{eff,dB}$ for L1 C/A and the pilot components of both L2 C (L2 CL) and L5 Q5, assuming a BLWN null-to-null jammer customized to each signal. Figure 9.2 shows the same thing for L1 P(Y) and the pilot component of L1 M (TDDM). It is important in both figures to recognize that the performance difference should not be based on the assumption that all signals have the same

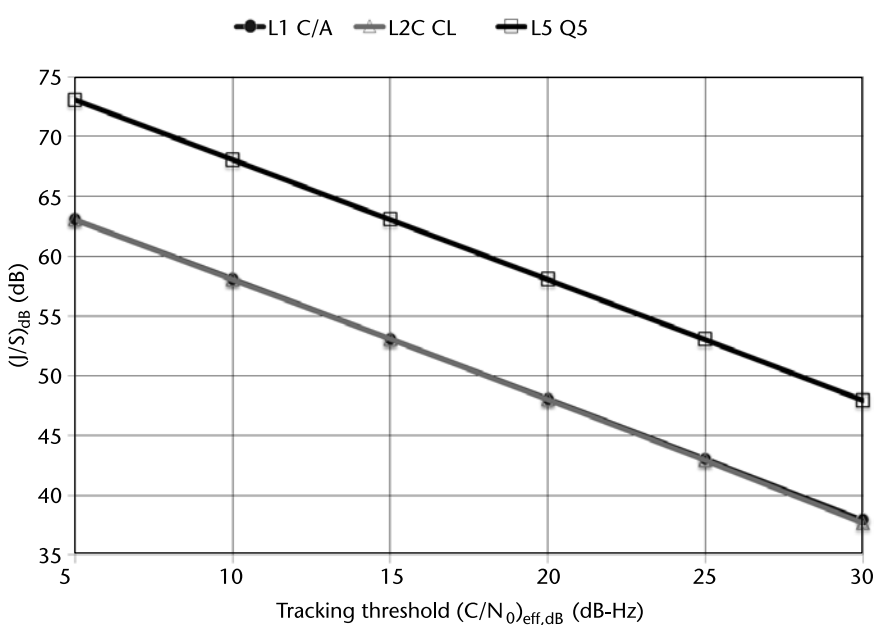


Figure 9.1 $(J/S)_{dB}$ as a function of tracking threshold for L1 C/A, L2 CL and L5 (Q5) signals.

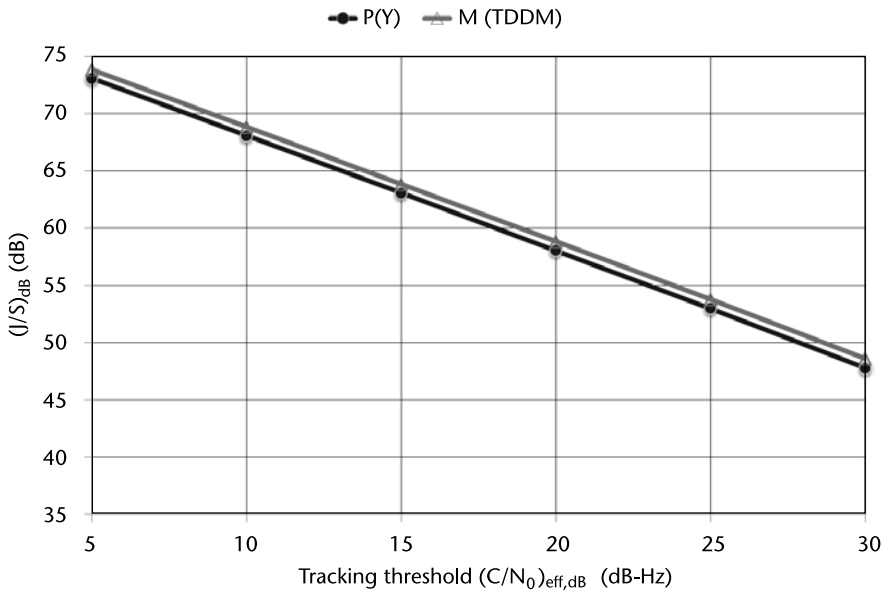


Figure 9.2 $(J/S)_{dB}$ as a function of tracking threshold for P(Y) and M (TDDM) signals.

$(C/N_0)_{eff,dB}$ tracking threshold. Keep in mind that, because of the pilot components in the modernized signals, the actual carrier and code tracking thresholds are superior to the Costas tracking of the traditional signals if the minimum received signal power is comparable.

Tolerable jamming (*tolerable* J_{dB}) is a better way than $(J/S)_{dB}$ to compare receiver jamming performance when there are multiple levels of minimum received signal power involved in the comparison. The equation for *tolerable* J_{dB} is simply

$$\textit{tolerable } J_{dB} = (J/S)_{dB} + (C_{Ri})_{dB} \quad \text{(dBW)} \tag{9.23}$$

Table 9.7 compares the receiver *tolerable* J_{dB} performance for L1 C/A, P(Y) and M code signals plus L2 CL and L5 Q5 signals for the same three types of jamming

Table 9.7 Tolerable Jamming Performance Comparisons with Appropriate *Q*

Signal/Jammer Type	*Tolerable* J_{dB} *(dBW),* [*Q (Dimensionless)*]				
PRN code	C/A	P(Y)	L1 M (TDDM)	L2 CL	L5 Q5
R_c *(chips/s)*	1.023×10^6	10.23×10^6	5.115×10^6	1.023×10^6	10.23×10^6
Modulation type	BPSK-R(1)	BPSK-R(10)	$BOC_s(10,5)$	BPSK-R(1)	BPSK-R(10)
$(C_{Ri})_{dB}$ *(dBW)*	–158.5	–161.5	–161.0	–163.0 (IIR,IIF)	–157.9 (IIF)
Wideband null-to-null	–117.5 [2.2]	–110.6 [2.2]	–97.9 [7.2]	–112.0 [2.2]	–96.8 [2.2]
Wideband matched spectrum	–119.2 [1.5]	–112.3 [1.5]	–100.4 [4.0]	–113.7 [1.5]	–98.6 [1.5]
Narrowband at spectral peak(s)	–121.0 [1.0]	–114.0 [1.0]	–102.8 [2.3]	–115.4 [1.0]	–103.3 [1.0]

(wideband, matched spectrum, and narrowband), using the values for $(J/S)_{dB}$ and $(C_{Ri})_{dB}$ from Table 9.4.

This comparison example reveals a more realistic (larger) separation between the actual threshold jammer power levels that differ more than the *J/S* metric indicates if there is more received signal power in one signal than another. For example, the M code receiver outperforms the P(Y) code receiver by 12.7, 11.9, and 11.2 dB for BLWN, matched spectrum and narrowband jammers, respectively. Likewise, L5 Q5 outperforms L1 C/A by 20.7, 20.6, and 17.7 dB as well as L2C(CL) by 15.2, 15.1, and 12.1 dB for BLWN, matched spectrum and narrowband jammers, respectively. The irony here is that the L5 pilot component is only 0.6 dB stronger than L1 C/A, but the L5 pilot tracking loop significantly outperforms the Costas L1 C/A tracking loop because the L5 pilot tracking loop has twice the PLL noise jitter tolerance plus it has zero squaring loss in the threshold area. In the case of the L5(Q5) to L2 CL comparison, the higher received signal power of L5 Q5 is solely responsible for its superior *tolerable* J_{dB}.

Figure 9.3 depicts the *tolerable* J_{dB} performance as a function of $(C/N_0)_{eff,dB}$ for L1 C/A and the pilot components of L2 CL and L5 Q5 for the third-order PLL tracking mode assuming BLWN null-to-null jammers customized to each signal. Figure 9.4 shows the same thing for L1 P(Y) and the pilot component of L1 M. In both figures, it would be incorrect to compare the *tolerable* J_{dB} performance by simply inspecting the same vertical axis intersections (i.e., using the same tracking threshold for every signal). Instead, determine the actual tracking threshold associated with each signal and then make the comparison.

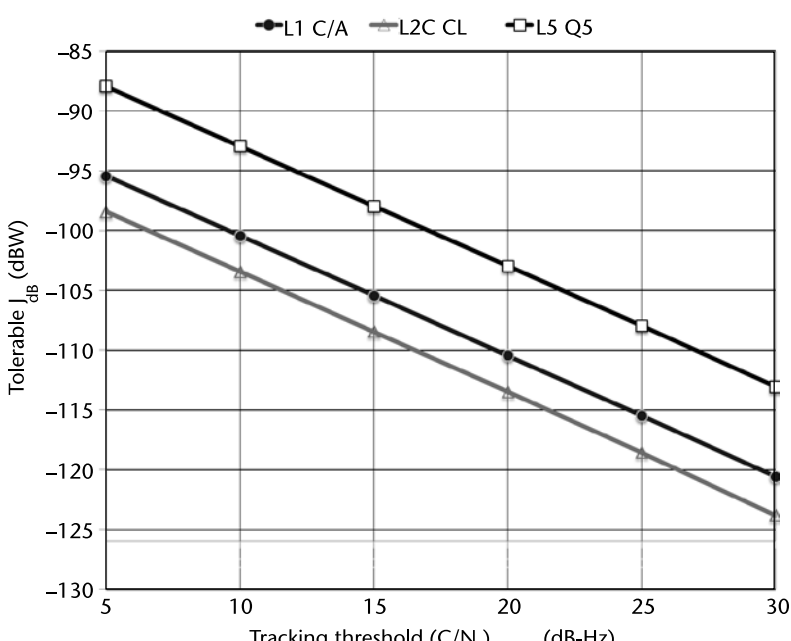


Figure 9.3 Tolerable J_{dB} as a function of tracking threshold for L1 C/A, L2 CL and L5 Q5 signals.

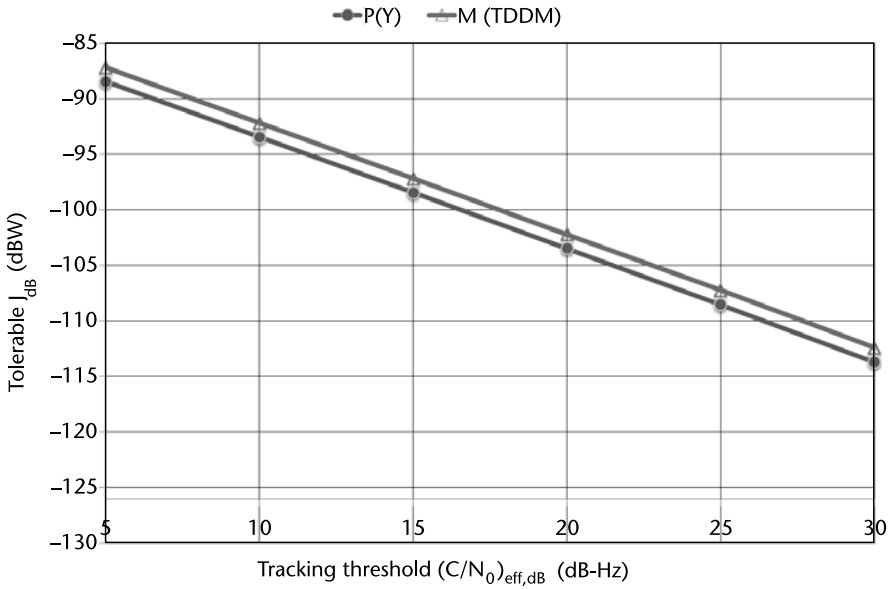


Figure 9.4 Tolerable J_{dB} as a function of tracking threshold for P(Y) and M (TDDM) signals.

9.2.2.3 Computing RF Interference Signal Levels

Even though the *J/S* performance of a GNSS receiver sounds impressive when the ratio is reported in decibels, it becomes less impressive when the actual jammer signal power at the antenna input that disrupts the receiver tracking is considered. This is because the GNSS signal power received at the antenna input is so small. To demonstrate how little jammer power is required at the input of a GNSS receiver to disable it when the receiver *J/S* performance in units of decibels has been determined, the following equation is required

$$(J/S)_{dB} = (J_r)_{dB} - (S_r)_{dB} = (J_r)_{dB} - (C_{Ri})_{dB} \tag{9.24}$$

where $(J_r)_{dB}$ = antenna input (incident) jammer power (dBW) and $(S_r)_{dB} = (C_{Ri})_{dB}$ = antenna input (incident) signal power (dBW)

Rearranging (9.24)

$$(J_r)_{dB} = (J/S)_{dB} + (C_{Ri})_{dB} = tolerable\ J_{dB} \quad (\text{dBW})$$

and since $(J_r)_{dB} = 10 \log_{10} J_r$, then the total power in watts at the antenna input at the receiver threshold jamming level is

$$J_r = 10^{\frac{(J/S)_{dB}+(C_{Ri})_{dB}}{10}} = 10^{\frac{tolerable\ J_{dB}}{10}} \ (\text{W}) \tag{9.25}$$

Using the *tolerable* $J_{dB} = -117.5$ dBW for the C/A code signal in the previous section where $(C_S/N_0)_{eff,dB} = 27$ dB-Hz, $(C_{Ri})_{dB} = -158.5$ dBW and $(J/S)_{dB} = 41$ dB

for a band-limited white noise (BLWN) null-to-null jammer, the incident jammer power is determined from (9.25) as follows:

$$J_r = 10^{\frac{-117.5}{10}} = 1.76 \times 10^{-12} \quad \text{W}$$

So this demonstrates that for the FRPA design assumed in the previous section, less than 2 pW of incident BLWN interference power is required to disable a C/A code signal receiver with a moderate $(J/S)_{dB}$ performance of 41 dB.

Table 9.8 depicts the extremely small disabling incident power using the *tolerable J* performance examples from Table 9.7 (including the associated *J/S* performance examples from Table 9.4 for easy comparison). Table 9.8 clearly illustrates how much more robust are the modernized signals than original signals in the presence of jamming. By inspection of the Table 9.8 entries, the small powers at the antenna input that disable the unaided GNSS receiver provides sobering insight into unaided GNSS receiver antijam performance. Recall that in all of these case examples it was assumed that the antenna gain pattern provided an additional 3-dB loss to the jammer power because it was assumed that it was arriving at a lower antenna elevation angle entry point than the GNSS signals.

9.2.2.4 Computing Range to RF Interference

Usually, the receiver operating range from the source of the RF interference is desired given the effective isotropic radiated power (EIRP) of the interference source. (Refer to Appendix C for greater insight into free-space propagation loss.)

The formula for the link budget for the transmitted jammer power up to the antenna input is given by

$$(EIRP)_{dB} = (J_r)_{dB} + (L_p)_{dB} \tag{9.26}$$

where

Table 9.8 Disabling Incident Power with Corresponding $(J/S)_{dB}$ Performance

Signal/Jammer Type	*Disabling incident power (pW), [$(J/S)_{dB}$ (dB)]*				
PRN code	C/A	P(Y)	L1 M (TDDM)	L2 CL	L5 Q5
R_c *(chips/s)*	1.023×10^6	10.23×10^6	5.115×10^6	1.023×10^6	10.23×10^6
Modulation type	BPSK-R(1)	BPSK-R(10)	$BOC_s(10,5)$	BPSK-R(1)	BPSK-R(10)
$(C_{Ri})_{dB}$ *(dBW)*	–158.5	–161.5	–161.0	–163.0 (IIR,IIF)	–157.9 (IIF)
Wideband null-to-null	1.8 [41.0]	8.8 [50.9]	164.1 [63.2]	6.4 [51.0]	205.0 [61.1]
Wideband matched spectrum	1.2 [39.3]	5.9 [49.2]	91.2 [60.6]	4.3 [49.3]	139.8 [59.4]
Narrowband at spectral peak(s)	0.8 [37.5]	4.0 [47.5]	52.4 [58.2]	2.9 [47.6]	93.2 [57.6]

$(EIRP)_{dB}$ = Effective isotropic radiated power of the jammer
$= (J_t)_{dB} + (G_t)_{dB}$

$(J_t)_{dB}$ = jammer transmit power into its antenna (dBW)
$= 10 \log_{10} J_t$ (J_t expressed in W)

$(G_t)_{dB}$ = jammer transmitter antenna gain (dBic)

$(J_t)_{dB}$ = incident (received) jammer power (dBW)
$= 10 \log_{10} J_r$ (J_r expressed in W)
$=$ *tolerable* J_{dB} for computing tolerable range to jammer

$(L_p)_{dB}$ = jammer power propagation loss (dB)

This link budget does not include what happens after the jammer signal arrives at the receiver antenna input. That is taken care of by the equation that computes $(J/S)_{dB}$ for the receiver.

If the jammer propagation path is air-to-air, air-to-ground or ground-to-air, then the free-space propagation loss equation can be used as a good approximation, so

$$\left(L_p\right)_{dB} = 10 \log_{10}\left(\frac{4\pi d}{\lambda_j}\right)^2 \qquad (9.27) \text{ (see Appendix C)}$$

where d = range to jammer (m) and λ_j = wavelength of jammer frequency (m). If the jammer propagation path is ground-to-ground, then modeling the jammer propagation loss is considerably more complex. Several ground-to-ground models for this case will be described later.

Assuming an essentially free-space jammer propagation path, a case example computation is provided for L1 C/A code signal and a BLWN null-to-null jammer. Assume that the jammer transmitter power $J_t = 2$W, so $(J_t)_{dB} = 10 \log_{10} 2 = 3.0$ dBW, and the jammer antenna is RHCP with gain $(G_t)_{dB} = 3$ dBic. Then $(EIRP_j)_{dB} = 6$ dBW and the effective isotropic radiated power is $EIRP_t = 10^{0.6} = 4.0$W. Since the jammer frequency is in-band, the jammer carrier wavelength, λ_j, will be assumed to be centered at L1. It is further assumed that the jammer carrier frequency is modulated by a white noise signal, then band-limited to about 2 MHz to become a null-to-null BLWN jammer. Using $(J_r)_{dB} =$ *tolerable* $J = -117.5$ dBW from Table 9.7 for the null-to-null BLWN case example of the L1 C/A signal, the line-of-sight range to the antenna at which the case example receiver reaches its loss of track threshold can now be determined from (9.28) rearranged to solve for the propagation loss as follows:

$$(L_p)_{dB} = (J_t)_{dB} + (G_t)_{dB} - (J_r)_{dB} = (EIRP)_{dB} - \textit{Tolerable } J \quad \text{(dB)}$$
$$= 6 + 117.5 = 123.5 \text{ dB}$$

Next, solve the free-space propagation equation for the range, *d,* as follows:

$$(L_p)_{dB} = (EIRP)_{dB} - Tolerable\ J = 10\log_{10}\left(\frac{4\pi d}{\lambda_j}\right)^2 = 20\log_{10}\left(\frac{4\pi d}{\lambda_j}\right) \quad \text{(dB)}$$

$$d = \frac{\lambda_j 10^{\frac{(EIRP)_{dB} - Tolerable\ J}{20}}}{4\pi} \quad \text{(m)}$$

This distance equation is the free-space range from the receiver antenna input to the jammer transmitter antenna output. This is the distance required to attenuate the jammer's power level to the power level corresponding to the tracking threshold level of the receiver case example. In this case example, it computes the free-space range between the jammer antenna and the receiver antenna that attenuates the jammer power, 6 dBW (4W), by 123.5 dB so that the arrival power level at the L1 C/A code signal receiver is its *tolerable J* level, –117.5 dBW (1.8 pW). Computing this range in kilometers by dividing the previous equation by 1,000 and then converting this range to nautical miles, obtains

$$d = \frac{\lambda_j 10^{(\frac{(EIRP)_{dB} - Tolerable\ J}{20}}}{4000\pi} = \frac{0.1903 \cdot 10^{\frac{123.5}{20}}}{12566.377} = 22.7 \text{ km } (12.2 \text{ nmi})$$

Table 9.9 illustrates the *tolerable* J_{dB} distance to the jammer for all case examples of Table 9.7, including all three types of jammers, assuming each jammer signal is RHCP and has an $(EIRP)_{dB}$ of 6 dBW (4W) of power. Note that the narrowband jammer is the most effective (lowest Q) for a given power, but it is also the easiest (least expensive antenna and receiver design) to mitigate.

Figure 9.5 depicts the free-space range to a wideband null-to-null jammer in kilometers as a function of EIRP in watts for the case examples of L1 C/A, L2 CL and L5 Q5 signals. Figure 9.6 shows the same for the case examples of P(Y) and M (TDDM) signals.

For ground-to-ground jammer paths the jammer signal experiences considerably more path loss than for free-space owing to varying ground attenuation effects. The variables are so unpredictable that exact path loss prediction (without experimental data in the actual ground area of operation) is virtually impossible.

Table 9.9 Tolerable J Distance to 4-W Jammer, Assuming Free-Space Propagation

Signal/Jammer Type	*Distance, km (nmi)*				
PRN code	C/A	P(Y)	L1 M (TDDM)	L2 CL	L5 Q5
R_c *(chips/s)*	1.023×10^6	10.23×10^6	5.115×10^6	1.023×10^6	10.23×10^6
Modulation type	BPSK-R(1)	BPSK-R(10)	$BOC_s(10,5)$	BPSK-R(1)	BPSK-R(10)
$(C_{Ri})_{dB}$ *(dBW)*	–158.5	–161.5	–161.0	–163.0 (IIR,IIF)	–157.9 (IIF)
Wideband null-to-null	22.7 (12.2)	10.2 (5.5)	2.4 (1.3)	15.4 (8.3)	2.8 (1.5)
Wideband matched spectrum	27.6 (14.9)	12.4 (6.7)	3.2 (1.7)	18.7 (10.1)	3.4 (1.8)
Narrowband at spectral peak(s)	33.8 (18.2)	15.2 (8.2)	4.2 (2.3)	17.8 (9.6)	3.1 (1.7)

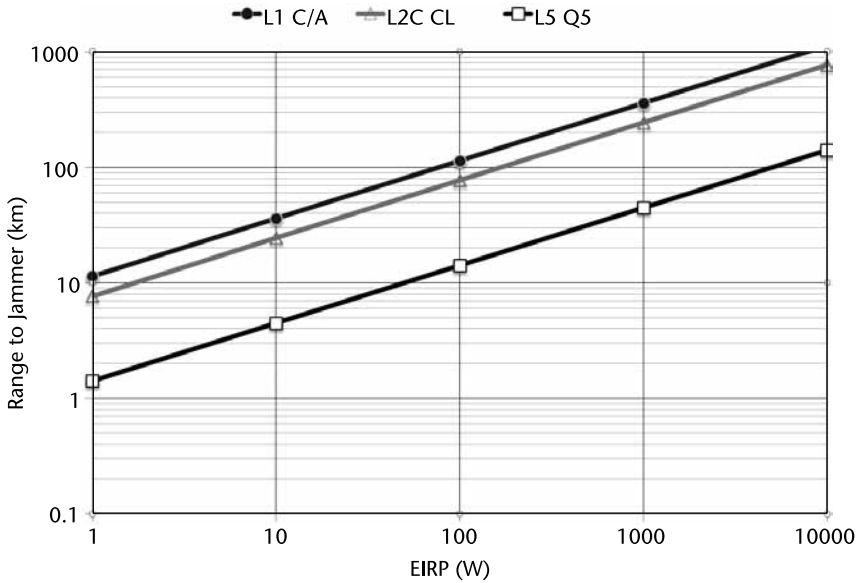


Figure 9.5 Free-space range to wideband null-to-null jammer as a function of EIRP for L1 C/A, L2 CL, and L5 Q5 signals.

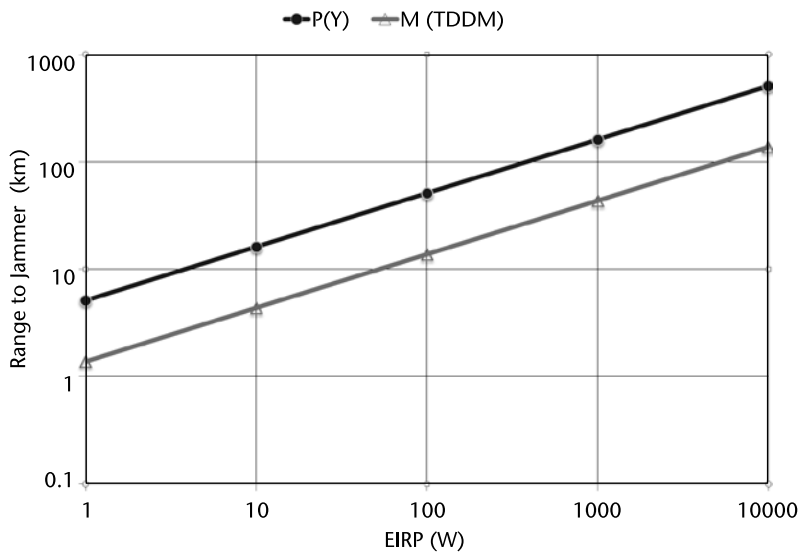


Figure 9.6 Free-space range to wideband null-to-null jammer as a function of EIRP for P(Y) and M (TDDM) signals.

However, an empirical model can provide a proxy for estimating typical ground-to-ground path loss. Okumura and Hata [7] published their combined results for use in mobile communications. Okamura obtained extensive mobile communications data from controlled field experiments in urban, small to medium city, large city, suburban and open country environments. The parameters of his extensive data included: (1) frequencies in the 150 to 1,500 MHz; (2) distances of 1 to 20 km; (3) base station antenna heights ranging from 30 to 200m; and (4) mobile antenna heights ranging from 1 to 10m. Hata used this data to develop an empirical

model of propagation loss for an urban environment of the form $(L)_{dB} = A + B \log_{10}(d[km])$, then added corrections to the urban model for the remaining environments. Later, Mogensen et al. [8] modified the Hata urban model. The Modified Hata urban model increases the path loss based on experimental observations in the 1,500 to 2,000-MHz frequency band, but no changes were made to the correction models. Keeping in mind the parameter ranges used to obtain these empirical models, the Modified Hata model is more suitable for path loss predictions in the L1 and higher frequency L-bands, while the original Hata model remains more suitable for L2 and lower frequency L-bands. The original Hata path loss equation in decibels for an urban area (most suitable for L2 and lower frequencies) is

$$\begin{aligned}(L_{p2})_{dB} &= 69.55 + 26.16 \cdot \log_{10}(f[\text{MHz}]) - 13.82 \cdot \log_{10}(h_{base}[m]) \\ &- a(h_{mobile}[m]) + \left[44.9 - 6.55 \cdot \log_{10}(h_{base}[m])\right] \cdot \log_{10}(d[km])\end{aligned} \tag{9.28}$$

The modified Hata path loss equation for an urban area most suitable for L1 and higher frequencies, modifies only the first two terms of the Hata model as follows

$$\begin{aligned}(L_{p1})_{dB} &= 46.3 + 33.9 \cdot \log_{10}(f[\text{MHz}]) - 13.82 \cdot \log_{10}(h_{base}[m]) \\ &- a(h_{mobile}[m]) + \left[44.9 - 6.55 \cdot \log_{10}(h_{base}[m])\right] \cdot \log_{10}(d[km])\end{aligned} \tag{9.29}$$

where

f [MHz] = transmission frequency (MHz);
h_{base} [m] = height of the base (transmitter) antenna (m);
d [km] = distance between base and mobile antennas (km);
$a(h_{mobile}[\text{m}])$= correction for height of the mobile antenna (dB).

There are two mobile antenna height correction equations. The correction for mobile antenna height for a small to medium city is

$$\begin{aligned}a(h_{mobile}[m]) = &\left[1.1 \cdot \log_{10}(f[MHz]) - 0.7\right] \cdot h_{mobile}[m] + \\ &0.8 - 1.56 \cdot \log_{10}(f[MHz]) \quad (\text{dB})\end{aligned} \tag{9.30}$$

and the correction for mobile antenna height for a large city for f [MHz] ≥ 400 MHz is

$$a(h_{mobile}[m]) = 3.2 \cdot \left(\log_{10} 11.75 \cdot h_{mobile}[m]\right)^2 - 4.97 \ (\text{dB}) \tag{9.31}$$

where h_{mobile}[m] = mobile antenna height (m).

Hata also provided correction terms for all other operating environments. Mogensen et al. did not change any of these correction terms. For a suburban area, L_{pN}, is corrected as

$$(L_{sN})_{dB} = (L_{pN})_{dB} - 2[\log_{10}(f[MHz]/28)]^2 - 5.4 \text{ (dB)} \tag{9.32}$$

where $N = 1$ or 2 as appropriate for f[MHz], and for an open area, L_{pN} is corrected as

$$\begin{aligned}(L_{oN})_{dB} = (L_{pN})_{dB} &- 4.78 \cdot (\log_{10} f[MHz])^2 + \\ &18.33 \cdot \log_{10} f[MHz] - 40.94 \text{ (dB)}\end{aligned} \tag{9.33}$$

As a computational example for the L1 C/A code signal using (9.29), the Modified Hata equation will be used assuming *tolerable J* = –117.5 dBW (determined earlier), jammer power $(EIRP)_{\text{dB}} = 6$ dB, that is, the same 4W BLWN jammer used for the free-space range to jammer computation example, mobile antenna height of 1.5m, and base antenna height of 30m. Further assumed is the urban area of a small to medium size city, so the correction equation for antenna height (9.30) is used. The problem is solved piecewise, beginning with (9.30), as follows

$$\begin{aligned}
&A = 46.3 + 33.9 \cdot \log_{10}(1575.42) = 154.692 \text{ (dB)} \\
&B = 13.82 \cdot \log_{10}(30) = 20.414 \text{ (dB)} \\
&a(h_{mobile}[m]) = \left[1.1 \cdot \log_{10}(1575.42) - 0.7\right] \cdot 1.5 \\
&\qquad + 0.8 - 1.56 \cdot \log_{10}(1575.42) = \ 0.038 \text{ (dB)} \\
&D \cdot \log_{10}(d[km]) = \left[44.9 - 6.55 \cdot \log_{10}(30)\right] \cdot \log_{10}(d[km]) = 35.225 \cdot \log_{10}(d[km]) \\
&(L_{p1})_{dB} = (EIRP)_{dB} - tolerable\ J = 6 + 117.5 = 123.5 \text{ (dB)} \\
&(L_{p1})_{dB} = 123.5 = A - B - a(h_{mobile}[m]) + D \cdot \log_{10}(d[km]) \text{ (dB)} \\
&d[km] = 10^{\frac{123.5 - A + B + a(h_{mobile}[m])}{D}} = 10^{\frac{123.5 - 154.692 + 20.414 + 0.038}{35.225}} = 0.497 \text{ (km)}
\end{aligned}$$

Note for this case example the significant increase in the over ground jammer signal attenuation resulting in a much shorter range of 0.50 km (0.27 nmi) as compared to the free-space jammer range of 22.7 km (12.2 nmi).

Table 9.10 summarizes the range to jammer distances for all five case examples for this over ground example assuming three different types of jamming.

Figure 9.7 depicts the over ground range in kilometers to a wideband null-to-null jammer as a function of EIRP in watts for the case examples of L1 C/A, L2 CL and L5 Q5 signals. Figure 9.8 shows the same for the case examples of P(Y) and M (TDDM) signals. Note the significant improvement in range to jammer for the over ground case due to attenuation of the jammer power by the ground.

Sklar [9] used the following equation if the path loss characteristic changes en route

$$L = L_0 + n \cdot 10 \log_{10}\left(\frac{d}{d_0}\right) + X_G \text{ (dB)} \tag{9.34}$$

where

Table 9.10 *Tolerable J* Distance to 4-W Jammer, Assuming over Ground Propagation in an Urban Area of a Small to Medium City

Signal/Jammer Type	*Distance, km (nmi)*				
PRN code	C/A	P(Y)	L1 M (TDDM)	L2 CL	L5 Q5
R_c *(chips/s)*	1.023×10^6	10.23×10^6	5.115×10^6	1.023×10^6	10.23×10^6
Modulation type	BPSK-R(1)	BPSK-R(10)	$BOC_s(10,5)$	BPSK-R(1)	BPSK-R(10)
$(C_{Ri})_{dB}$ *(dBW)*	–158.5	–161.5	–161.0	–163.0 (IIR,IIF)	–157.9 (IIF)
Wideband null-to-null	0.496 (0.268)	0.315 (0.170)	0.137 (0.074)	0.458 (0.247)	0.176 (0.095)
Wideband matched spectrum	0.554 (0.299)	0.352 (0.190)	0.162 (0.087)	0.511 (0.276)	0.197 (0.106)
Narrowband at spectral peak(s)	0.622 (0.336)	0.395 (0.213)	0.190 (0.102)	0.574 (0.310)	0.221(0.119)

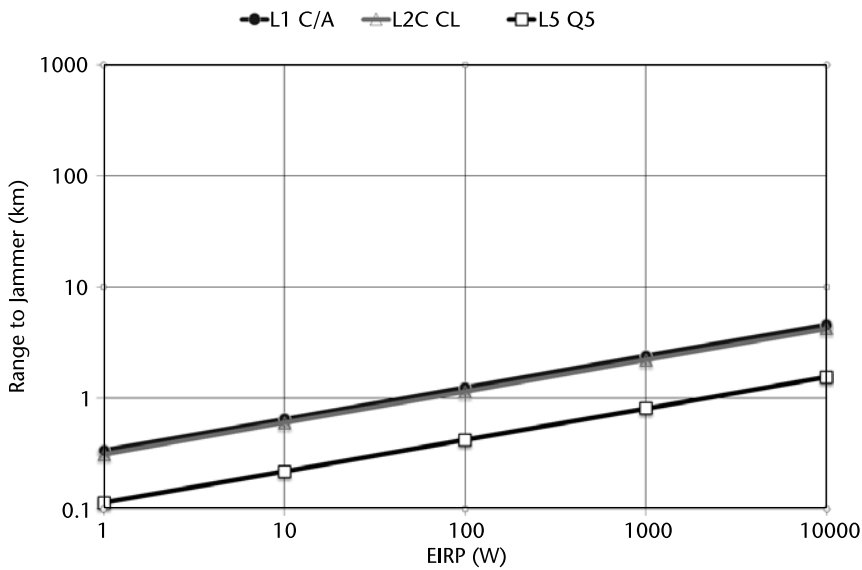


Figure 9.7 Over ground range to wideband null-to-null jammer as a function of EIRP in small to medium city urban area for L1 C/A, L2 CL, and L5 Q5 signals.

L_0 = path loss up to d_0 (dB);

n = path loss factor (typically 2 or higher) beyond d_0 (dimensionless);

$\frac{d}{d_0}$ = ratio of d (distance beyond d_0) to d_0 where both are in meters (dimensionless);

X_G = constant to account for other known (e.g., system) losses (dB).

Sklar also used a popular air-to-air path loss model that is basically the same equation as the free space model but modifies the square loss exponent using a larger value than 2.

9.2.2.5 Vulnerability of C/A Code Signal to CW Interference

The range-to-jammer case examples in the previous section assumed that the quality factor Q holds up uniformly for CW as well as for other (wider) narrowband

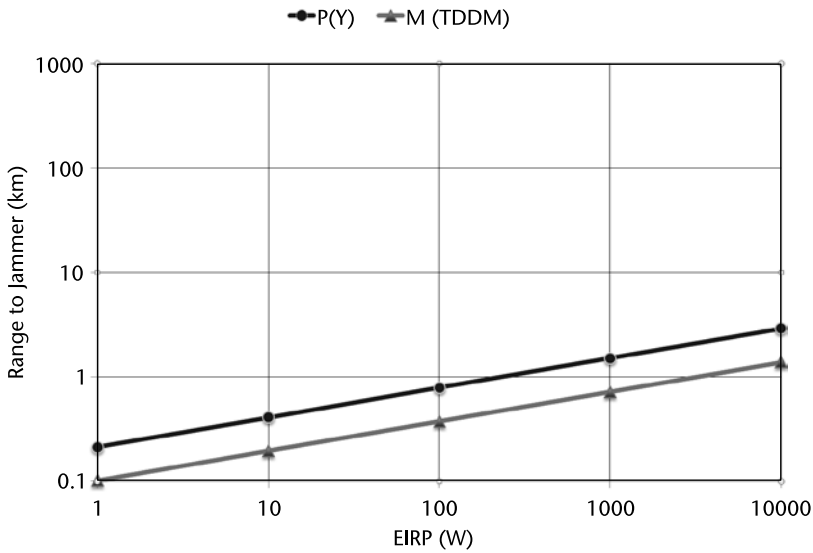


Figure 9.8 Over ground range to wideband null-to-null jammer as a function of EIRP in small to medium city urban area for P(Y) and M (TDDM) signals.

interference. In particular, the GPS L1 C/A code signal is more vulnerable to the line spectrum of CW interference than to wider narrowband interference. This is due to the fact that GPS C/A spreading code is a Gold code with a short 1-ms period (i.e., the PRN sequence repeats every 1 ms). Therefore, the C/A code signal (neglecting the navigation data) has a line spectrum with lines that are 1 kHz apart [10]. The adoption of a secondary short synchronization code, called the Neuman-Hofman code, is one among many GNSS signal synthesis innovations that plays an important role on spectral separation, bit synchronization, and narrowband interference protection at the receiver end [11].

Although it is typical for each line in the C/A code signal power spectrum to be –24 dB or more negative with respect to the total power, there are some lines in every C/A code signal that provide less attenuation (i.e., stronger than –24 dB). The C/A code signal line spectrum characteristic is inferior to a maximum length PRN sequence with the same number of shift register bits [12]. As a result, a CW jammer can mix with a strong C/A code signal line and leak through the correlator. Table 9.11 summarizes the worst line frequency and the worst-line (strongest) amplitude for all 37 original PRNs of the GPS C/A code signal [13]. The modernized GPS satellites, QZSS satellites, and satellite-based augmentation system GEO satellites have many more C/A code PRN numbers. This table is shown to demonstrate typical levels of worst lines and the fact that they occur at various line numbers. It should be recognized that even the weaker C/A code signal spectral lines are not very robust against any CW jamming that aligns with any line. This leak-through problem (CW momentarily matching a C/A code signal spectral line) causes a transient increase in CW effectiveness in comparison with the norm. These transient CW phenomena usually cause more of a problem during C/A code signal search and acquisition modes than during tracking modes.

If the receiver has a CW jammer detector, this can provide a warning that special (time-consuming) search measures must be taken such as increasing the search

Table 9.11 Worst Line for Each of the 37 Original GPS C/A Codes

C/A Code PRN Number	*Worst-Line Frequency (kHz)*	*Worst-Line Amplitude (dB)*	*C/A Code PRN Number*	*Worst-Line Frequency (kHz)*	*Worst-Line Amplitude (dB)*
1	42	–22.71	20	30	–22.78
2	263	–23.12	21	55	–23.51
3	108	–22.04	22	12	–22.12
4	122	–22.98	23	127	–23.08
5	23	–21.53	24	123	–21.26
6	227	–21.29	25	151	–23.78
7	78	–23.27	26	102	–23.06
8	66	–21.5	27	132	–21.68
9	173	–22.09	28	203	–21.73
10	16	–22.45	29	176	–22.22
11	123	–22.64	30	63	–22.14
12	199	–22.08	31	72	–23.13
13	214	–23.52	32	74	–23.58
14	120	–22.01	33	82	–21.82
15	69	–21.9	34	55	–24.13
16	154	–22.58	35	43	–21.71
17	138	–22.5	36	23	–22.23
18	183	–21.4	37	55	–24.13
19	211	–21.77	—	—	—

dwell time and adjusting the search detector parameters for best C/A code signal search operation in the presence of CW. The modernized signals with lower chipping rates such as L2C have design features that minimize this vulnerability. The line spectra of higher chipping rate signals such as L5, P(Y), and M signals have lines each having much lower power, and spaced more closely, so that they essentially take on attributes of continuous spectrum, so these signals do not exhibit this vulnerability.

Even if an adaptive antenna array or temporal filtering are used to reduce CW interference to the thermal noise level, there remains a vulnerability of C/A code signal to CW interference. The thermal noise floor can be determined from the following equation

$$\left(N_t\right)_{dB} = \left(N_0\right)_{dB} + 10\log_{10}\left(B_{fe}\right) \text{ dBW} \tag{9.35}$$

where B_{fe} = receiver front end bandwidth (Hz).

Assume that the C/A code signal receiver is a narrow correlator design with a 15-MHz bandwidth. Substituting the thermal noise density, $(N_0)_{dB}$, value from the example in Section 9.2.2.2 into (9.36) yields

$$\left(N_t\right)_{dB} = -204.3 + 71.76 = -132.5 \text{ dBW}$$

If an adaptive antenna array or temporal filter takes the CW interference down to this thermal noise floor, then $(J_r)_{dB} = (N_t)_{dB}$. Substituting this into (9.26) and using the minimum received L1 C/A code signal received power $(S_r)_{dB} = -158.5$ dBW gives

$$(J/S)_{dB} = (J_r)_{dB} - (S_r)_{dB} = (N_t)_{dB} - (C_{Ri})_{dB} = -132.5 - (-158.5) = 26.0 \text{ dB}$$

This would not be a problem for most unaided C/A code signal receiver designs if the source were wideband noise RF interference or even narrowband RF interference if the bandwidth were, say, 10 kHz or wider. However, CW interference at this level could cause problems with the C/A code signal receiver because of the leak-through phenomena described earlier. For example, compare $(J/S)_{dB} = 26$ dB with the worst-case leak-through levels shown in Table 9.11. If the C/A code signal receiver were a standard correlator design, then $B_{fe} = 1.7$ MHz and $(J/S)_{dB}$ decreases to 16.5 dB. Obviously, increasing the receiver front-end bandwidth increases the intrinsic vulnerability of C/A code signals to CW interference.

A C/A code (Gold code) signal jammer can also be a problem for this same situation because temporal side lobes are produced. In both cases, the problem is more serious during C/A code signal search and acquisition modes than for tracking modes.

9.2.2.6 Effects of RF Interference on Code Tracking

The effect of RF interference on code tracking is different from its effect on signal acquisition, carrier tracking, and data demodulation. While the latter three functions depend on the output signal-to-noise-plus-interference ratio (SNIR) at the output of a prompt correlator, as described in Section 9.2.2, code tracking relies on the difference between an early correlator and a late correlator, as described in Section 8.7.

The interference considered here is modeled as Gaussian and zero-mean, but not necessarily having a white (flat) spectrum. The analysis assumes that the receiver front end does not saturate or respond nonlinearly in some other way to the interference, as discussed in Section 8.3, and that there is no multipath, so that code tracking errors are caused by noise and interference. While the effects of white noise on code tracking error are considered in Section 8.7, this section evaluates the effect of nonwhite interference that produces additional random, zero-mean, code tracking error. The effect of interference is quantified in terms of the standard deviation of the code tracking error.

As described in Section 8.7, there are many different designs for discriminators and tracking loops, and interference may have different effects on each. However, a lower bound on the code tracking error has been developed that is independent of code tracking circuit design, yet is a tight bound in the sense that it provides reasonably accurate predictions of code tracking performance for well-designed tracking circuits. This lower bound (in units of seconds) is given by [14]

$$\sigma_{\mathrm{LB}} \cong \frac{\sqrt{B_n}}{2\pi\sqrt{\int_{-\beta_r/2}^{\beta_r/2} f^2 \left[\frac{S_s(f)}{\left(\frac{C_s}{N_0}\right)^{-1} + \frac{C_t}{C_s}S_t(f)}\right] df}} \tag{9.36}$$

where the code-tracking loop has a (one-sided) equivalent rectangular bandwidth of B_n Hz that is much smaller than the reciprocal of the correlation integration time, the power spectral density of white noise and any spectrally flat interference is N_0 W/Hz, and the nonwhite component of the interference has power spectral density $C_t S_t(f)$ W/Hz, with normalized power spectral density $\int_{-\infty}^{\infty} S_t(f)\,df = 1$, and interference power over infinite bandwidth of C_t W (the aggregate interference carrier power and power spectral density may result from the aggregation of multiple interfering signals). The signal component being tracked has power spectral density $S_S(f)$ normalized to unit power over infinite bandwidth, $\int_{-\infty}^{\infty} S_s(f)\,df = 1$, and C_s is the recovered desired signal power, also defined over an infinite bandwidth, so that the signal has a carrier power to noise density ratio of C_s/N_0 Hz, in white noise. The ratio of interference power to signal power is C_t/C_s It is assumed that the power spectral densities are symmetric about $f = 0$. Precorrelation filtering in the receiver is approximated by an ideal filter with linear phase and rectangular passband having total bandwidth β_r Hz.

Now consider a code-tracking loop whose discriminator uses coherent early-late processing, where the carrier phase of the reference signal tracks that of the received signal, so that the in-phase or real outputs of early and late correlations drive the discriminator, with early-to-late spacing of D spreading code periods. Using the same notation and assumptions as in (9.36), the standard deviation (in units of seconds) for the coherent early-late processing (CELP) in interference is [14]

$$\begin{aligned}\sigma_{\mathrm{CELP}} &\cong \frac{\sqrt{B_n}}{2\pi\int_{-\beta_r/2}^{\beta_r/2} f S_s(f)\sin(\pi f D T_c)\,df}\sqrt{\int_{-\beta_r/2}^{\beta_r/2}\left[\left(\frac{C_s}{N_0}\right)^{-1} + \frac{C_t}{C_s}S_t(f)\right]S_s(f)\sin^2(\pi f D T_c)\,df} \\ &= \frac{\sqrt{B_n}}{2\pi\int_{-\beta_r/2}^{\beta_r/2} f S_s(f)\sin(\pi f D T_c)\,df} \\ &\quad\times\sqrt{\left(\frac{C_s}{N_0}\right)^{-1}\int_{-\beta_r/2}^{\beta_r/2} S_s(f)\sin^2(\pi f D T_c)\,df + \frac{C_t}{C_s}\int_{-\beta_r/2}^{\beta_r/2} S_t(f)S_s(f)\sin^2(\pi f D T_c)\,df}\end{aligned} \tag{9.37}$$

The second line in (9.37) shows that the code tracking error is the root-sum-squared of a term that only involves the signal in white noise, and a term that

involves the spectra of the interference and the desired signal, scaled by the ratio of interference power to signal power.

In the limit as D becomes vanishingly small (in practice, how small D needs to be depends upon the specific spectra of signal and interference; examination of the Taylor series expansions shows that the criterion $DT_c\beta_r << \frac{2\sqrt{3}}{\pi} \cong 1.1$ is sufficient but not always necessary), the trigonometric expressions in (9.37) can be replaced by Taylor Series expansions around $D = 0$, and (9.37) becomes

$$\sigma_{\text{CELP},D\to 0} \cong \frac{\sqrt{B_n}}{2\pi\beta_s}\left[\left(\frac{C_s}{N_0}\right)^{-1} + \frac{C_\iota}{C_s}\frac{\chi_{\iota s}}{\beta_s^{\,2}}\right]^{1/2} \tag{9.38}$$

where

$$\beta_s = \sqrt{\int_{-\beta_r/2}^{\beta_r/2} f^2 S_s(f)\,df} \tag{9.39}$$

is the RMS bandwidth of the signal computed over the precorrelation bandwidth and $\chi_{\iota s}$ is the code tracking spectral separation coefficient (SSC) defined by

$$\chi_{\iota s} = \int_{-\beta_r/2}^{\beta_r/2} f^2 S_\iota(f) S_s(f)\,df \tag{9.40}$$

which includes a frequency-squared weighting in the integral that is not found in the SSC used for correlator output SNR defined in (9.8).

The expression (9.38) shows that neither the output SNIR nor merely the RMS bandwidth of the modulation is sufficient to describe code-tracking accuracy for CELP; instead the quantity $\frac{C_\iota}{C_s}\frac{\chi_{\iota s}}{\beta_s^{\,2}}$ is needed. When this quantity is small, CELP with small early-late spacing approaches the lower bound on code-tracking error.

The interference spectrum affects code-tracking accuracy in a fundamentally different way from the way it affects effective C/N_0. The frequency-squared weighting inside the integral in (9.40) indicates that interference power away from the center frequency can have much greater effect on code-tracking accuracy than on effective C/N_0, which has no such frequency-squared weighting.

In many applications, early-late processing uses the power difference between early and late taps, rather than relying on phase locked loop (PLL) tracking in the carrier tracking loop to support coherent delay locked loop (DLL) processing in the code tracking loop. The code tracking error for the resulting noncoherent early-late processing (NELP) is [14]

$$\sigma_{\text{NELP}} \cong \sigma_{\text{CELP}} \sqrt{1 + \frac{\int_{-\beta_r/2}^{\beta_r/2} S_s(f)\cos^2(\pi f D T_c)\,df}{T\frac{C_s}{N_0}\left(\int_{-\beta_r/2}^{\beta_r/2} S_s(f)\cos(\pi f D T_c)\,df\right)^2} + \frac{\int_{-\beta_r/2}^{\beta_r/2} S_t(f) S_s(f)\cos^2(\pi f D T_c)\,df}{T\frac{C_s}{C_t}\left(\int_{-\beta_r/2}^{\beta_r/2} S_s(f)\cos(\pi f D T_c)\,df\right)^2}} \tag{9.41}$$

that reveals the same behavior of NELP that is well-known for infinite front-end bandwidth and white noise—the standard deviation of NELP code tracking error is the product of the standard deviation of CELP code tracking error and a squaring loss that is greater than unity, but approaches unity as the signal power increases relative to both the white noise level and the interference power.

In the limit as D becomes vanishingly small, the trigonometric expressions in (9.41) can be replaced by Taylor series expansions around zero, and (9.41) becomes

$$\begin{aligned}\sigma_{\text{NELP},D\to 0} &\cong \sigma_{\text{CELP},D\to 0}\left[1 + \frac{\int_{-\beta_r/2}^{\beta_r/2} S_t(f) S_s(f)\,df}{TC_s\left(\int_{-\beta_r/2}^{\beta_r/2} S_s(f)\,df\right)^2}\right]^{1/2} \\ &= \sigma_{\text{CELP},D\to 0}\left[1 + \frac{1}{T\frac{C_s}{N_0}\eta} + \frac{\kappa_{ts}}{T\frac{C_s}{C_t}\eta^2}\right]^{1/2}\end{aligned} \tag{9.42}$$

where η is the fraction of signal power passed by the precorrelation bandwidth,

$$\eta = \int_{-\beta_r/2}^{\beta_r/2} S_s(f)\,df, \tag{9.43}$$

and κ_{ts} is the SSC describing the effect of interference on correlator output SNR, defined in (9.8).

Clearly, quantifying the effect of interference on code tracking accuracy is different and more complicated than evaluating its effect on signal acquisition, carrier tracking, and data demodulation. Not only does the effect depend on the spectra of signal and interference and on the precorrelation filter, but also on details of the discriminator design and the bandwidth of the code tracking loop.

As an example, consider narrowband interference centered at $\pm f_t$, whose spectrum is modeled as $S_t(f) = 0.5[\delta(f + f_t) + \delta(f - f_t)]$, where $\delta(\cdot)$ is the Dirac function having infinite amplitude, vanishing width, and unit area. Substituting for this interference power spectral density in the code tracking SSC (9.40), assuming the interference is within the precorrelation bandwidth, yields

$$\chi_{ts} = f_t^2 S_s(f_t) \tag{9.44}$$

The lower bound on code tracking accuracy with narrowband interference is obtained by substituting the interference spectrum into (9.36), yielding

$$\sigma_{\mathrm{LB}} \cong \frac{\sqrt{B_n}}{2\pi\sqrt{\dfrac{C_s}{N_0}\displaystyle\int_{-\beta_r/2}^{\beta_r/2} f^2 S_s(f)\,df}} = \frac{1}{2\pi\beta_s}\sqrt{\frac{B_n}{C_s/N_0}}, \tag{9.45}$$

This result shows that optimal code tracking in narrowband interference produces the same code tracking error as with no narrowband interference. It is readily shown that this processing is closely approximated by narrowband excision followed by CELP with very small early-late correlator spacing.

When narrowband excision is not employed and NELP is used, the effect of narrowband interference is obtained using (9.42) and (9.38), assuming small early-late spacing,

$$\sigma_{\mathrm{NELP},D\to 0} \cong \frac{\sqrt{B_n}}{2\pi\beta_s}\sqrt{\left[\left(\frac{C_s}{N_0}\right)^{-1} + \frac{C_t}{C_s}\frac{f_t^2 S_s(f_t)}{\beta_s^2}\right]\left[1 + \frac{1}{T\dfrac{C_s}{N_0}\eta} + \frac{S_s(f_t)}{T\dfrac{C_s}{C_t}\eta^2}\right]}. \tag{9.46}$$

Figure 9.9 plots (9.45) and (9.46) for four different modulations, calculated with B_n of 0.1 Hz, $(C_s/N_0)_{dB}$ of 30 dB-Hz, $(C_t/C_s)_{dB}$ of 40 dB, precorrelation bandwidth of 24 MHz, correlation integration time of 20 ms, and very small early-late spacing. The results for NELP approach the lower bound for certain interference frequencies. Interference ve-ry near band center degrades NELP code tracking accuracy less than interference further away from band center. The oscillatory behavior of the NELP error for BPSK-R(1) and BOC(1,1) demonstrates that narrowband interference away from band center can have the same effect on code tracking error as interference nearer to band center, reflecting the frequency-squared weighting in (9.40). The result for NELP BPSK-R(10) shows that the maximum error occurs when the narrowband interference is placed half way between the spectral peak at band center, and the first spectral null at 10.23 MHz.

9.2.3 Interference Mitigation

In addition to optimizing the designs of the unaided (stand-alone) GNSS receiver for robust performance in the presences of RFI, the following specific interference types and mitigation techniques should be recognized and the use of mitigation countermeasures should be considered: (1) adjacent-band interference from ground-based transmitters [3] that can only be mitigated by extreme front-end stopband filtering techniques; (2) colocated transmitter harmonic interference requiring signal blanking during transmissions; (3) narrowband interference that can be mitigated by

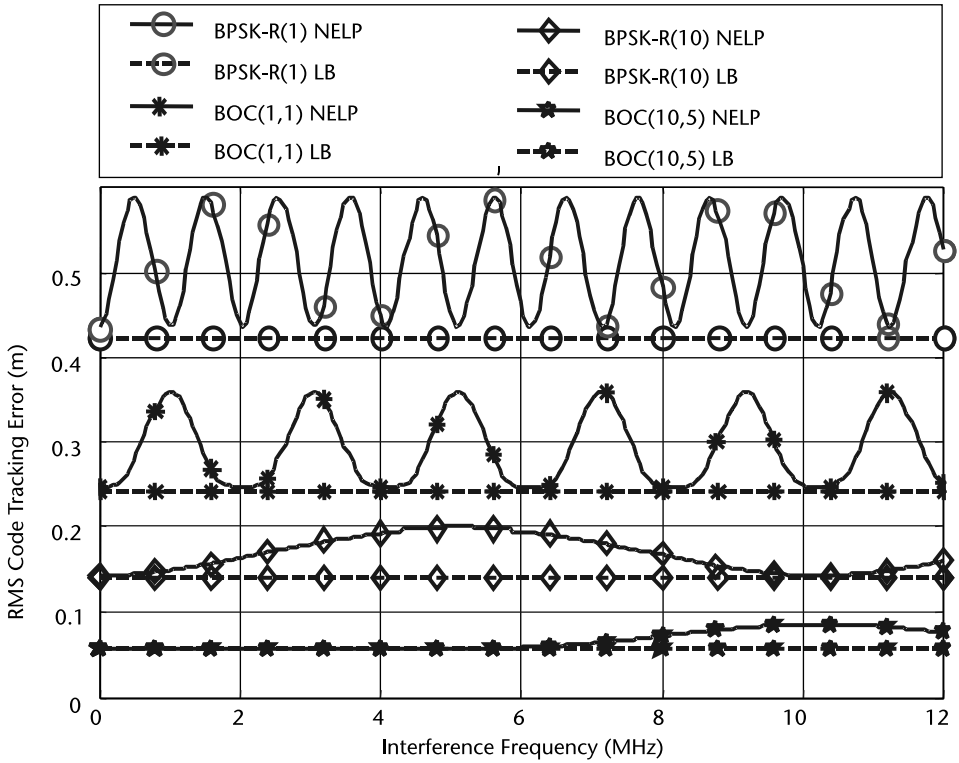


Figure 9.9 Noncoherent early-late processing (NELP) and lower bound (LB) code tracking error of different modulations in narrowband interference, for different frequency interference, with 0 MHz corresponding to band center, and 12 MHz corresponding to the edge of the band.

various signal processing techniques prior to the carrier wipe-off stage as described in Section 9.2.3.1; (4) wideband pulse interference (typically from radars and DME transmitters) that can be mitigated by autonomous front-end fast blanking [15] and recovery techniques (see Section 9.2.3.1); and (5) wideband matched spectrum or Gaussian noise interference that can be mitigated using a CRPA (in place of the FRPA) to steer nulls toward jammers and gain toward SVs (see Section 9.2.3.2).

In addition to these controlled design features, operate-through augmentations are sometimes required. For examples, an IMU that is intrinsically impervious to RFI can continue to navigate, with quadratic position accuracy deterioration (drift) as a function of time, and a chip-scale-atomic clock (CSAC) that can continue to maintain time with some time accuracy deterioration as a function of time. Such operate-through features are essential if total loss of PVT in the presence of excessive RFI can have catastrophic consequences. The use of eLoran (a two-dimensional ground-based PVT system) as an operate-through system has found considerable GNSS synergism for some applications because the low frequency of eLoran makes it impossible to build a compact RFI antenna plus the much stronger received power of eLoran makes it intrinsically more robust to RFI than GNSS. Unfortunately, eLoran is not a global navigation system, but is regaining popularity in many parts of the world because of its low cost, robustness to RFI and remarkable accuracy [16].

There are natural barriers to GNSS RFI that limit its effectiveness. RFI can only have full effect on a GNSS receiver if it is in the line of sight of the receiver antenna and unobstructed. Of course, there is the possibility of RFI being reflected into the receiver antenna, thereby taking a nondirect route. The most universal example of natural RFI blockage is the Earth's curvature that blocks most low-elevation sources of RFI more than 50 km (27 nmi) away. That is why intentional RFI (jamming) utilizes high elevation for longer-range effectiveness. One example of intentional use of a natural barrier is the military strategy to operate with the handheld receiver antenna below ground level in a foxhole that permits visibility of the SVs overhead but effectively masks line-of-sight ground-based jammers. An example of a coincidental natural barrier is an aviation receiver with the antenna located on top of the aircraft. The aircraft body provides some masking of ground-based RFI and the gain pattern of the antenna rolls off significantly below the aircraft horizon. But this is not a significant barrier against strong ground-based jammers.

9.2.3.1 Mitigating Narrowband and Pulse RF Interference

Since the Q factors (in Table 9.3) for narrowband interference for all GNSS signals are always much smaller than for wider band interference, this is the most lethal form of interference (i.e., more disabling capability for the same amount of EIRP from the transmitter). Fortunately, narrowband interference is also the easiest to mitigate because it is observable by the receiver when it is above the thermal noise level and not harmful to modernized GNSS signals when it is at or below the thermal noise level. Reference [17] described and evaluated modern types of narrowband suppression techniques for GPS receivers: overlapped FFT (OFFT), filter bank (FB), an extension of OFFT to further reduce the weighting loss of the OFFT, and adaptive transversal filter (ATF). The OFFT has the fastest response time, so the most likely modernized narrowband suppression will be some form of FFT technique.

Various hardware-based techniques have been used in the past and these are worthy of further discussion. One of the least complex hardware-based techniques uses a nonlinear analog-to-digital converter (ADC) along with digital automatic gain control (AGC) to observe (but not suppress) narrowband RFI. This nonlinear ADC technique was originally developed for communications applications by Amoroso as reported in [18, 19]. The technique was first adapted for GPS signal use by Scott [20]. Scott's nonlinear ADC adaptation along with his innovation of digital AGC to provide interference situational awareness was included in [21, 22]. Scott's nonlinear ADC technique for the GPS L1 C/A signal (equally applicable to other GNSS signals) utilizes the fact that the nature of CW interference allows a substantial portion of the desired signal to be correlated at and near the peaks of the CW signal and it is relatively simple to detect the presence of CW in that signal.

Another hardware-based technique uses a sophisticated application specific integrated circuit (ASIC) implementation of an ATF that required a 12-bit ADC at the low IF output of the receiver. That ASIC was provided as government furnished equipment (GFE) to military GPS manufacturers during the development and manufacturing era of the SAASM (Selective Availability/Anti-Spoofing Module). The

GFE ASIC design provided more than 70 dB of narrowband interference suppression. The transversal filter detects the presence of any narrowband energy that exists above the thermal noise level and suppresses that energy down to the thermal noise level. That process also suppresses the signal in those frequency regions, but this loss of energy has only a small order effect on the receiver's tracking performance because only a small percentage of the total signal spectrum is suppressed. As described in [17], the ATF technique has the least insertion loss but the longest response time, so the OFFT (or some suitable form of FFT) technique should be considered for modernized applications.

Modernized approaches to narrowband RFI mitigation use digital signal processing in the receiver channels prior to the carrier wipe-off stage are described in more detail in Section 8.3. As explained in Section 9.2.2.5, L1 C/A signal receivers can still experience acquisition problems with CW interference even if suppressed to the thermal noise level due to the strong spectral lines of the C/A signals.

Pulsed interference can be easily mitigated by instant recovery analog design techniques by preventing front-end gain compression and saturation with fast attack, fast recovery AGC design (consistent with AGC stability criteria). *Pulse blanking*, or the zeroing of the received signal when pulsed interference is detected, is a particularly effective mitigation technique [15]. The receiver cannot correlate with the signals during these bursts, but the duty cycle of most burst jammers is usually so low that correlations take place most of the time, unless gain compression or saturation is permitted to take place in the front end that results in slow recovery back to linear operation. Thus, a well-designed receiver front end renders the overall receiver immune to most burst jammers (e.g., a pulse jammer with 50% duty cycle blanks out half the received signal power that degrades the $(C_s/N_0)_{dB}$ by 3 dB, but the duty cycle is usually much smaller). It is relatively inexpensive in terms of cost or size, weight and power to build-in pulse jamming mitigation features in a receiver, but most commercial receivers do not have such protection.

9.2.3.2 Mitigating Gaussian and Spectrum Matching Wideband Interference

Encrypted GNSS signals such as GPS P(Y) and M, Galileo E1 PRS, or BeiDou B1Q signals provide no intrinsic advantage against enemy jamming owing to encryption. The encryption is there to prevent spoofing, thereby ensuring signal integrity and data authentication, by denying access to unauthorized users. The two most difficult military jamming threats to mitigate are generally referred to as bandlimited white noise (BLWN) and matched spectrum jammers. Specifically, the BLWN jammer threat generates Gaussian noise whose bandwidth is designed to span the null-to-null signal spectrum of the target signal. The matched-spectrum jammer threat generates the same spectrum characteristics of the target signal spectrum. Encryption does not offer any protection against the jammer's ability to perfectly match the encrypted signal power spectrum. For these two types of wideband jammers, there are only three mitigation techniques, given in the order of increasing mitigation capability and cost: (1) receiver tracking threshold enhancements; (2) external velocity aiding, historically from an inertial measurement unit (IMU); and (3) antenna directional gain control, historically from a controlled reception pattern antenna (CRPA) [23] that steers deep nulls toward the jammers while providing some gain toward the SVs. All three techniques should be used synergistically to

achieve maximum jamming robustness. These have been described in more detail in Chapter 8, but some additional insight is presented here on the latter two in this wideband interference mitigation context.

Because of improved performance under dynamic stress, there has also been increased commercial use of IMU aiding on dynamic platforms. Because the likelihood of RFI threats has increased, there are also commercial CRPAs available for applications where there is concern that such threats might be encountered. In general, the commercial capabilities are less sophisticated than their military counterparts.

Historically, IMU aiding for military applications has been of two basic types: (1) loosely coupled aiding that permits the weak link of (unaided) carrier tracking to be opened while the IMU aiding provides an estimate of the line-of-sight Doppler to each SV with sufficient accuracy that the code tracking loop can continue to operate, thereby providing the added tracking robustness of the code loop operating with minimal dynamic stress; and (2) tightly coupled aiding into the closed carrier tracking loop that sustains the carrier tracking loop because dynamic stress has been largely removed by the external aiding. The latter provides the highest navigation accuracy, including the opportunity for the common satellite/IMU navigation system to provide on-going calibration of the IMU, as well as retaining the potential for data demodulation (because the carrier tracking loop is closed). Increased jamming causes this mode to give way to the loosely coupled mode (open carrier tracking loop) that can still prevent the IMU from drifting, but not good enough for calibration. This transition happens when the jamming level exceeds the improved carrier-tracking threshold under minimal dynamic stress. When the jamming level exceeds the improved code tracking threshold under minimal dynamic stress, navigation reverts to free-inertial mode (IMU drifting), but this is vastly superior to losing navigation altogether as in the unaided case. The transitions are determined not by the observed loss of the current operating mode (that corrupts the measurements), but by the observation and precise measurement of the jamming level and a priori knowledge of when the transition must be made to a more robust mode. Tightly coupled IMU aiding is significantly more difficult to properly implement, but has demonstrated significant accuracy payoff for smart weapons applications where the weapon is approaching jammer threats but sustains the IMU accuracy longer. This is because the less accurate loosely coupled operation time becomes shorter before the receiver is forced into free-inertial mode. A more robust GNSS receiver/IMU coupling technique is *ultratight* (sometimes referred to as *deeply integrated*) and is discussed in detail in Section 13.2.8.3. Sophisticated IMU aiding along with superior receiver tracking enhancements have achieved up to 70 dB of J/S performance, but lessons have been learned that this falls short of the jamming robustness mark for many military operations. The military solution to the required wideband jamming robustness improvement to date has been the use of a CRPA that replaces the usual FRPA to provide antenna directional null-steering of –30 to –45 dB toward the jammers and gain-steering of 1.5 to 5.0 dB toward the SVs. This can improve the total J/S performance by up to 120 dB. The most advanced integrated technology that synergistically combines both IMU and CRPA techniques is called space-time adaptive processing (STAP) and space-frequency adaptive processing (SFAP), the former involving time domain signal processing and the latter involving frequency domain signal processing. It is possible to steer

nulls toward the jammers without assistance from the receiver, but to steer gain toward the SVs requires direction-cosines (or the equivalent). This, in turn, requires an attitude and heading reference system for the CRPA, historically from an IMU if the platform is not stationary.

The typical military CRPA uses 7 elements and can therefore steer up to 6 nulls toward enemy jammers. With so few antenna elements, STAP or SFAP cannot provide much beam steering (gain) toward each SV. The IMU information can be used to minimize the adverse effects of the jammer null-steering process. Obviously, if the jammer direction is colocated with the SV direction, that SV will be lost, but it is much better to lose that SV than all of the SVs as would be the case with a FRPA. Also, the most advanced CRPA signal processing design will first remove all narrowband jamming energy by frequency excision techniques from each antenna element so as not to lose wideband nulls to narrowband jammers.

There are also low-cost antenna null-steering techniques that depend on the jammers to be ground located (or some a priori known location). One example is the analog cancellation technique that senses the jamming energy from a bow-tie antenna element with a sector antenna gain coverage at and slightly above the horizon and then subtracts this energy from the FRPA portion of the GPS antenna (with overlapping gain coverage) using analog RF techniques. Another technique assumes and senses nonpolarized jammer energy and removes it from the right-hand circularly polarized GPS signals. All of these low-cost CRPA techniques expect that the enemy will cooperate with the a priori restrictions of the antenna design. Airborne jammers are often more vulnerable and difficult to sustain over long periods of time, but it should be apparent that these are by far the most effective jammers.

9.3 Ionospheric Scintillation

Irregularities in the ionospheric layer of the Earth's atmosphere can at times lead to rapid fading in received signal power levels [24–26]. This phenomenon, referred to as ionospheric scintillation, can lead to a receiver being unable to track one or more visible satellites for short periods of time. This section describes the causes of ionospheric scintillation, characterizes the fading associated with scintillation, details the effects of scintillation upon the performance of a GNSS receiver, and lastly describes mitigations.

9.3.1 Underlying Physics

The ionosphere is a region of the Earth's atmosphere from roughly 50 km up to several Earth radii where incident solar radiation separates a small fraction of the normally neutral constituents into positively charged ions and free electrons. The maximum density of free electrons occurs at an altitude of around 350 km above the surface of the Earth in the daytime. Most of the time, the principal effect of the presence of free electrons in the ionosphere is to impart a delay on the signals (see Section 10.2.4.1). However, irregularities in the electron density occasionally arise that cause constructive and destructive interference among each signal. Such irregularities are most common and severe after sunset in the equatorial region (within

±20° from the geomagnetic equator). High-latitude regions also experience scintillation, which is generally less severe than in the equatorial region, but may persist for long periods of time. Scintillation is also more common and severe during the peak of the 11-year solar cycle.

9.3.2 Amplitude Fading and Phase Perturbations

In the absence of scintillation, a simplified model for one particular signal as seen by a receiver is:

$$r(t) = \sqrt{2P}s(t)\cos(\omega t + \phi) + n(t) \tag{9.47}$$

where P is the received signal power, ω is the carrier frequency (in radians/s), $s(t)$ is the normalized transmitted signal, and $n(t)$ is noise.

Scintillation causes a perturbation to both the received signal amplitude and phase, and the received signal in the presence of scintillation may be modeled as [27]:

$$r(t) = \sqrt{2P \bullet \delta P} \bullet s(t)\cos(\omega t + \phi + \delta\phi) + n(t) \tag{9.48}$$

where $\sqrt{\delta P}$ is a positive, unitless parameter that characterizes amplitude fading due to scintillation, and $\delta\phi$ is a parameter with units of radians that represents phase variations due to scintillation. The power fluctuation, δP, is generally modeled as following a Nakagami-m probability density function given by:

$$p(\delta P) = \frac{m^m \delta P^{m-1}}{\Gamma(m)} e^{-m\delta P}, \delta P \geq 0 \tag{9.49}$$

with mean value of one and variance of $1/m$. The strength of amplitude fading due to scintillation is characterized using a parameter referred to as the S_4 *index*, which is equal to the standard deviation of the power variation δP:

$$S_4 = \sqrt{\frac{1}{m}} \tag{9.50}$$

Due to the properties of the Nakagami-m distribution, the S_4 index cannot exceed $\sqrt{2}$.

Power fluctuations are highly correlated over short time intervals. Measured power spectral densities of scintillation-induced power fluctuations fall off with increasing frequency with a level proportional to f^{-p} with p in the range of 2.5 to 5.5 [24]. The spectral density of the power fluctuations also tends to fall off at extremely low frequencies (below around 0.1 Hz). Figure 9.10 shows simulated receiver power fluctuations due to strong scintillation ($S_4 = 0.9$).

Phase variations due to scintillation are most commonly modeled as following a zero-mean Gaussian distribution:

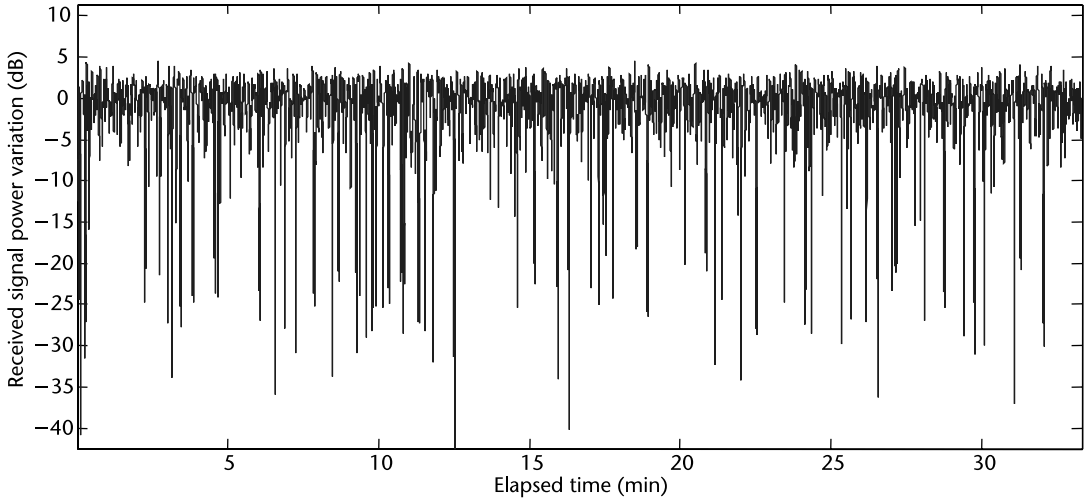


Figure 9.10 Simulated effects of strong scintillation ($S_4 = 0.9$) on received signal level.

$$p(\delta\phi) = \frac{1}{\sqrt{2\pi}\sigma_\phi} e^{-\frac{\delta\phi^2}{2\sigma_\phi{}^2}} \tag{9.51}$$

with standard deviation σ_ϕ. Phase variations are highly correlated over short periods of time, with observed power spectral densities approximately following the form Tf^{-p} with p in the range of 2.0 to 3.0 [24] and where T is a strength parameter (in units of rad^2/Hz).

9.3.3 Receiver Impacts

Scintillation can lead to intermittent GNSS receiver signal tracking outages in two different ways. First, if an amplitude fade is of sufficient depth and time duration, from a receiver perspective, the desired signal is absent and loss of lock of the code and carrier phase tracking loops is inevitable. If the desired signal is being received at a very high level, such as 50 dB-Hz, a 20-dB fade is generally tolerable, but a much deeper fade will typically cause an outage if the fade persists longer than the time constant of the tracking loops. At low signal-to-noise ratios, even a 5–10-dB fade can cause a disruption in tracking. Second, strong phase scintillation can cause loss of phase lock within the receiver if the phase variations introduce a level of dynamics that is greater than the phase lock loop can accommodate (see discussion in Section 8.9.

Fortunately, scintillation rarely occurs on all visible satellites simultaneously. The irregularities that cause scintillation are not generally present within the ionosphere in the vicinity of each of the points where the signals from the visible satellites intersect the ionosphere. Thus, scintillation tends to only impact one or at most a few satellites simultaneously.

Both S_4 index and phase standard deviation are a function of carrier frequency:

$$S_4 \propto \frac{1}{f^{1.5}} \tag{9.52}$$

$$\sigma_\phi \propto \frac{1}{f} \tag{9.53}$$

so that, for instance, when fading due to ionospheric scintillation occurs the observed S_4 index for a signal on GPS L2 is approximately 1.45 times greater than the S_4 index for a GPS L1 signal and observed σ_ϕ for GPS L2 is approximately 1.28 times greater than for GPS L1. The implication of this carrier frequency dependency is that scintillation would be much more likely to cause outages for GNSS signals in lower L-band (1,164–1,300 MHz), than for GNSS signals in upper L-band (1,559–1,610 MHz), provided that the signals were at similar power levels and with similar design features (e.g., with or without pilot components).

9.3.4 Mitigation

Many of the techniques discussed in Section 9.2.3.2 to mitigate wideband interference can be also used to successfully mitigate the impacts of ionospheric scintillation amplitude fading. These include: (1) receiver tracking threshold enhancements, (2) external velocity aiding, for example, from an IMU [28], and (3) beam-forming antennas. Reduction of carrier phase tracking loop bandwidths can improve performance when scintillation results in deep amplitude fading. However, at times when fades are not deep but phase variations are significant due to ionospheric scintillation, the opposite approach of increasing carrier loop bandwidths can be helpful. An adaptive carrier loop design to mitigate ionospheric scintillation impacts is presented in [29]. Cross-aiding of carrier tracking loops amongst multiple frequency signals broadcast from each satellite [30] can also be beneficial since it is somewhat rare for deep fades to happen simultaneously on multiple frequencies [31, 32].

Receivers capable of tracking most or all of the GNSS satellites as opposed to just a subset of them are more robust against ionospheric scintillation [33] since, as noted in Section 9.3.3, scintillation is rarely seen in all directions above the local horizon simultaneously.

9.4 Signal Blockage

Signal blockage, also known as shadowing, occurs when propagating electromagnetic waves encounter physical objects in the direct path between transmit antenna and receive antenna. In some cases, such as if only a few leaves are in the path, the effect may be negligible. In other cases, such as when a large building is between the receiver and the satellite, the electromagnetic waves can be absorbed or reflected, introducing such large attenuation that this direct path signal is unusable by even the most sensitive GNSS receiver.

GNSS is being used in an increasingly wide variety of situations, as signals are becoming more robust, and receivers are becoming more sensitive. Consequently, it is increasingly common for situations to occur where blockage of some type occurs, and yet a usable, even if degraded, signal is received.

Since actual results in specific situations can vary significantly from those predicted by available models, it is important to employ significant amounts of margin in assessing performance. If one is hoping for shadowing of interference, it may be important to consider situations where vegetation is sparse or dry, terrain is flat, and structures are dry lumber or have ample windows; in these cases, blockage of interference may be modest. Conversely, if one is seeking signal reception under blockage conditions, it may be prudent to plan for situations of dense and wet vegetation, significant hills or mountains, and structures having walls and ceilings of thick concrete and steel; in these cases, blockage of desired signals may be significant.

This section provides guidelines for assessing the effects of signal blockage on GNSS signals or sources of interference to a GNSS receiver. Section 9.4.1 considers the effects of vegetation, while Section 9.4.2 addresses terrain effects. Finally, Section 9.4.3 discusses man-made structures. It is important to recognize that actual effects are highly dependent on specific conditions, so the models and results provided here are only nominal, and not exact.

9.4.1 Vegetation

Vegetation, which typically consists of a combination of woody branches and trunks along with foliage, or leaves, can cause a combination of refraction and diffraction. The result can be three types of effects on received signals:

- Delay spread, as small-scale multipaths (see Section 9.5);
- Multiple angles of arrival, as the waves refract around various dense structures in the vegetation;
- Attenuation, where the signal undergoes excess attenuation due to absorption and reflection of energy by the vegetation.

All of these effects vary with type of vegetation, humidity, whether foliage is present or not, and whether there is moisture on the vegetation from rain or dew.

In addition, these effects are time varying when the transmitter or receiver (or both) are moving, or when wind-induced motion of vegetation occurs. In particular, there can be deep amplitude fades and significant carrier phase fluctuations over time scales of a fraction of a second, when either the transmitter or the receiver is moving, and there is vegetation in the path.

An excellent compilation of data and models concerning vegetation effects on propagation is provided in [34]. It reports delay spreads of 10 ns or less. For attenuation considerations [34] suggests separately addressing horizontal path propagation (when transmitter and receiver are both below the height of vegetation) and slant path propagation (with transmitter or receiver above the height of vegetation). References [34] and Chapters 2 and 3 of [35] provide more detail on the following discussion, albeit there appears to be an absence of data on angles of arrival

effects. This type of effect may depend upon the type of vegetation as well as the proximity of the receive antenna to the vegetation.

Empirical models have been developed, where a parametric form is assumed and extensive data is taken and then analyzed to determine model parameter values that best fit the data. A commonly used empirical model for horizontal path loss in decibels is of the form:

$$L = Af^B d^C \text{ (dB)} \tag{9.54}$$

where A, B, and C are parameters typically determined by fitting data to this parametric form. Here, f is the frequency and d is the distance, each with units that vary with different models.

The various models summarized in [34] show very different values for these parameters. Some show distinctly different values for A depending upon whether trees are in-leaf or out-of-leaf, whereas even the sign of B is different in different models, with some models suggesting increasing loss with higher frequency, while others claiming the opposite. Results in [35] suggest that, at L-band, attenuation in decibels due to vegetation with foliage is between 24% and 35% larger than attenuation due to the same vegetation with no foliage.

The frequency dependence coefficient tends to be small over the range of frequencies used for GNSS signals, with the largest value of B reported as 0.5. Thus, given the uncertainties in the conditions and models, modeling excess losses due to vegetation as constant over L-band frequencies used for GNSS signals appears to be within the margin of modeling error.

Numerical results in [35] are provided for (9.54) with $C = 1$ and Af^B treated as a single numerical value (the loss per meter). The results indicate that the losses per meter are much greater for one or several trees, than for tens or hundreds of meters of vegetation. For long horizontal paths with distances of 100 km or greater and intermittent (not continuously dense) vegetation over the path, an attenuation coefficient of 0.3 dB/m is recommended. Over horizontal paths with distances greater than 100 m, this simple model predicts greater vegetation loss than does the ITU-R model in [34], which is of the form (9.54) with $A = 0.2$, $B = 0.3$, and $C = 0.6$ for f in megahertz and d in meters, with $d < 400$ m.

Slant path propagation involves shorter paths through vegetation—typically fewer than five trees. Reference [35] provided results for propagation through a small number of trees, indicating that attenuation coefficients vary with different types of trees, with values at L-band ranging from 0.7 dB/m to 2.0 dB/m, and an average over tree types of 1.3 dB/m. The total L-band attenuation through single trees of different types varies from 3.5 dB for poplar to 20.1 dB for white spruce, and an average over different tree types of 11.0 dB. Other data indicate that L-band attenuation for single trees tends to range between 10 dB and 20 dB, with attenuation coefficients typically between 1.0 dB/m and 2.0 dB/m.

The attenuation tends to decrease with higher elevation angles to the transmitter. A variety of results in [35] indicate that the attenuation for elevation angles between 15° and 45° is between 16 dB and 12 dB, with the larger value for full foliage and the smaller value for bare trees.

The vegetation losses described in this section are typically added (in decibels) to those for no vegetation losses.

9.4.2 Terrain

Terrain is typically considered impervious to propagation of electromagnetic waves at L-band. When terrain blocks the path between transmitter and receiver, any signal energy at the receiver arrives because the waves bend, or diffract, over the terrain.

Propagation losses due to terrain vary considerably with specific path geometry and terrain profile. Physics-based propagation models account for terrain effects by calculating propagation losses over a specific path with known terrain profile. These models employ first-principles physics representations of interactions between the electromagnetic waves and the terrain, typically not accounting for other impediments to propagation such as vegetation and buildings. One such widely used model is the Terrain-Integrated Rough-Earth Model (TIREM) [36]. TIREM employs analytical models of knife-edge refraction to predict the effects of propagation over terrain. Computing terrain propagation losses using TIREM requires access to a terrain model and the TIREM software, yielding results for propagation over a specific path and terrain profile.

An alternative for more generally assessing propagation over terrain is to employ empirical propagation models, primarily developed for commercial wireless communications. These models are developed using extensive data sets of measured propagation loss. Parametric models are then fit to this data. These models are less accurate than TIREM for propagation over a specific terrain profile, but provide generally representative results over a variety of propagation paths and conditions. They implicitly include effects of building and vegetation, to the extent these were present when the data was taken, as well as terrain effects. Several such empirical models were discussed in [37, 38]; two specific models are addressed in more detail here.

The *Erceg model* described in [39] provides an estimate of propagation loss of the form

$$L = A + 10\gamma \log_{10}\left(d / d_0\right) + s \text{ (dB)} \tag{9.55}$$

where

- A is the free space propagation loss from the source to a reference point $d_0 = 100$ m, $A = 20\log_{10}(4\pi f d_0/c)$, where f is in hertz and c is the speed of light in a vacuum;
- γ is the propagation loss coefficient, modeled as a Gaussian random variable having mean $\mu_\gamma = a - bh_t + c/h_t$ with h_t the height of the transmit antenna, and standard deviation σ_γ that varies for different transmitter and receiver locations;
- s is the shadow fading component that varies for different transmitter and receiver locations, having mean μ_s and standard deviation σ_s

Numerical values for these parameters are based on data obtained at 1,900 MHz in suburban environments, so generally apply for L-band within several decibels, which is typically within the margin of error for empirical models. The receive antenna height is always 2m, perhaps an optimistic value for handheld GNSS

receivers. Transmit antenna heights range from 10m to 80m, with separations from the receive antenna from 0.1 to 8 km. Application of the Erceg model to GNSS signal blockage is thus limited to situations where the user is a couple of meters above the ground in a suburban environment and the satellite is visible below around 40° elevation angle with a maximum signal path length through the suburban blockage environment of around 8 km.

The formulation in (9.55) has a number of attractive aspects. It is a generalization of free-space propagation, for which $\gamma = 2$ and s is identically zero, as well as generalizing the two-ray flat reflective earth model [40] for which $A = 0$, $\gamma = 4$, and s is identically zero. Using the mean values for γ and s in (9.55) provides typical values for propagation loss, using the mean values plus or minus two- or three-sigma values for these parameters yields extreme values for propagation losses that account for fading.

Table 9.12 lists the parameter values given in [39]. With a low transmit antenna height of 10m, the mean propagation exponent exceeds 5.5 for all of the terrain categories examined, and the mean shadowing loss exceeds 8 dB.

As an alternative, one of the most widely used empirical propagation models is known as the COST-231 Hata model, a refinement by the Coopération Européenne dans le domaine de la recherché Scientifique et technique (COST)-231 group of a model originally developed by Hata and then updated [38]. The COST-231 Hata model for propagation loss in decibels is

$$\begin{aligned} L &= 46.3 + 33.9\log_{10}(f) - 13.82\log_{10}(h_t) - R \\ &+ \left[44.9 - 6.55\log_{10}(h_r)\right]\log_{10}(d) + E \quad \text{(dB)} \end{aligned} \tag{9.56}$$

where

- f is the frequency in megahertz, with $500 < f < 2000$;
- h_t is the transmitter height above the Earth in meters, with $30 < h_t < 200$;
- h_r is the receiver height above the Earth in meters, $1 < h_r < 10$;
- d is the range between transmitter and receiver in kilometers, $1 < d < 10$,

Table 9.12 Parameter Values for the Erceg Propagation Loss Model

	Terrain Category		
Model Parameter	*Hilly with Moderate-to-Heavy Tree Density*	*Hilly with Light Tree Density or Flat with Moderate-to-Heavy Tree Density*	*Flat with Light Tree Density*
a	4.6	4.0	3.6
b (m^{-1})	0.0075	0.0065	0.0050
c (m)	12.6	17.1	20.0
σ_γ	0.57	0.75	0.59
μ_s	10.6	9.6	8.2
σ_s	2.3	3.0	1.6

- R is the receiver term, $R = [1.1\log_{10}(f) - 0.7]\, h_r - [1.56\log_{10}(f) - 0.8]$,
- E is a constant that depends on the type of environment: for urban areas $E = 3$ while for flat rural or suburban areas $E = 0$.

Compared to the model of [39], the COST 231-Hata model has the advantage of explicitly accounting for frequency and for height of the receive antenna. However, it only applies for higher transmit antennas, and provides the median propagation loss, rather than the statistics available when the standard deviations are used in the Erceg model [39].

Figure 9.11 shows propagation losses computed using these two models, with results for the Erceg model [39] including two-sigma values of propagation loss exponent and shadowing loss, which are shown in dash-dot lines. If many measurements were made for different transmitter and receiver locations, the measured values would typically fall between the envelope formed by these dash-dot lines. There is a tremendous variation in propagation loss at distances greater than a few hundred meters from the transmitter. Some of this variation involves the terrain and vegetation conditions. But even for a given terrain and vegetation condition, propagation losses can vary by more than 50 dB at a distance of 2 km, accounting for different propagation paths and fading. For assured delivery of signal power, the upper values of the propagation loss envelope should be used. However, if the

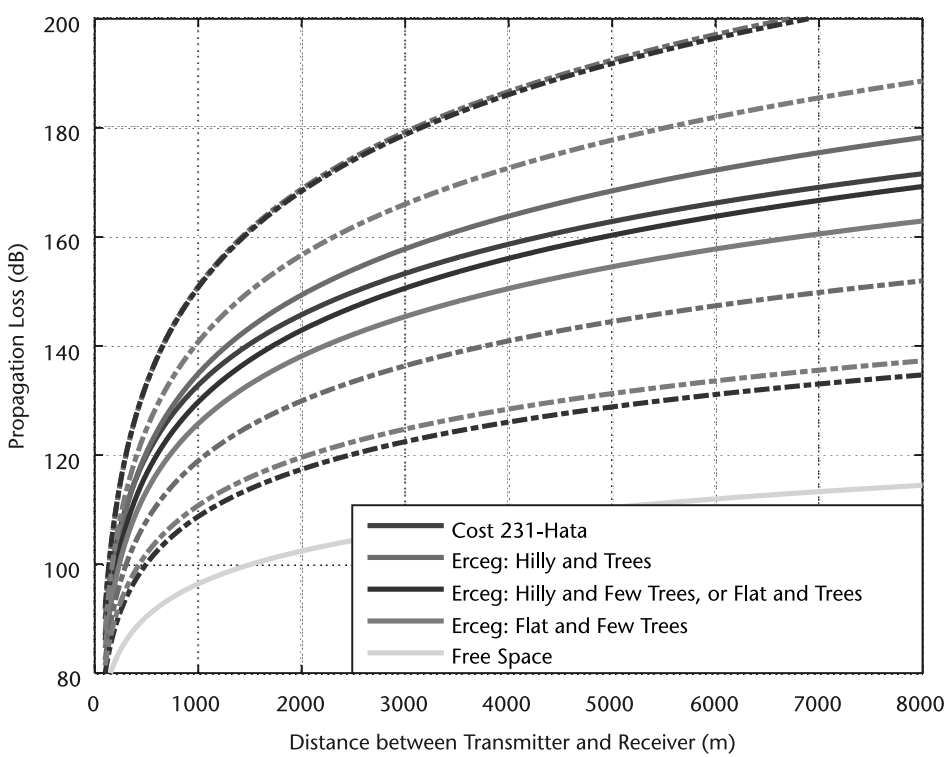


Figure 9.11 Propagation losses computed using Erceg model and the COST 231-Hata model; transmit antenna height 30m, receive antenna height 2m; 1,575.42 MHz frequency for COST 231-Hata model; solid lines are median loss, while dash-dot lines are median plus or minus two-sigma propagation loss exponent and shadowing loss.

transmitter is an interference source, the lower value of the propagation loss envelope should be used for assured maximum received interference power. The lower values of the envelopes have slopes similar to free space, indicating propagation loss coefficients near two can occur. However, the upper values of the envelopes have propagation loss coefficients approaching six. Hilly conditions with trees typically produce the greatest propagation loss, while median propagation loss is 10 dB to 20 dB less at distances of kilometers when there is flat terrain and few trees. Results using the COST 231-Hata model are very similar to those for the Erceg model [39] with hilly terrain and few trees or flat terrain with trees.

The same types of results are shown in Figure 9.12, where the only change is that the transmit antenna height is reduced to 10m. This one change in parameter value causes significant effects. The propagation loss coefficients are much greater than two, even for the lower values of the envelopes. Values for the three different terrain categories of the Erceg model are grouped, and median values for this model are reasonably close to those for the COST-231-Hata model, although the latter model predicts lower mean loss. In both of these figures, the free space propagation model significant under predicts propagation loss for this case.

Empirical models apply only to limited conditions corresponding to those where sufficient data was gathered and analyzed. The model [39] applies at 1,900 MHz, outside the range of L-band GNSS. Since the COST 231-Hata model predicts that median propagation loss changes by less than 3 dB between 1,900 MHz

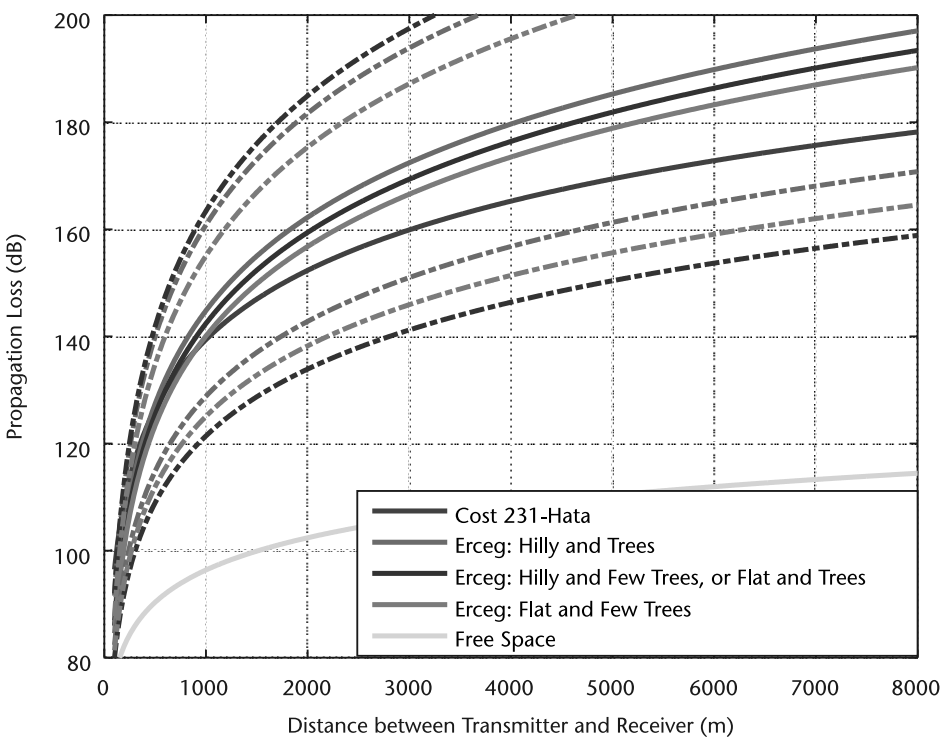


Figure 9.12 Propagation losses computed using Erceg model and the COST 231-Hata model; transmit antenna height 10m, receive antenna height 2m; 1,575.42-MHz frequency for COST 231-Hata model; solid lines are median loss, while dash-dot lines are median plus or minus two-sigma propagation loss exponent and shadowing loss.

and 1,575.42 MHz, however, it appears that the Erceg model is still appropriate at this lower frequency. However, a more significant limitation is the transmitter antenna height. Since the empirical model data was gathered to support cellular communications laydowns, data gathering for the COST 231-Hata model only used transmit antennas heights as small as 30m, while the Erceg model applies to transmit antenna heights as low as 10m. Extrapolating the results to lower antenna heights carries some risk to accuracy of the predictions.

9.4.3 Man-Made Structures

If either the transmitter or the receiver, or both, are inside man-made structures, then additional propagation losses typically occur. These losses should be added (in decibels) to the propagation losses computed using the models described previously.

Building penetration losses vary considerably with building construction, materials, structure, and the location of the receiver or transmitter within the building. The presence or absence of windows, and even the difference between metal window frames and wooden window frames, can make a significant difference in the propagation loss into a room. Besides propagation loss, signals received within a building often experience significant multipath. Building penetration losses are discussed extensively [41]. The following discussion is based on [40].

The excess loss in decibels due to building penetration is typically modeled, for signals arriving from above the building, as

$$L = L_{roof} + n_{floor} L_{floor} \tag{9.57}$$

where

- L_{roof} is the roof penetration loss, which can range from 1 dB to 30 dB at L-band;
- n_{floor} is the number of floors penetrated;
- L_{floor} is the loss per floor, which can range from 1 dB to 10 dB at L-band.

For building penetration through walls, the excess loss in decibels is similarly typically modeled as

$$L = L_{ext} + n_{int} L_{int} \tag{9.58}$$

where

- L_{ext} is the exterior wall penetration loss, which can range from 1 dB to 30 dB at L-band;
- n_{int} is the number of interior walls penetrated;
- L_{int} is the loss per interior wall, which can range from 1 dB to 10 dB at L-band.

Table 9.13 lists representative losses for different building materials, drawing from an extensive set of measurements reported in [42].

Table 9.13 Measured Losses in Decibels for Different Building Materials

	Frequency (MHz)	
Material	*1,176.45*	*1,575.42*
Brick	3 to 7	5 to 9
Composite brick/concrete	14 to 25	17 to 33
Brick/masonry block	11	10
Poured concrete	12 to 45	14 to 44
Reinforced concrete	27 to 30	30 to 35
Masonry block	11 to 27	11 to 30
Drywall	0.2 to 0.5	0.4 to 0.7
Glass	0.8 to 3	1.2 to 4
Dry lumber	3 to 6	3.5 to 8
Wet lumber	3.5 to 7.5	6 to 10

9.5 Multipath

Improvements due to GNSS augmentations and GNSS modernization are reducing many sources of error, leaving multipath and shadowing as significant and often dominant contributors to error. This section discusses these sources of error, their effects, and ways to mitigate their effects.

Multipath is the reception of multiple reflected or diffracted replicas of the desired signal, along with the direct path signal. Since the path traveled by a multipath is always longer than the direct path, multipath arrivals are delayed relative to the direct path. When the multipath delay is large (e.g., greater than twice the spreading code symbol period for a BPSK-R modulation), receivers typically can readily resolve and reject the multipath. As long as the receiver tracks the direct path (which always arrives earlier than any multipath), such resolvable multipaths have little effect on performance. However, multipath reflections from nearby objects, or even grazing multipaths reflected from distant objects, can arrive at short delays (e.g., tens or hundreds of nanoseconds) after the arrival of the direct path. Such multipaths distort the correlation function between the received composite (direct path plus multipaths) signal and the locally generated reference in the receiver, and also distort the phase of the composite received signal, introducing errors in pseudorange and carrier phase measurements that are different among the signals from different satellites, and thus produce errors in position, velocity, and time.

The effects of multipath are commonly assessed when the direct path signal is received unattenuated, so that multipath power is lower than direct path power. When blockage or shadowing of the direct path occurs along with multipath, the received power of the multipath may be even greater than the received power of the shadowed direct path. Such a phenomenon can occur in outdoor situations as portrayed in Figure 9.13, and also in indoor situations, such as when the direct path is significantly attenuated while passing through walls or ceiling and roof, while the multipath is reflected from another building and arrives with little attenuation through a window or other opening. Consequently, shadowing of the direct path and multipath has combined effects on the relative amplitudes of direct path and

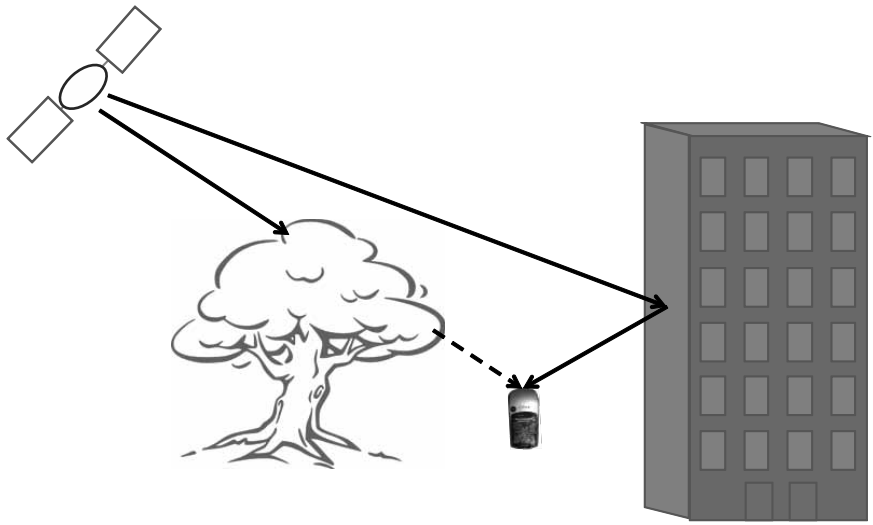


Figure 9.13 Outdoor multipath and shadowing situation.

multipaths. In some cases, shadowing of the direct path may be so severe that the receiver only tracks the multipath(s).

When the receiver can track the direct path, the error introduced by multipaths depends upon their delays, power and carrier phase relative to those of the direct path. Multipaths with received power much less than that of the direct path produce little distortion of the received signal, and consequently produce little error.

Typically, consideration of multipath in a GNSS context emphasizes its effect on signal code and carrier tracking accuracies, since these receiver functions are more sensitive to multipath degradation than signal acquisition or data demodulation. Under most situations, multipath conditions that would cause observable degradation to acquisition or data demodulation also introduce large degradations to pseudorange accuracy. Effects on acquisition and data demodulation are assessed using techniques developed in digital communications [43], and so the remainder of this discussion focuses on tracking performance in the presence of multipath.

Section 9.5.1 describes different models and characteristics of multipath. Section 9.5.2 relates the effect of multipath on signal tracking accuracy for situations involving different signal modulations, different precorrelation bandwidths, and different early-late spacings in the code tracking discriminator. Section 9.5.3 discusses some specialized techniques for multipath mitigation.

9.5.1 Multipath Characteristics and Models

The simplest model of multipath is a set of discrete reflected signals having larger delays and different amplitudes and carrier phases from the direct path. If the signal with no multipath is described in analytic signal form as

$$s(t) = \alpha_0 x(t - \tau_0) e^{-j\phi_0} e^{j2\pi f_c (t-\tau_0)}, \tag{9.59}$$

where $x(t)$ is the complex envelope of the transmitted signal, τ_0 is the time in seconds for the signal to propagate from satellite to receiver via the direct path, and f_c

is the carrier frequency in hertz, then a simple model for the complex envelope of a received signal with multipath (neglecting noise and interference) after frequency downconversion (neglecting any intentional IF) is

$$r(t) = \alpha_0 e^{-j\phi_0} x(t-\tau_0) e^{-j2\pi f_c \tau_0} + \sum_{n=1}^{N} \alpha_n e^{-j\phi_n} x(t-\tau_n) e^{j2\pi f_n t}, \tag{9.60}$$

where there are N multipaths, α_0 is the received amplitude of the direct path and α_n are the received amplitudes of the multipath returns, τ_0 is the propagation delay of the direct path and τ_n are the propagation delays of the multipath returns, ϕ_0 is the received carrier phase of the direct path and ϕ_n are the received carrier phases of the multipath returns, and f_n are the received frequencies of the multipath returns relative to the carrier frequency.

In general, each of the parameters in (9.60) is time-varying due to motion of the satellites and the receiver, as well as motion of objects that produce the multipath. This time variation is not shown explicitly in (9.60) because it complicates the notation. However, it is accounted for in some of the multipath models discussed below.

The expression (9.60) can be rewritten using parameters that relate the multipaths to the direct path:

$$r(t) = \alpha_0 e^{-j\phi_0} \left[x(t-\tau_0) + \sum_{n=1}^{N} \tilde{\alpha}_n e^{-j\tilde{\phi}_n} x(t-\tau_0-\tilde{\tau}_n) \right] \tag{9.61}$$

where $\tilde{\alpha}_n = \alpha_n / \alpha_0$ is the multipath-to-direct ratio (MDR) of amplitudes, $\tilde{\tau}_n = \tau_n - \tau_0$ is the excess delay of the multipath returns, and $\tilde{\phi}_n$ are the relative received carrier phases of the different signal components. The multipath profile producing (9.61) can be portrayed graphically as a power-delay profile (PDP) by plotting the points $\left\{ \left(\tilde{\tau}_n, \tilde{\alpha}_n^2 \right) \right\}\Big|_{n=1}^{N}$.

As noted earlier, the parameters in the expression (9.61) can be time varying. Linearly variations in the $\tilde{\phi}_n$ with time can be used to model situations where the received carrier frequencies of the multipaths are not the same as the received carrier frequency of the direct path. In the illustrative examples to follow, the $\tilde{\phi}_n$ terms will be assumed to be constant. This representation may not be adequate when relative motion between satellites, scatterers, and receiver is different from relative motion between satellites and receiver, causing multipath arrivals at different Doppler shifts from the direct path. As these Doppler differences increase and become greater than the reciprocal of the coherent integration time in the correlator, they cause the received multipath signals to be less correlated with the direct path.

A special case of (9.61) occurs when the propagation geometry is such that the direct path is nearly tangent to the Earth's surface (such as when the satellite is near the horizon). Then there can be a single dominant multipath arrival that reflects from a large object near the horizon, with excess delay orders of magnitude less than the reciprocal of the signal bandwidth and only a small fraction of the carrier period—often smaller than a picosecond. When the reflection coefficient is sufficiently high and there are no other multipaths, then $x(t-\tau_0-\tilde{\tau}_n) \cong x(t-\tau_0)$.

Consequently, (9.61) can be approximated (when the reflection introduces a 180° rotation of the carrier phase) as

$$r(t) \cong \alpha_0 e^{-j\phi_0}\left[1 - \tilde{\alpha}_1 e^{-j2\pi f_c \tilde{\tau}_1}\right] x(t - \tau_0) \tag{9.62}$$

where $f_c\tilde{\tau}_1$ is very small, so that when the reflection is strong enough that $\tilde{\alpha}_1$ is near unity, the magnitude of the quantity in square brackets is very much less than unity. The delay of this multipath is so small that it causes negligible pseudorange error, but by nearly canceling the direct path, it causes significant reduction in received signal power, relative to what would be observed with free-space propagation. This phenomenon is well-known in land mobile radio [44], and not addressed further in this section.

More general models of multipath channels [43] do not represent the fine structure as in models discussed previously, but instead represent the effect of the multipath channel [in our case, relative to the direct path, as in (9.61)] as a slowly time-varying linear system. The impulse response falls off with excess delay, and the range of excess delays where the impulse response is essentially nonzero is called the channel's multipath spread. In turn, the multipath spread can be represented by the root-mean squared (RMS) delay spread of the channel. This linear system has a time-varying transfer function that describes how it passes different frequency components of the signal.

Since the transfer function at a given frequency randomly varies over time, the correlation between transfer functions at different times and the same frequency [43] describes the time variation of the channel. If the time variation is fast relative to time constants in the receiver tracking loops, the multipath errors are smoothed by the receiver processing. Otherwise, they produce a time-invariant error or bias. The power spectral density resulting from the Fourier transform of this correlation is called the Doppler power spectrum of the channel, and the range of frequencies over which it is essentially nonzero is called the Doppler spread of the channel. The reciprocal of the Doppler spread is the coherence time of the channel—the time over which the multipath structure does not change much relative to the direct path. Two fundamental quantities introduced by this channel model—the multipath spread and the Doppler spread—provide succinct yet useful high-level representations of the multipath characteristics.

Despite its limited realism, the expression (9.61) with $N = 1$ and time-invariant parameters is widely used in theoretical assessments of multipath performance due to its ease of use. This time-invariant distortion produces a bias error in pseudorange. If the multipath is specular, the MDR remains independent of range from receiver to reflector, and hence independent of the multipath's excess delay. For a reflection to be truly specular, the reflector must be many wavelengths large, the reflecting surface must be smooth (surface roughness less than a few centimeters for L-band signals), and have consistent electrical properties. Observe that the one-path specular multipath model provides the limiting case of zero Doppler spread (time-invariant impulse response) and infinite delay spread.

On airplanes at altitude, multipath typically involves reflections from surfaces such as the wings and tail, sometimes accompanied by creeping waves over the aircraft skin. Aircraft multipath may be characterized as a discrete number of

reflections all occurring with relative delays less than 20 ns and relative amplitudes less than 0.3 for vehicles as large as a Boeing 747 [45]. The model (9.61) can be employed for this situation; since the reflecting surfaces are close to the receive antenna and share the same motion, the multipath parameters, including the phases $\tilde{\phi}_n$, may remain constant over time periods exceeding the reciprocal of tracking loop bandwidths, motivating use of time-invariant parameters over durations longer than the reciprocal of the signal tracking loops. For this case, the delay spread is very short (20 ns) and the Doppler spread is also small (perhaps thousandths of a hertz).

In terrestrial applications, there have been extensive efforts to measure, model, and predict the diverse multipath environments that may be encountered. For some applications, multipath can be characterized as a large number of reflections from objects in the proximity of the user. A general model for this diffuse multipath is presented in [46]. In this model, 500 small reflectors are randomly located within 100m of the user. Since the reflectors are small, each emanates a spherical wave and thus the received power from each reflector varies with the square of the distance between the reflector and the user. Moreover, the large number of signal reflections, spaced so closely in delay, make the multipath arrivals appear to result from passing through a linear filter with continuous impulse response amplitude decreasing with excess delay, rather than the discrete delays in (9.60) and (9.61). This diffuse scatterer model has been found to closely represent measured multipath for an aviation differential GNSS reference station application with the receiver located in an open environment. Here, the delay spread is hundreds of nanoseconds, and the Doppler spread is tenths or hundredths of a hertz.

Among many attempts to measure and model real-world multipath environments, [47] stands out as offering a particularly comprehensive and useful representation of complex terrestrial multipath. As shown in Figure 9.14, the parametric model is based on (9.60), with the arrivals grouped into three components: the direct path, a discrete set of near echoes, and a discrete set of far echoes. Shadowing of the direct path is represented by a Rice distribution of amplitude when line-of-sight (LOS) visibility exists between the receiver and the transmitter, and a Rayleigh distribution when LOS visibility does not exist. The mean received power of the near echoes falls off exponentially with delay. The number of far echoes is typically much smaller than the number of near echoes, and the mean value of the far echoes does not vary over the range of delays. The numbers of near echoes and number of far echoes are each Poisson distributed, described by different Poisson parameters. Multipath phases are modeled as independent and identically distributed over

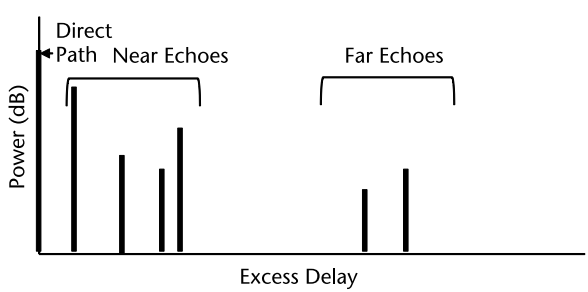


Figure 9.14 Canonical power-delay-profile for land-mobile satellite channel.

360°. Extensive tables of statistical parameters for these components are provided in [47] for many different environments (e.g., open, rural, urban, highway) and satellite elevations. Time variation of the multipath characteristics is described in [47] using second-order statistics based on Doppler spectra, with bandwidth established by the movement of satellites and the receiver.

As noted in [47], the line of sight does not always exist between the receiver and transmitter, particularly for low-elevation angles. For instance, trees or buildings along a road may block signals from below a certain elevation angle. In urban environments, 97% of signals were blocked when the transmitter was at an elevation angle of 15°, and blockage of lower-elevation satellites was also not uncommon even in rural environments, due to shadowing by trees. In these circumstances, it is entirely possible for a receiver to track a reflection rather than the direct signal, causing large pseudorange errors.

Over the range of environments and elevation angles considered in [47], the average power of the near echoes never exceeds –16.5 dB relative to the average power of the direct path. The mean power levels of the near echoes fall off at a wide variety of rates, ranging from 1 to 37 dB/μs depending on the elevation angle and environment. The range of delays associated with the near echoes is from 0 to 0.6 μs. No significant far echoes occur beyond 5 to 15 μs and the mean power levels of the far echoes are within the range of –20 to –30 dB (relative to an unshadowed direct path). Doppler spreads are dominated either by the satellite motion or the receiver motion. Delay spreads are often multiple microseconds, while Doppler spreads for a stationary receiver can be tenths of hertz, but for a receiver in a vehicle can be many hertz, particularly for multipaths with small excess delay.

Indoor multipath has very different characteristics depending on the placement of the building relative to other buildings, satellite elevation, whether the receiver is in an interior area deep within the building or near a window, what floor the receiver is on, and the building materials. Except in cases where the direct path is shadowed, multipath with significant values of MDR typically arises from reflections near the receive antenna, thus having small excess delay. Indoor data discussed in [48] have RMS delay spread less than 50 ns, with delay spread less than 250 ns. The Doppler spread is often dominated by the motion of the receiver and can be fractions of a hertz for stationary receivers or multipaths with large excess delay and hertz for multipaths with small excess delay and receivers being carried by a person.

While it is difficult to make any generalizations about phenomena as highly variable as multipath and shadowing, several observations can be made. Shadowing exacerbates any multipath effects, and severe shadowing can cause the receiver to track a multipath rather than a direct path, causing potentially large ranging errors. Near-in multipaths are often the most stable over time for a receiver that does not move relative to its local environment, but the fastest varying in time for a receiver that moves relative to its local environment. Near-in multipaths often have the greatest MDR, but typically introduce smaller ranging errors than multipaths with larger excess delays.

9.5.2 Effects of Multipath on Receiver Performance

Since received signals from different satellites typically encounter different multipath channels, the resulting pseudorange errors are not common to signals received from different satellites, and thus produce errors in position, velocity, and time. Further, the size of the multipath errors in tracking different satellites may also be very different, since signals received from higher-elevation satellites tend to experience less multipath in many applications. Ironically, the contributions of lower-elevation satellites to improved dilution of precision can provide an important incentive to use these signals, in spite of their larger multipath errors.

As discussed in Section 9.5.1, actual multipath environments are both complicated and diverse, making it difficult to quantify the effects of multipath in ways that are both generally applicable yet accurate. Computer simulations that synthesize waveforms, and then employ high-fidelity channel models and specific receiver processing approaches can provide accurate and realistic assessments, yet provide little insight into underlying issues and characteristics. In contrast, the multipath model (9.61) has limited realism, but provides useful insights. In fact, extensive assessments have been made using the one-path specular multipath version of (9.61). While the numerical results obtained are often not representative of real-world multipath conditions, they provide useful diagnostic insights. Further, it is sufficient, although it may not be necessary, to perform well under these simple conditions.

For the one-path specular multipath model, (9.61) can be rewritten for $N = 1$ (continuing to neglect noise and interference), as

$$r(t) = \alpha_0 e^{-j\phi_0}\left[x(t-\tau_0) + \tilde{\alpha}_1 e^{-j\tilde{\phi}_1} x(t-\tau_0-\tilde{\tau}_1)\right] \tag{9.63}$$

When the locally generated replica $x(t)e^{-j\theta}$ is correlated against this received signal, the statistical mean of the result is

$$\begin{aligned}\overline{\lambda}(\tau) &= \alpha_0 e^{-j(\tilde{\phi}_0-\theta)}\left[R_x(\tau-\tau_0) + \tilde{\alpha}_1 e^{-j\tilde{\phi}_1} R_x(\tau-\tau_0-\tilde{\tau}_1)\right] \\ &= \alpha_0 e^{-j(\tilde{\phi}_0-\theta)}\hat{R}_x(\tau-\tau_0)\end{aligned} \tag{9.64}$$

The term $\hat{R}_x(\tau-\tau_0) = R_x(\tau-\tau_0) + \tilde{\alpha}_1 e^{-j\tilde{\phi}_1} R_x(\tau-\tau_0-\tilde{\tau}_1)$ is a composite correlation function that is the sum of the ideal correlation function and a second version of the ideal correlation function that is scaled in amplitude, rotated in phase, and delayed. When the receiver attempts to estimate delay and carrier phase from this composite correlation function, its estimates are in error, even in the absence of noise and interference.

Figure 9.15 illustrates the effect of one-path multipath on noncoherent early-late processing, for a signal with BPSK-R(1) modulation strictly bandlimited to 4 MHz. The top row shows results with no multipath, while subsequent rows show results for multipath phase (relative to phase of the direct path) of 0°, 90°, and 180°. The left columns show magnitude-squared correlation functions, while the right columns show the pseudorange error introduced by the multipath, for different values of early-late spacing. The phase of the multipath dictates whether the error is positive or negative. With the narrow precorrelation bandwidth in this

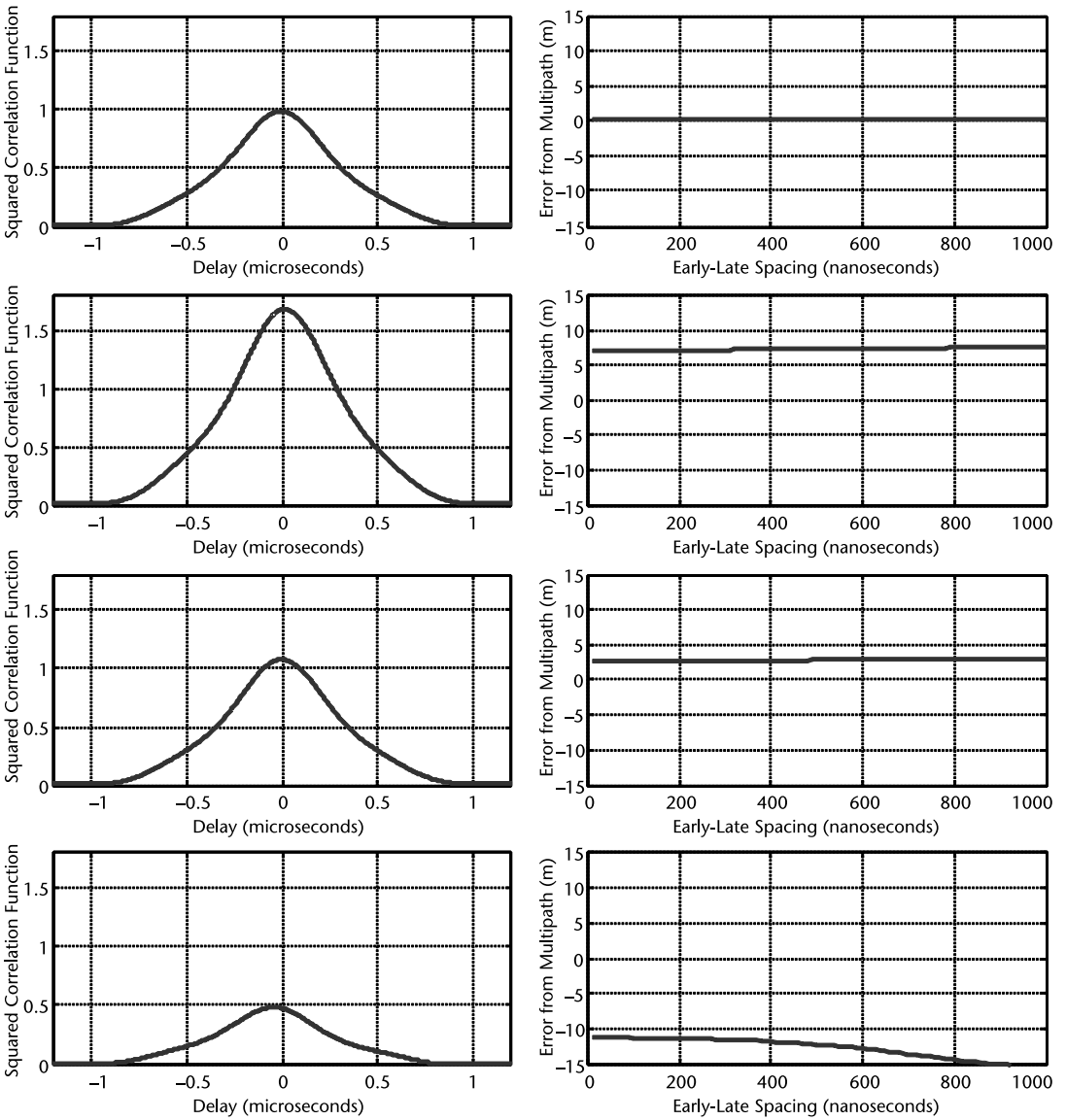


Figure 9.15 Effects of one-path multipath on pseudorange estimation, BPSK-R(1) modulation strictly bandlimited to 4 MHz. The top row shows no multipath, while in subsequent rows MDR is –10 dB and excess delay is 0.1 μs, and phase is 0°, 90°, and 180°. The left column shows distorted correlation functions, while the right column shows dependence of range error on early-late spacing.

case, narrower early-late spacings examined here have little effect on the error in most cases.

Figure 9.16 shows the same results as in Figure 9.15, for a signal with BPSK-R(1) modulation bandlimited to 24 MHz. With the wider precorrelation bandwidth, narrower early-late spacing significantly reduces the error in most cases.

Figure 9.17 shows the same results as in Figure 9.16, for a signal with BPSK-R(10) modulation bandlimited to 24 MHz. Since the sharper correlation function peak resolves this multipath better, the ranging errors tend to be smaller. When these results are repeated for multipath excess delay of 0.4 μs, the errors for BPSK-R(1)

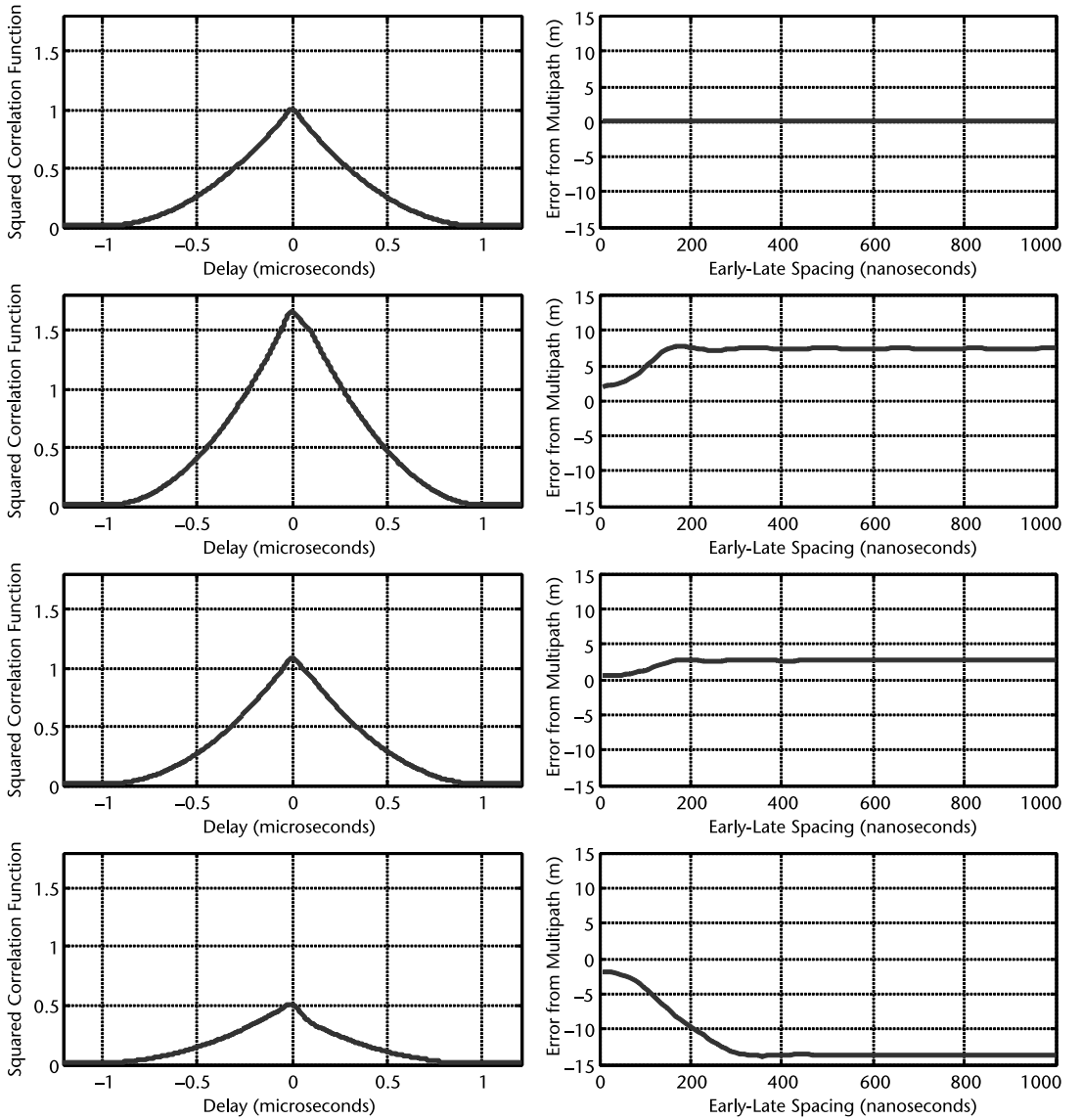


Figure 9.16 Effects of one-path multipath on pseudorange estimation, BPSK-R(1) modulation strictly band-limited to 24 MHz. Top row shows no multipath, while in subsequent rows MDR is –10 dB and excess delay is 0.1 μs, and phase is 0°, 90°, and 180°. The left column shows distorted correlation functions, while the right column shows dependence of range error on early-late spacing.

modulation are similar to those in Figures 9.15 and 9.16, while BPSK-R(10) modulation displays no errors, since its sharper correlation function peak completely resolves the multipath with larger excess delay.

For a more comprehensive depiction of ranging error caused by one-path multipath, recognize that, for a given modulation and receiver design (including pre-correlation bandwidth and code tracking discriminator), multipath error is determined by the MDR, phase, and delay of the multipath. When the MDR is constant (independent of delay and phase) as in specular multipath, the error for a given multipath delay varies with multipath phase, as seen in Figures 9.15, 9.16, and 9.17. For a given MDR, the maximum and minimum errors at each delay are taken

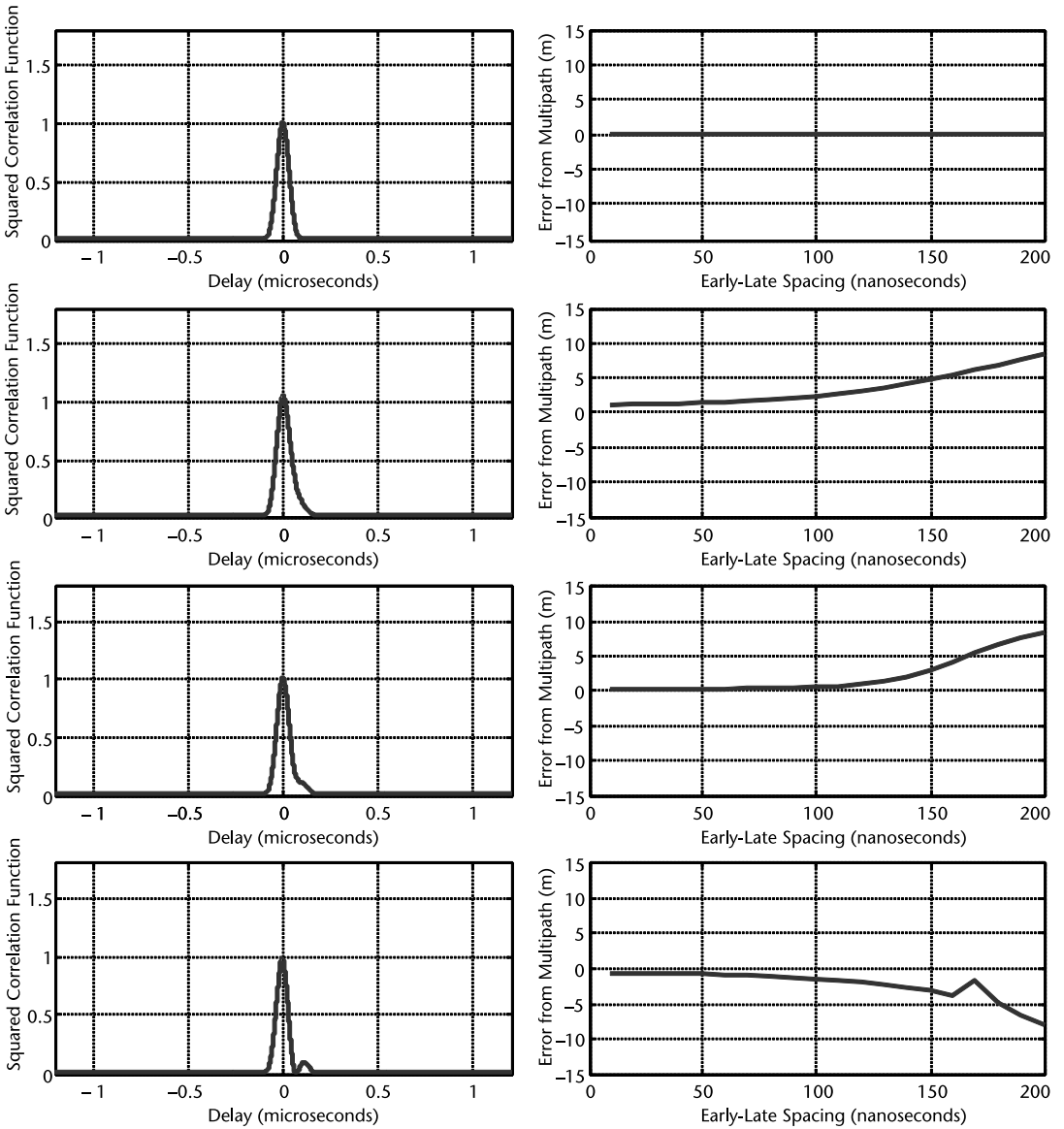


Figure 9.17 Effects of one-path multipath on pseudorange estimation, BPSK-R(10) modulation strictly band-limited to 24 MHz. The top row shows no multipath, while in subsequent rows MDR is –10 dB and excess delay is 0.1 μs, and phase is 0°, 90°, and 180°. The left column shows distorted correlation functions, while the right column shows dependence of range error on early-late spacing.

over all multipath phase values, producing a range of possible delay estimates for each value of excess delay. If $\hat{\tau}\left(\tilde{\alpha}_1, \tilde{\tau}_1, \tilde{\phi}_1\right)$ is the estimated delay for a specific MDR, excess delay, and multipath phase, then denote the error in delay estimation by $\varepsilon\left(\tilde{\alpha}_1, \tilde{\tau}_1, \tilde{\phi}_1\right) = \tau_0 - \hat{\tau}\left(\tilde{\alpha}_1, \tilde{\tau}_1, \tilde{\phi}_1\right)$. The maximum and minimum errors for a specific MDR and excess delay are, respectively, $\max_{\tilde{\phi}_1 \in [0, 2\pi]} \varepsilon\left(\tilde{\alpha}_1, \tilde{\tau}_1, \tilde{\phi}_1\right)$ and $\min_{\tilde{\phi}_1 \in [0, 2\pi]} \varepsilon\left(\tilde{\alpha}_1, \tilde{\tau}_1, \tilde{\phi}_1\right)$, where the latter takes on negative values; the envelope of delay errors at a given excess delay is defined by

$$\left(\max_{\tilde{\phi}_1\in(0,2\pi]}\varepsilon\left(\tilde{\alpha}_1,\tilde{\tau}_1,\tilde{\phi}_1\right),\min_{\tilde{\phi}_1\in(0,2\pi]}\varepsilon\left(\tilde{\alpha}_1,\tilde{\tau}_1,\tilde{\phi}_1\right)\right) \tag{9.65}$$

The resulting envelope of ranging errors is obtained by multiplying the envelope of delay errors by the speed of light.

Figure 9.18 shows multipath ranging error envelopes for BPSK-R(1) modulation with two different precorrelation bandwidths, and BPSK-R(10), all with early-late spacing of 50 ns. For the BPSK-R(1) modulation, the wider precorrelation bandwidth, combined with the narrow early-late spacing, provides smaller error, as recognized in [49]. The BPSK-R(10) modulation provides even smaller errors.

To assess multipath performance over a range of possible delay values; define the average range error envelope as $\frac{1}{2\tilde{\tau}_1}\int_0^{\tilde{\tau}_1}\left[\max_{\tilde{\phi}_1}\varepsilon(\tilde{\alpha}_1,u,\tilde{\phi}_1)-\min_{\tilde{\phi}_1}\varepsilon(\tilde{\alpha}_1,u,\tilde{\phi}_1)\right]du$. The average envelope can provide useful insights, particularly for modulations whose range error envelopes oscillate with delay, such as some BOC modulations.

To assess the effect of one-path multipath on carrier phase estimation, consider further the composite correlation function obtained in (9.64) resulting from one-path multipath, $\hat{R}_x(\tau-\tau_0)=R_x(\tau-\tau_0)+\tilde{\alpha}_1e^{-j\tilde{\phi}_1}R_x(\tau-\tau_0-\tilde{\tau}_1)$. It has the real part $R_x(\tau-\tau_0)+\tilde{\alpha}_1\cos(\tilde{\phi}_1)R_x(\tau-\tau_0-\tilde{\tau}_1)$ and the imaginary part $\tilde{\alpha}_1\sin(\tilde{\phi}_1)R_x(\tau-\tau_0-\tilde{\tau}_1)$. The carrier phase angle of the composite correlation function relative to that of the direct path is given by

$$\psi=\tan^{-1}\left[\frac{\tilde{\alpha}_1\sin\left(\tilde{\phi}_1\right)R_x\left(\tau-\tau_0-\tilde{\tau}_1\right)}{R_x\left(\tau-\tau_0\right)+\tilde{\alpha}_1\cos\left(\tilde{\phi}_1\right)R_x\left(\tau-\tau_0-\tilde{\tau}_1\right)}\right] \tag{9.66}$$

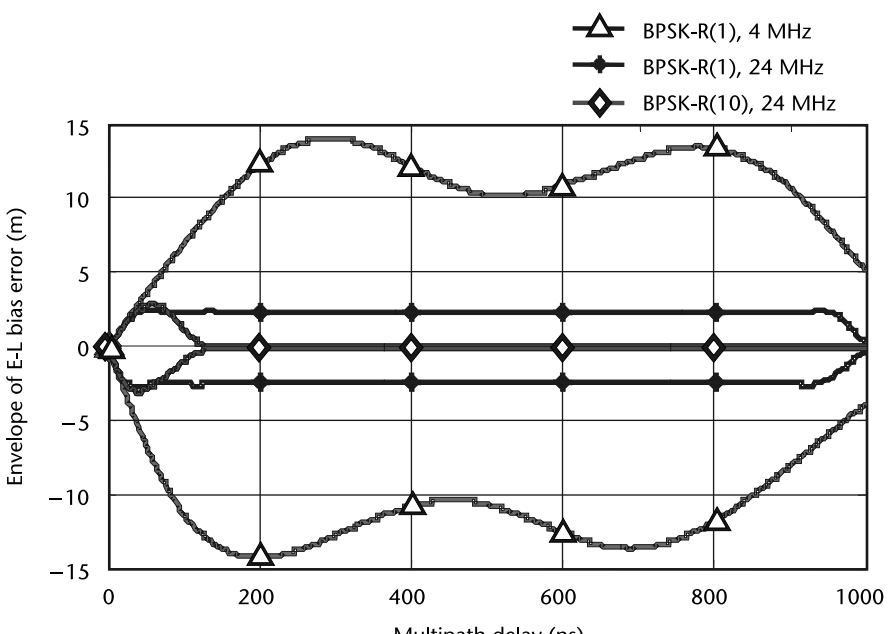


Figure 9.18 Multipath ranging error envelops showing the maximum and minimum code tracking error for one-path multipath with MDR –10 dB, at different multipath delays.

Thus, ψ is the carrier phase error introduced by the multipath. The carrier phase error is then a function of the multipath characteristics and the delay error.

Observe that when $\tilde{\tau}_1$ is very small, $\hat{R}_x(\tau-\tau_0) \cong R_x(\tau-\tau_0-\tilde{\tau}_1)$, and $\psi \cong \tan^{-1}\left[\dfrac{\tilde{\alpha}_1 \sin(\tilde{\phi}_1)}{1+\tilde{\alpha}_1 \cos(\tilde{\phi}_1)}\right]$. When the multipath power is equal to the power in the direct path, the carrier phase error is greatest when the multipath carrier phase is 180° relative to the direct path, producing a carrier phase error of 90°. As long as the MDR is less than or equal to unity, and the delay-locked loop maintains track on the correlation function of the direct path, the magnitude of the carrier phase error is less than or equal to 90°.

For a given MDR and excess delay, the carrier-phase error varies with the multipath phase and the error in delay estimate. The resulting minimum and maximum carrier-phase errors are given, respectively, by

$$\begin{aligned}
\psi_{\min}(\tilde{\alpha}_1,\tilde{\tau}_1) &= \min_{\tilde{\phi}_1\in(0,2\pi]} \tan^{-1}\left[\frac{\tilde{\alpha}_1 \sin(\tilde{\phi}_1) R_x\left(\varepsilon(\tilde{\alpha}_1,\tilde{\tau}_1,\tilde{\phi}_1)-\tilde{\tau}_1\right)}{R_x\left(\varepsilon(\tilde{\alpha}_1,\tilde{\tau}_1,\tilde{\phi}_1)\right)+\tilde{\alpha}_1 \cos(\tilde{\phi}_1) R_x\left(\varepsilon(\tilde{\alpha}_1,\tilde{\tau}_1,\tilde{\phi}_1)-\tilde{\tau}_1\right)}\right] \\
\psi_{\max}(\tilde{\alpha}_1,\tilde{\tau}_1) &= \max_{\tilde{\phi}_1\in(0,2\pi]} \tan^{-1}\left[\frac{\tilde{\alpha}_1 \sin(\tilde{\phi}_1) R_x\left(\varepsilon(\tilde{\alpha}_1,\tilde{\tau}_1,\tilde{\phi}_1)-\tilde{\tau}_1\right)}{R_x\left(\varepsilon(\tilde{\alpha}_1,\tilde{\tau}_1,\tilde{\phi}_1)\right)+\tilde{\alpha}_1 \cos(\tilde{\phi}_1) R_x\left(\varepsilon(\tilde{\alpha}_1,\tilde{\tau}_1,\tilde{\phi}_1)-\tilde{\tau}_1\right)}\right]
\end{aligned} \tag{9.67}$$

so the resulting envelope of carrier phase error is given by $(\psi_{\max}(\tilde{\alpha}_1,\tilde{\tau}_1), \psi_{\min}(\tilde{\alpha}_1,\tilde{\tau}_1))$ Figure 9.19 shows multipath carrier phase error envelops for the same conditions as Figure 9.18: BPSK-R(1) modulation with two different precorrelation bandwidths, and BPSK-R(10), all with early-late spacing of 50 ns. While the smaller ranging error envelope for BPSK-R(1) with wider precorrelation bandwidth translates into a somewhat smaller carrier phase error envelope, the distinctly sharper correlation function of BPSK-R(10) produces much better performance for this multipath model.

Since this one-path specular multipath model has limited realism (only one multipath with delay-invariant MDR and no time variation) and the processing

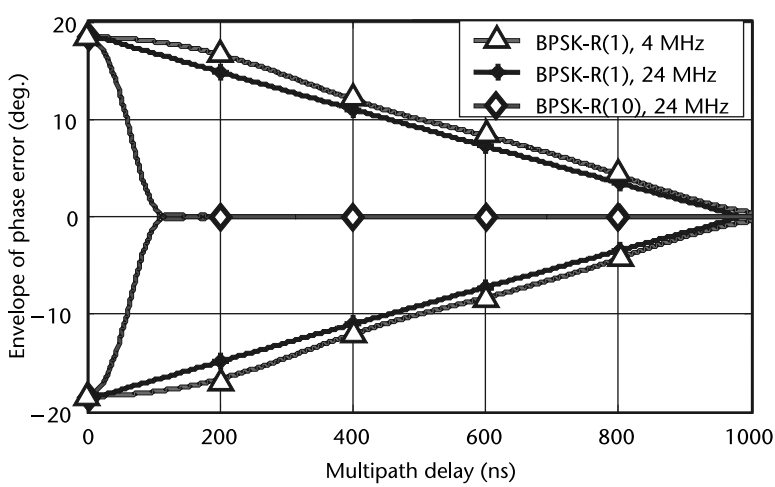


Figure 9.19 Multipath carrier phase error envelopes showing the maximum and minimum code tracking error for one-path multipath with MDR –10 dB, at different multipath delays.

model does not include smoothing of time varying errors by loop filters in the receiver, the quantitative results in Figures 9.15 through 9.19 tend not to represent actual errors in actual multipath environments. However, the qualitative reduction of multipath errors through use of small early-late spacings, wider precorrelation bandwidths, and modulations with sharper correlation function peaks is borne out in practice. Further, it is sufficient but not necessary in all situations to mitigate the errors caused by fixed multipath, since signal tracking loops integrate out some of the multipath error when the rate of multipath variation exceeds the loop bandwidth. This relatively rapid multipath variation occurs particularly when the receiver moves relative to the scatterers reflecting the multipath, so that at least the received multipath phase varies differently than the received phase of the direct path. However, when the receiver is stationary, multipaths from stationary scatterers can produce errors that vary little over typical loop filter time constants in a receiver; particularly for nearby scatterers.

As discussed in Section 9.5.1, multipath models other than the one-path static model are often more realistic, and thus provide more realistic quantitative results. Figure 9.20 shows range errors computed using the diffuse multipath model [50]. These simulated results are similar to measured results, and confirm the previous qualitative conclusions that wider precorrelation bandwidths are the most important way to obtain lower errors in multipath, and wider bandwidth modulations also provide benefits.

The corresponding RMS carrier phase error is shown in Figure 9.21. The differences are not as significant as for code tracking error.

The results in this section demonstrate that smaller errors from multipath can be obtained through use of wider signal bandwidths, wider precorrelation bandwidths, and narrower early-late spacing (when used in conjunction with wider precorrelation bandwidths). An extensive set of results comparing different spreading modulations, bandwidths, and early-late spacings is provided in Chapter 22 of [40]. The quantitative amount of improvement depends critically on the specific multipath environment, including the time variation of the multipath and smoothing of errors within the receiver processing. In some applications, shadowing of the direct path is common, and the errors that result from tracking of a multipath can be more important than the errors from multipath when the direct path is present.

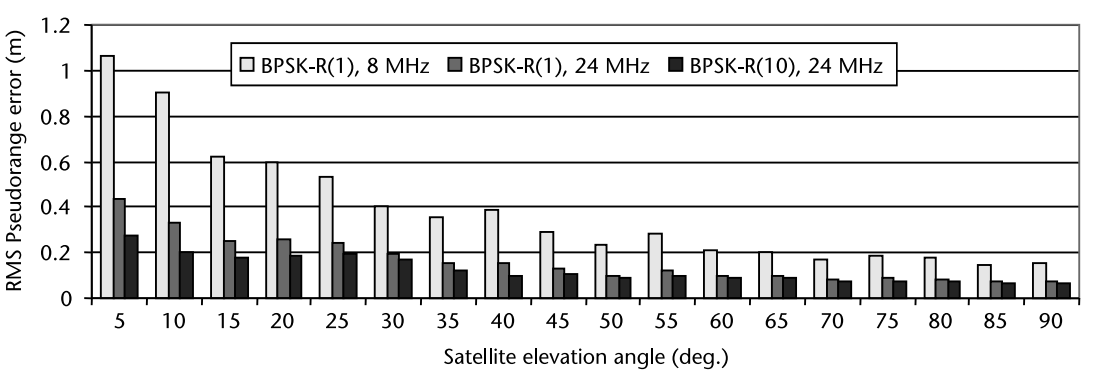


Figure 9.20 Root-mean squared ranging error for diffuse multipath model, for signals received from satellites at different elevation angles (numerical results from [50]).

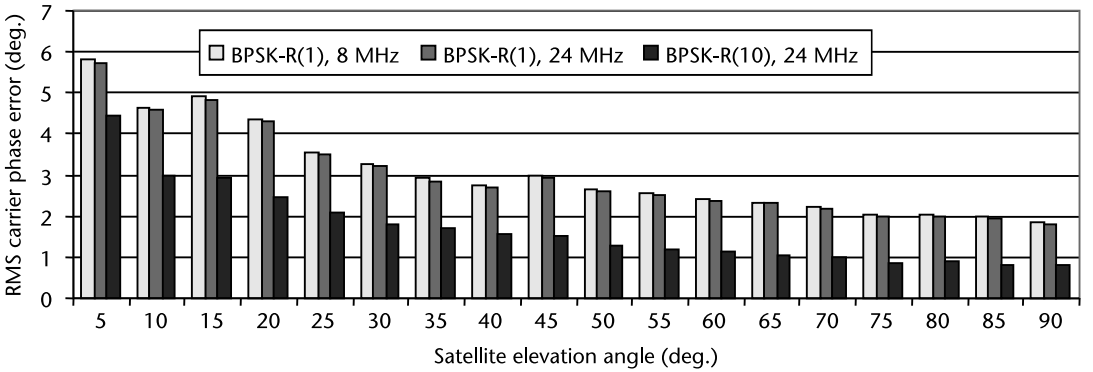


Figure 9.21 Root-mean squared carrier phase tracking error for diffuse multipath model, for signals received from satellites at different elevation angles (numerical results from [50]).

When wider precorrelation bandwidths and narrower-early-late spacing are employed, the complexity of receiver processing increases due to the higher sampling rates.

9.5.3 Multipath Mitigation

The dominance of multipath-induced errors in some applications has motivated considerable investigation into development of multipath mitigation techniques that go beyond the straightforward strategies described in Section 9.5.2. Some multipath mitigation techniques have been incorporated into production receivers, while others remain research topics. Chapter 17 in [51] provides a useful overview of advanced multipath mitigation techniques, complementing the discussion in this section.

A number of considerations arise in assessing multipath mitigation techniques. Good performance in realistic multipath conditions must be provided. Robustness is also important, ensuring that performance is satisfactory over the range of environmental conditions (including noise and interference) in which the receiver must operate. Implementation complexity is also a factor, as are any restrictions on how the receiver would be employed (such as requirements for multipath characteristics to be time-invariant over long periods of time, or restrictions to use by fixed receivers). While multipath mitigation techniques remain important research topics, this section outlines some of the strategies that have been pursued and remain areas of active work.

One important group of multipath mitigation techniques attempts to reduce the reception of multipath signals, reducing the need to discriminate against these multipaths by the receiver processing. Antenna siting, and even removal or modification (e.g., coating with RF-absorptive materials) of reflective structures in the vicinity of the antenna can produce significant benefits. In benign environments such as an open field, placing an antenna closer to the ground can decrease observed multipath errors. The reason is that, with the antenna closer to the ground, the multipath reflections from the ground experience shorter excess path delays that tend to produce smaller multipath errors, as shown in Figure 9.18. Conversely, in environments with obstacles near the horizon, the opposite course of action is

often beneficial—raising the antenna decreases antenna gain at elevation angles corresponding to dominant reflectors that produce multipath.

Antennas can also be designed to attenuate multipath reflections, particularly multipaths that arrive at elevation angles near or below the horizon, where desired signals are not expected to arrive. Choke ring antennas have been particularly successful for mitigating multipath arrivals from the ground or low-elevation scatterers. Since reflections reverse the polarization of electromagnetic waves, these choke ring antennas can also be designed to reject left-hand circularly polarized waves. In short-baseline differential systems, multipath errors at a fixed reference station can also be reduced through calibrations that measure multipath error based on satellite position [52].

Techniques for multipath mitigation receiver processing can be divided into nonparametric and parametric processing. Nonparametric processing employs discriminator designs that are less sensitive to multipath-induced errors, while parametric processing attempts to estimate parameters associated with the multipath and then correct for their effect on the estimate of the direct path's time of arrival.

Some nonparametric techniques, such as those in [53, 54], rely on precise prior knowledge of the signal's correlation function, and employ novel receiver processing approaches that attempt to match the ideal correlation function to the observed correlation function in multipath. Nonparametric techniques in most common use, however, are based on variations of early-late processing described in Chapter 8. However, they go beyond the narrow correlator approach considered in Section 9.5.2 by either time-gating the reference signal or by computing two pairs of early and late correlations with different early-late spacings. A number of similar techniques have been developed and implemented in different brands of receivers.

An excellent overview of these modified-reference techniques, their capabilities, and their limitations, is provided in [55]. One interpretation is that this processing is equivalent to generating a modified locally-generated reference signal that does not replicate the desired signal, but rather approximates the derivative of the desired signal. The resulting correlation between the received signal and the modified reference has a much sharper correlation peak (along with, for some approaches, small artifacts at larger delays) than the original signal, providing better resolution of multipaths just as the P(Y)-code signal provides better resolution as shown in Section 9.5.2. These approaches provide little or no benefit for multipaths with very small [a few tens of nanoseconds for BPSK-R(1) modulations] excess delay, but do provide enhanced performance for multipaths with larger delays, compensating in part for the limitations of narrower-bandwidth modulations (as long as the precorrelation bandwidth is wide). However, as discussed later in this section, their benefits are offset to some degree by poorer performance in noise and interference, compared to use of conventional early-late processing with more capable modulations and the same precorrelation bandwidth.

Most parametric approaches rely on the discrete model of multipath defined in (9.60) or (9.61). A parametric algorithm either estimates or assumes the number of multipaths, and then estimates nuisance parameters such as MDR, excess delay, and relative carrier phase of each multipath. Typically, these parametric approaches employ carrier-coherent processing and use very long coherent integration times (greater than 1 second), requiring the received multipath characteristics (including phase relative to the direct path) to be stable over the integration time. One such

approach is the Multipath Estimating Delay Lock Loop (MEDLL) [56], which applies the maximum likelihood estimation theory to minimize the mean-squared error between the received signal [modeled as in (9.60)] and the locally generated reference signal. Other approaches have been proposed [57] and shown to minimize mean-squared error and root-mean squared error for specific multipath models.

Only limited evaluations have been published to describe the effect of noise and interference on performance of multipath mitigation techniques. The analysis in [57] shows that modified-reference processing degrades the postcorrelation signal-to-noise ratio by large amounts. However, this degradation is readily overcome by use of a conventional prompt correlator with a locally-generated reference signal matched to the transmitted signal. The results in [55] also demonstrate that the code tracking accuracy of modified reference techniques in white noise is degraded relative to conventional early-late processing by an amount equivalent to reducing the signal power by 3 dB at higher input signal-to-noise conditions, and perhaps greater amounts at C/N_0 less than 35 dB-Hz. While the effects of nonwhite interference have not been evaluated, it can be expected that performance would be degraded more by interference with power concentrated away from band center, compared to conventional early-late processing. This increased sensitivity to noise and interference can be offset by use of narrower loop bandwidths, although the effect of dynamics on these techniques has not been documented, and narrower loop bandwidths would further degrade performance in dynamics.

Multipath mitigation remains an area of active research interest. Designs of new GNSS signals provide opportunities for new modulation designs, and better performance in multipath can be one consideration. However, there are many other constraints and factors that must be considered in GNSS modulation design, including issues that arise in sharing frequency bands with multiple signals. The increasing opportunity to process signals at multiple frequency bands opens up new potential for multipath mitigation processing that takes advantage of multiple carrier frequencies and multipath's frequency-selective characteristics. There are also opportunities to explore processing of multiple polarizations, although many antenna designs exhibit predominantly linear polarization response at low elevation angles of many multipath arrivals. Improved receiver processing techniques may still be developed, as may better theoretical understanding of capabilities and limitations of multipath mitigation. Most multipath mitigation techniques incur practical consequences, such as increased receiver complexity and poorer performance in noise and interference, which must be evaluated on a case-by-case basis.

References

[1] Dovis, F., (ed.), *GNSS Interference Threats and Countermeasures*, Norwood, MA: Artech House, 2015.

[2] Hegarty, C. J., et al., "An Overview of the Effects of Out-of-Band Interference on GNSS Receivers," *Proc. of the 24th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2011)*, Portland, OR, September 2011, pp. 1941–1956.

[3] Hegarty, C. J., "Considerations for GPS Spectrum Interference Standards," *Proc. of the 25th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2012)*, Nashville, TN, September 2012, pp. 2921–2929.

[4] Betz, J. W., "Effect of Narrowband Interference on GPS Code Tracking Accuracy," *Proc. of ION 2000 National Technical Meeting*, Institute of Navigation, January 2000.

[5] Betz, J. W., and D. B. Goldstein, "Candidate Designs for an Additional Civil Signal in GPS Spectral Bands," *Proc. of The Institute of Navigation National Technical Meeting*, San Diego, CA, January 2002.

[6] Ward, P. W., "GPS Receiver RF Interference Monitoring, Mitigation, and Analysis Techniques," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 41, No. 4, Winter 1994-95, pp. 367–391.

[7] Hata, M., "Empirical Formula for Propagation Loss in Land Mobile Radio Service," *IEEE Trans. on Vehicular Technology*. Vol.29, August 1980, pp. 317–325.

[8] Mogensen, P. E., et al., "Urban Area Radio Propagation Measurements at 955 and 1845 MHz for Small and Micro Cells," *Proc. of IEEE Globecom*, 1991.

[9] Sklar, B., *Digital Communications: Fundamentals and Applications*, Upper Saddle River, NJ: Prentice Hall, 2000.

[10] Betz, J. W., "On the Power Spectral Density of GNSS Signals, with Applications," *Proc. of the 2010 International Technical Meeting of The Institute of Navigation*, San Diego, CA, January 2010, pp. 859–871.

[11] Spilker, J. J., Jr., Van Dierendonck, A. J., "Proposed New L5 Civil GPS Codes," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 48, No. 3, Fall 2001, pp. 135–144.

[12] Spilker, J. J., Jr., "GPS Signal Structure and Performance Characteristics," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 25, No. 2, 1978.

[13] Scott, H. L., "GPS Principles and Practices," The George Washington University Course 1081, Vol. I, Washington, D.C., March 1994.

[14] Betz, J. W., and K. R. Kolodziejski, "Generalized Theory of Code-Tracking with an Early-Late Discriminator," *IEEE Trans. on Aerospace and Electronic Systems*, Vol. 45, No. 4, October 2009.

[15] Hegarty, C., et al., "Suppression of Pulsed Interference Through Blanking," *Proc. of The International Association of Institutes of Navigation 25th World Congress/The Institute of Navigation 56th Annual Meeting*, San Diego, CA, June 2000.

[16] Cameron, A., "eLoran Progresses Toward GPS Back-Up Role in U.S., Europe," *GPS World*, June 25, 2015.

[17] Rifkin, R. J., and J. Vaccaro, *Comparison of Narrowband Adaptive Filter Technologies for GPS*, MITRE Technical Report MTR 00B0000015, https://www.mitre.org/sites/default/files/pdf/rifkin_comparison.pdf.

[18] Amoroso, F., "Adaptive A/D Converter to Suppress CW Interference in DSPN Spread-Spectrum Communications," *IEEE Trans. on Communications*, Vol. Com-31, No. 10, October 1983, pp. 1117–1123.

[19] Amoroso, F., and J. L. Bricker, "Performance of the Adaptive A/D Converter in Combined CW and Gaussian Interference," *IEEE Trans. on Communications*, Vol. Com-34, No. 3, March 1986, pp. 209–213.

[20] Scott, H. Logan, Originally adapted Amoroso's nonuniform ADC design for military GPS receivers. This design was first used in the TI 4XOP family of military GPS receiver designs, Texas Instruments Incorporated, 1985.

[21] Ward, P., "Interference & Jamming: (Un)intended Consequences," GNSS Receivers," TLS NovAtel's Thought Leadership Series, *Inside GNSS*, September-October 2012, pp. 28–29.

[22] Ward, P. W., "Simple Techniques for RFI Situational Awareness and Characterization in GNSS Receivers," *Proc. of the 2008 National Technical Meeting of The Institute of Navigation*, San Diego, CA, January 2008, pp. 154–163.

[23] Rao, R., et al., *GPS/GNSS Antennas*, Norwood, MA: Artech House, 2012.

[24] Basu, S., et al., "250 MHz/GHz Scintillation Parameters in the Equatorial, Polar, and Auroral Environments," *IEEE Selected Areas in Communication*, Vol. SAC-5, No. 2, 1987, pp. 102–115.

[25] Aarons, J., and S. Basu, "Ionospheric Amplitude and Phase Fluctuations at the GPS Frequencies," *Proc. of The Institute of Navigation ION GPS-94*, Salt Lake City, UT, 1994, pp. 1569–1578.

[26] Klobuchar, **J.**, "Ionospheric Effects on GPS," *Global Positioning System: Theory and Applications – Volume I*, Washington, D.C.: American Institute of Aeronautics and Astronautics, Inc., 1996, pp. 485–515.

[27] Hegarty, C., et al., "Scintillation Modeling for GPS-Wide Area Augmentation System Receivers," *Radio Science*, Vol. 36, No. 5, September/October 2001, pp. 1221–1231.

[28] Chiou, T. Y., "Model Analysis on the Performance for an Inertial Aided FLL-Assisted-PLL Carrier-Tracking Loop in the Presence of Ionospheric Scintillation," *Proc. ION NTM 2007*, 2007, pp. 2895–2910.

[29] Shallberg, K., et al., "Dynamic Phase Lock Loop for Robust Receiver Carrier Phase Tracking," *Proc. of the 2009 International Technical Meeting of The Institute of Navigation*, Anaheim, CA, January 2009, pp. 924–936.

[30] Ashwitha, L. T., and G. Lachapelle, "Development of a Multi-Frequency Adaptive Kalman Filter-Based Tracking Loop for Ionospheric Scintillation Monitoring Receiver," *Proc. of the 28th International Technical Meeting of The Satellite Division of the Institute of Navigation* (*ION GNSS+ 2015*), Tampa, FL, September 2015, pp. 3797–3807.

[31] Jiao, Y., et al., "A Comparative Study of Triple Frequency GPS Scintillation Signal Amplitude Fading Characteristics at Low Latitudes," *Proc. of the 28th International Technical Meeting of The Satellite Division of the Institute of Navigation* (*ION GNSS+ 2015*), Tampa, FL, September 2015, pp. 3819–3825.

[32] Jiao, Y., et al., "Equatorial Scintillation Amplitude Fading Characteristics Across the GPS Frequency Bands," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 63, No. 3, Fall 2016, pp. 267–281.

[33] Delay, S. H., et al., "A Statistical Comparison of Satellite Tracking Performances During Ionospheric Scintillation for the GNSS Constellations GPS, Galileo and GLONASS," *Proc. of the 2016 International Technical Meeting of The Institute of Navigation*, Monterey, CA, January 2016, pp. 540–548.

[34] Meng, Y. S., and Y. H. Lee, "Investigations of Foliage Effect on Modern Wireless Communication Systems: A Review," *Progress in Electromagnetics Research*, Vol. 105, 2010, pp. 313–332.

[35] Goldhirsh, J., and W. J. Vogel, *Handbook of Propagation Effects for Vehicular and Personal Mobile Satellite Systems: Overview of Experimental and Modeling Results*, December 1998. http://rsl.ece.ubc.ca/archive/MobSatSysHandbook.pdf. Accessed February 19, 2016.

[36] Eppink, D., and W. Kuebler, *TIREM/SEM Handbook*, Department of Defense Electromagnetic Compatibility Analysis Center Handbook HDBK-93-076, March 1994.

[37] Seybold, J. S., *Introduction to RF Propagation*, New York: John Wiley and Sons, 2005.

[38] Abhayawardhana, V., et al., "Comparison of Empirical Propagation Path Loss Models for Fixed Wireless Access Systems," *Proc. of 61st IEEE Vehicular Technology Conference* (*VTC*), 2005.

[39] Erceg, V., et al., "An Empirically Based Path Loss Model for Wireless Channels in Suburban Environments," *IEEE Journal on Selected Areas in Communications*, Vol. 17, No. 7, July 1999, pp. 1205–1211.

[40] Betz, J. W., *Engineering Satellite-Based Navigation and Timing: Global Navigation Satellite Systems, Signals, and Receivers*, New York: Wiley-IEEE Press, 2016.

[41] Doble, J., *Introduction to Radio Propagation for Fixed and Mobile Communications*, Norwood, MA: Artech House, 1996.

[42] Stone, W. C., *Electromagnetic Signal Attenuation in Construction Materials*, NISTIR 6055, NIST Construction Automation Program, Report No. 3, October 1997.

[43] Proakis, J. G., *Digital Communications*, New York: McGraw-Hill, 1995.

[44] Parsons, J. D., *The Mobile Radio Propagation Channel*, 2nd ed., New York: John Wiley and Sons, 2000.

[45] Braasch, M., "GPS Multipath Model Validation," *Proc. of the IEEE Position, Location and Navigation Symposium, PLANS 96*, Atlanta, GA, April 22–25, 1996.

[46] Brenner, M., R. Reuter, and B. Schipper, "GPS Landing System Multipath Evaluation Techniques and Results," *Proc. of The Institute of Navigation ION GPS-98*, Nashville, TN, September 1998.

[47] Jahn, A., H. Bischl, and G. Heiß, "Channel Characterisation for Spread Spectrum Satellite Communications," *Proc. of the IEEE 4th International Symposium on Spread Spectrum Techniques and Applications* (*ISSSTA'96*), Mainz, Germany, September 1996.

[48] O'Donnell, M., et al., "A Study of Galileo Performance: GPS Interoperability and Discriminators for Urban and Indoor Environments," *Proc. of The Institute of Navigation's ION-GPS/GNSS-2002*, Portland, OR, September 2002.

[49] Van Dierendonck, A. J., P. Fenton, and T. Ford, "Theory and Performance of Narrow Correlator Spacing in a GPS Receiver," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 38, No. 3, Fall 1992, pp. 265–283.

[50] Hegarty, C. J., et al., "Multipath Performance of the New GNSS Signals," *Proc. of The Institute of Navigation National Technical Meeting*, San Diego, CA, January 2004.

[51] Jin, S., *Global Navigation Satellite Systems: Signal, Theory and Applications*, Croatia: In-Tech, 2012.

[52] Wanninger, L., and M. May, "Carrier Phase Multipath Calibration of GPS Reference Stations," *Proc. of The Institute of Navigation ION-GPS-2000*, Salt Lake City, UT, September 2000.

[53] Phelts, R. E., and P. Enge, "The Multipath Invariance Approach for Code Multipath Mitigation," *Proc. of The Institute of Navigation ION-GPS-2002*, Portland, OR, September 2002.

[54] Fante, R. L., and J. J. Vaccaro, "Multipath and Reduction of Multipath-Induced Bias on GPS Time-of-Arrival," *IEEE Trans. on Aerospace and Electronic Systems*, Vol. 39, No. 3, July 2003.

[55] McGraw, G. A., and M. S. Braasch, "GNSS Multipath Mitigation Using Gated and High Resolution Correlator Concepts," *Proc. of The Institute of Navigation ION-NTM-99*, January 1999.

[56] Townsend, B., et al., "Performance Evaluation of the Multipath Estimating Delay Lock Loop," *Proc. of The Institute of Navigation ION GPS-94*, Salt Lake City, UT, September 1994.

[57] Weill, L. R., "Multipath Mitigation Using Modernized GPS Signals: How Good Can It Get?" *Proceedings of the Institute of Navigation ION-GPS-2002*, September 2002.

CHAPTER 10

GNSS Errors

Christopher J. Hegarty, Elliott D. Kaplan, Maarten Uijt de Haag, and Ron Cosentino

10.1 Introduction

This chapter describes the most significant sources of error that corrupt the range measurements (pseudorange and carrier phase) made by GNSS receivers. As will be detailed in Chapter 11, for stand-alone positioning with GNSS there are two major factors in determining overall position accuracy: (1) the quality of the range measurements, and (2) the quality of the satellite geometry (e.g., how many satellites are visible and how well spread out in azimuth and elevation in the sky). Statistically, for a pseudorange-only position solution:

$$(\text{error in GNSS solution}) = (\text{geometry factor}) \times (\text{pseudorange error factor}) \quad (10.1)$$

Chapter 11 will formally derive this relationship and characterize the geometry factor for the GNSS constellations as well as the statistics of the position solution errors. This chapter focuses on the measurement error factor.

Importantly, before using the raw measurements to determine PVT, a stand-alone GNSS receiver corrects the pseudorange and carrier-phase measurements using a variety of techniques including utilizing elements of the broadcast navigation data from each satellite and mathematical models (e.g., for atmospheric errors). Thus, within (10.1) for standalone GNSS positioning, it is the statistics of the residual errors after corrections are applied within the receiver that influences overall performance.

For increased performance, many users augment GNSS with data from either differential or precise point positioning (PPP) ground networks. Such augmentations utilize one or more *reference stations*, which are essentially GNSS receivers at known locations, to measure GNSS errors and provide this data to the end user. As will be emphasized in Chapter 12 , with differential or PPP techniques, the absolute magnitude of the measurement errors made by the end user is far less important than how different these errors are than the measurement errors experienced by the differential or PPP reference stations. Errors that are seen completely in common between the reference receiver(s) and end user completely cancel when the

differential or PPP data is applied. The largest residual errors typically result for error sources such as multipath that are well characterized as statistically independent between receivers, even when they are separated only by short distances. With differential or PPP, rapidly changing errors can also deteriorate performance since these systems always have some latency in their provision of corrections to the end user. To pave the way for the discussion of such considerations in Chapter 12, this chapter includes descriptions of the spatial and temporal correlation characteristics of each GNSS error source.

Following this brief introduction, the remainder of the chapter is organized as follows. Section 10.2 describes the major sources of GNSS measurement errors, typical correction methods, and the characteristics of the residual errors after correction. Section 10.2 also describes the spatial and temporal correlation characteristics of each error source, which, as discussed above, are the most important characteristics for differential and PPP applications. Section 10.3 develops representative error budgets for stand-alone single-frequency and dual-frequency users.

10.2 Measurement Errors

The satellite and receiver clock offsets discussed in Chapter 2 directly translate into pseudorange and carrier-phase errors. The ranging code component of the satellite signal experiences delays as it propagates through the atmosphere making the pseudorange larger than it would be if the signal propagated in a vacuum. The carrier component of the signal is delayed by the troposphere but is actually advanced in phase by the ionosphere in a phenomenon referred to as *ionospheric divergence* that will be discussed in more detail in Section 10.2.4.1. Further, reflections (i.e., multipath) and hardware effects between the user's antenna phase center and receiver code correlation point may delay (or advance) the signal components [1]. The total time offset due to all of these effects on the ranging component of each received signal is:

$$\delta t_D = \delta t_{atm} + \delta t_{noise\&\text{int}} + \delta t_{mp} + \delta t_{hw} \tag{10.2}$$

where δt_{atm} = delays due to the atmosphere, $\delta t_{noise\&int}$ = errors due to receiver noise and interference, δt_{mp} = multipath offset, and δt_{hw} = receiver hardware offsets. A delay expression with the same form as (10.2), but with generally different numerical values, is incurred on the RF carrier component of each signal.

The pseudorange time equivalent is the difference between the receiver clock reading when the signal (i.e., a particular code phase) was received and the satellite clock reading when the signal was sent. These timing relationships are shown in Figure 10.1, where

Δt = geometric range time equivalent;

T_s = system time at which the signal left the satellite;

T_u = system time at which the signal would have reached the user receiver in the absence of errors (i.e., with δt_D equal to zero)

T_u' = system time at which the signal reached the user receiver with δt_D;

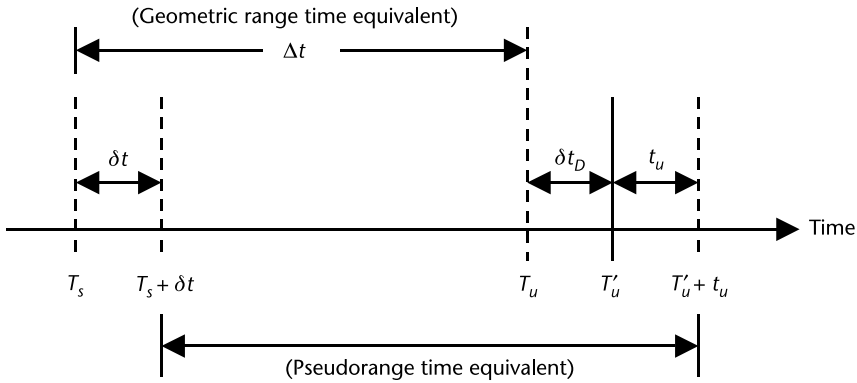


Figure 10.1 Range measurement timing relationships.

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 time which the signal left the satellite;

$T'_u + t_u$ = user receiver clock reading at time when the signal reached the user receiver;

c = speed of light.

It is observed that the pseudorange ρ is:

$$\begin{aligned}\rho &= c\left[\left(T'_u + t_u\right) - \left(T_s + \delta t\right)\right] \\ &= c\left(T'_u - T_s\right) + c\left(t_u - \delta t\right) \\ &= c\left(T_u + \delta t_D - T_s\right) + c\left(t_u - \delta t\right) \\ &= r + c\left(t_u - \delta t + \delta t_D\right)\end{aligned}$$

where r is the geometric range

$$r = c\left(T_u - T_s\right) = c\Delta t$$

A similar expression can be derived for the carrier-phase measurement when the raw measurement (see Section 8.10), usually computed in units of cycles, is converted to units of meters by multiplying by the carrier wavelength in meters. As noted above, the error terms are in general different for the carrier-phase measurement. Further, as discussed in Section 8.10, the carrier-phase measurement includes an ambiguity that is an integer multiple of a wavelength. Elaboration on the pseudorange and carrier-phase error sources, including relativistic effects, is provided in the following sections.

10.2.1 Satellite Clock Error

As discussed in Chapter 2, the GNSS satellites contain atomic clocks that control all onboard timing operations including broadcast signal generation. Although these

clocks are highly stable, they are not typically perfectly synchronized with their respective system times (e.g., GPS system time for the GPS satellites and GLONASS system time for the GLONASS satellites). Rather, the time read on the satellite clocks, referred to as SV time, is allowed to float within a certain tolerable range and clock correction fields in the navigation data are supplied to adjust for the deviation between SV time and GNSS time. Each GNSS CS determines and transmits clock correction parameters to the satellites for rebroadcast in the GNSS navigation message. Five of the six SATNAV systems described in this book (GPS, Galileo, BeiDou, QZSS, and NAVIC) utilize clock corrections based upon a second-order polynomial of the form [2–6]:

$$\Delta t_{SV} = a_{f0} + a_{f1}(t - t_{oc}) + a_{f2}(t - t_{oc})^2 + \Delta t_r \tag{10.3}$$

where

a_{f0} = clock bias (seconds);
a_{f1}= clock drift (s/s);
a_{f2} = frequency drift (i.e., aging) (s/s^2);
t_{oc} = clock data reference time (seconds);
t = current time epoch (seconds);
Δt_r = correction due to relativistic effects (seconds).

The specifications for GPS, QZSS, and IRNSS use the exact notation of (10.3). The Galileo interface control document (ICD) [3] uses the same form and notation as (10.3) except that it denotes the reference time as t_{0c} rather than t_{oc}. The BeiDou ICD [4] also uses the exact form of (10.3) but with slightly different notation: it refers to the a_{f0}, a_{f1}, and a_{f2} terms as, respectively, a_0, a_1, and a_2. The correction Δt_r compensates for one of the three relativistic effects discussed in Section 10.2.3.

GLONASS uses a polynomial clock correction similar to that in (10.3), but only of the first order. The GLONASS clock bias correction is referred to as τ_n and the clock drift correction as γ_n [7]. Also, within GLONASS these broadcast bias and drift correction terms already include relativistic adjustments, so no further correction for relativistic effects are needed by the receiver.

If the broadcast clock corrections are not applied by a GNSS receiver, extremely large pseudorange and carrier phase measurements can result. The error can be positive or negative, and the maximum magnitude of this error is approximately the range of the clock bias term in the broadcast clock correction parameters. For instance, for the GPS legacy signals, the a_{f0} term is an 11-bit signed two's complement number with a least significant bit (LSB) value of 2^{-20} seconds. Thus, the clock correction can be as large as nearly 1 ms in magnitude, yielding pseudorange and carrier phase errors of up to 300 km if the clock correction is not applied.

When the GNSS clock corrections are applied by the receiver, the pseudorange and carrier phase errors that result from the residual satellite clock error are typically small (i.e., not greater than several meters). The residual error has several contributors: (1) the clock correction data is generated by the respective CS using noisy measurements made by the CS monitor stations, and thus the CS estimates

of satellite clock errors are never perfect; (2) in the operational SATNAV systems, uploads are performed infrequently (e.g., once/day for GPS), so the broadcast clock correction data is necessarily based upon a prediction made up to a day in the past by the CS; and (3) the broadcast clock correction data is usually generated by a curve fit and truncated to conform to finite-length fields within the broadcast navigation message.

Due to the above factors, the residual clock error, δt, results in ranging errors that depend on the design of the CS, type of satellite, age of the broadcast data as well as data field representation adequacy. Range errors due to residual clock errors are generally the smallest following a CS upload to a satellite, and then slowly degrade over time until the next upload. As an example, see Figure 10.2, which depicts GPS SV residual clock error statistics as a function of time since last navigation upload for various GPS SV blocks. Nominal upload frequencies vary among the SATNAV systems. For instance, for each satellite the typical interval between uploads is 15 minutes for QZSS, 100 minutes for Galileo, 12 hours for GLONASS, and 24 hours for GPS. User equipment that is tracking all visible GNSS satellites will generally observe satellites with age-of-data (AOD)'s varying from 0 to 24 hours. It is thus appropriate, in the development of a statistical model for clock errors suitable for position or time error budgets to average over AOD.

Figure 10.3 shows the distribution of satellite clock errors for GPS based upon an analysis of navigation data broadcast by the legacy GPS signals from 2008 to 2014 [8]. The observed 1-sigma clock error averaged across the GPS constellation and for all AODs for this 7-year period was 0.5m. The mean clock error over the period was very small, –0.1 cm. One contributor to GPS clock errors at present is the fairly coarse LSB (approximately 0.5 ns) of the legacy GPS signal navigation data clock bias term, which equates to 14 cm in range. The modernized GPS signals

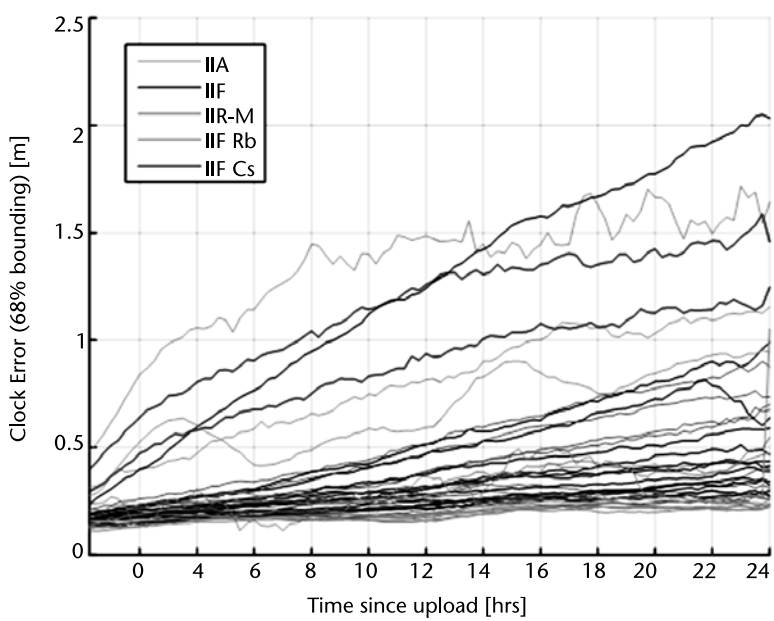


Figure 10.2 GPS satellite clock error (68% bounding) versus time since upload by satellite block, 2013–2016. (Courtesy of Kazuma Gunning/Stanford University.)

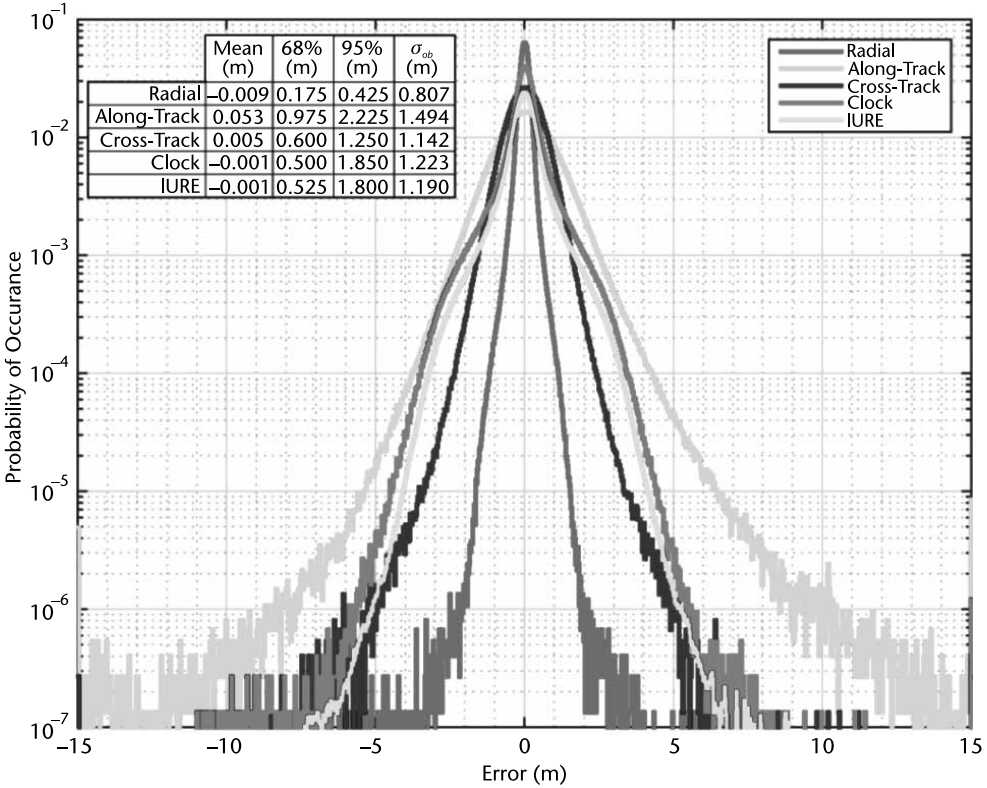


Figure 10.3 GPS satellite clock and ephemeris error statistics for the 7-year period from 2008 to 2014 [8].

use a much finer LSB (0.03 ns) for a_{f0}, which will reduce this error contributor in the future.

Data collected and analyzed in [9] for the period from 2009–2011 shows that clock errors dominate GLONASS signal-in-space (SIS) URE, which ranged from 1.5–4m over this period. SIS URE is a one-sigma error statistic that includes both clock errors and ephemeris errors (discussed in Section 10.2.2). GLONASS residual clock errors are improving as better performing clocks are introduced into the constellation and also as the CS is expanded. A more recent assessment of GLONASS SIS URE indicated that the typical 1-sigma SIS URE was 1.5–1.9m in 2015 [10]. AOD is typically under 12 hours for GLONASS based upon twice daily uploads. Quantization errors for the GLONASS FDMA signals are a significant contributor to GLONASS clock errors due to a coarse LSB of approximately 0.9 ns (0.3m) in the broadcast clock bias correction.

Data from [10] and other sources indicates that Galileo, BeiDou, QZSS, and NAVIC clock errors are already under 1m, 1-sigma, and are expected to diminish as these systems mature.

10.2.1.1 Spatial Correlation

Satellite clock errors contribute to overall pseudorange and carrier-phase measurement errors to exactly the same extent for all GNSS receivers, independent of their

location on or near Earth, making this GNSS error one of the simplest to correct using differential techniques. For instance, if a satellite clock (after application of the broadcast navigation data corrections) is in error by 10 ns, it will result in a 3-m pseudorange and carrier-phase measurement error for a user at any location.

10.2.1.2 Temporal Correlation

Today, GNSS satellite clock errors vary extremely slowly with time (i.e., the temporal correlation is very high), making latency in the delivery of corrections from a typical differential system an insignificant error source. As an example, Figure 5.8 in Chapter 5 provides Allan deviation measurements for the operational master clocks on-board the operational Galileo satellites for the period of October 2015 to January 2016. It can be observed that for a time interval of 200 seconds, the Allan deviations are below 2×10^{-13} s/s for both the passive hydrogen maser clocks (PHM) and rubidium atomic frequency standard (RAFS). This Allan deviation level corresponds to less than 0.06 mm/s rate of change over this interval. The stability of the clocks onboard other GNSS satellites are similar, yielding rates of change of the same order of magnitude.

10.2.2 Ephemeris Error

Estimates of ephemerides for all GNSS satellites are computed by their respective CS and uplinked to the satellites with other navigation data message parameters for rebroadcast to the user. As in the case of the satellite clock corrections, these corrections are generated using a curve fit of the CS's best prediction of each satellite's orbit at the time of upload. (See, for example, Section 3.3.1.4 for a description of the GPS curve fit process and sample results.) The residual satellite position error is a vector that is depicted in Figure 10.4.

The effective pseudorange and carrier-phase errors due to ephemeris prediction errors can be computed by projecting the satellite position error vector onto the satellite-to-user line-of-sight vector. Ephemeris errors are generally smallest in the radial (from the satellite towards the center of the Earth) direction. The components of ephemeris errors in the along-track (the instantaneous direction of travel of the satellite) and cross-track (perpendicular to the along-track and radial) directions are much larger. Along-track and cross-track components are more difficult for the CS to observe through its monitors on the surface of the Earth, since these components do not project significantly onto line of sights towards the Earth. Fortunately, the user does not experience large measurement errors due to the largest ephemeris error components for the same reason.

The distribution of GPS ephemeris errors over 2008–2014 is shown in Figure 10.3. Over this period, the radial, along-track, and cross-track components of the error had one-sigma levels of approximately 18 cm, 98 cm, and 60 cm, respectively. The overall GPS SIS URE was at the 52 cm level (shown on the figure as IURE, which is an acronym for instantaneous URE). As with clock errors, ephemeris errors generally grow with increasing AOD. This characteristic is illustrated in Figure 10.5, which plots GPS ephemeris errors as a function of time since upload. It can be observed that the broadcast ephemeris one-sigma levels are approximately linearly proportional to AOD. Figure 10.6 shows the overall SIS URE versus AOD.

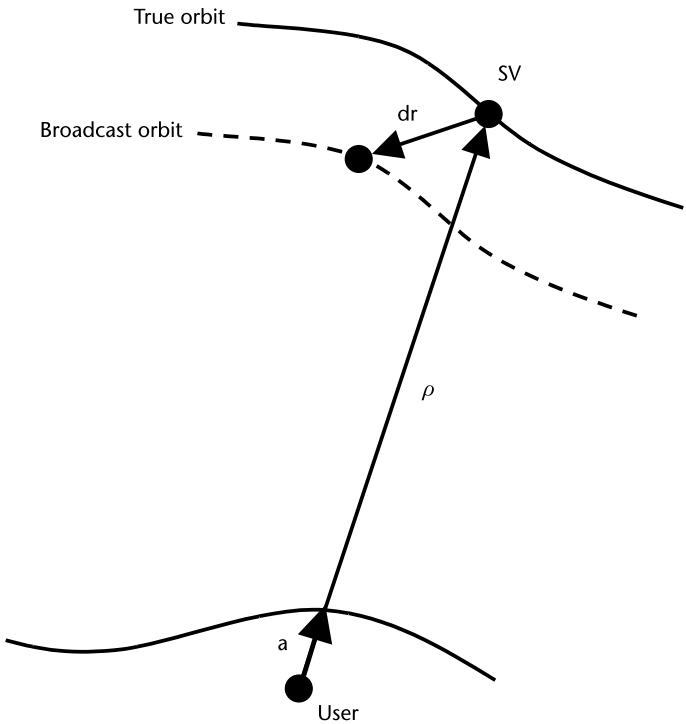


Figure 10.4 Ephemeris error.

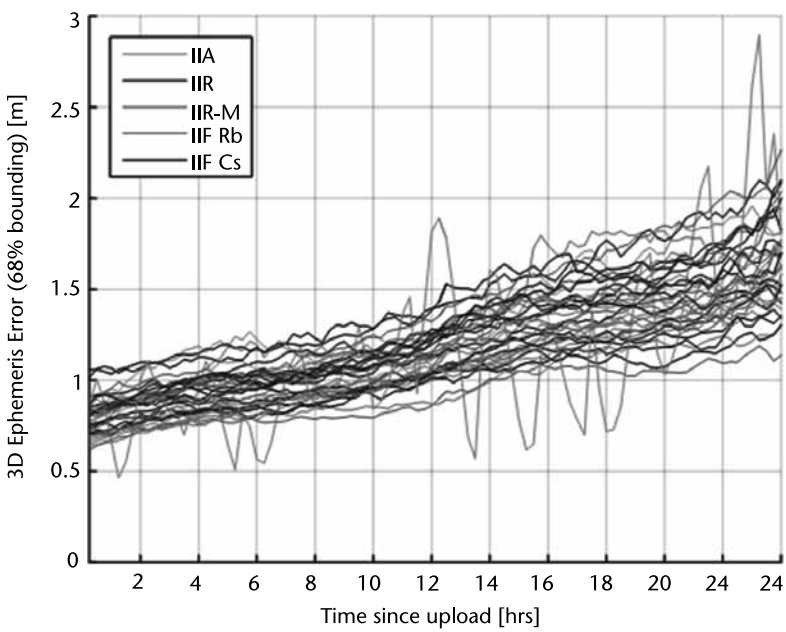


Figure 10.5 GPS ephemeris error (68% bounding) versus time since upload by satellite block, 2013–2016. (Courtesy of Kazuma Gunning/Stanford University.)

Galileo ephemeris errors were assessed in [10]. For the month of March 2015, the radial, along-track, and cross-track errors had one-sigma levels of 44 cm, 1.83

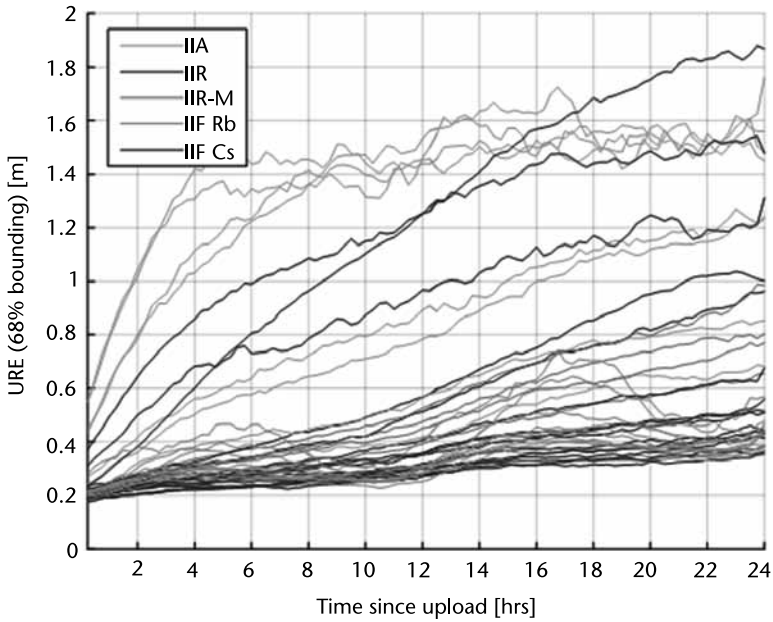


Figure 10.6 URE (68% bounding) versus time since upload by satellite block, 2013–2016. (Courtesy of Kazuma Gunning/Stanford University.)

cm, and 88 cm, respectively. The overall Galileo SIS URE was assessed to be at the 0.7-m level.

It should be noted that the geographical extent of CS monitoring networks plays a role in ephemeris determination. Geographic separation of monitoring stations enables more accurate orbit estimation. For instance, in GLONASS the monitoring station network was originally entirely within the Russian territory (see Section 4.3). Prior to 2012, there was on average only a 53% chance for a GLONASS satellite to be in view of at least one monitoring station if a 0° mask angle is assumed [9]. The GLONASS satellite ephemeris errors were observed to be dependent on whether the satellite is monitored. Figure 10.7 depicts the URE for monitored versus unmonitored SVs. As discussed in Chapter 4, Russia is expanding the GLONASS monitoring station network to outside of the Russian territory. The data in [9] suggests one-sigma GLONASS ephemeris errors for the period from 2009 to 2011 at approximately the half-meter level for radial, and meter-level for along-track and cross-track. The overall GLONASS SIS URE was at the 1–4-m level.

Data from [10] and other sources indicate that radial BeiDou, QZSS, and NAVIC ephemeris errors are submeter.

10.2.2.1 Spatial Correlation

Errors in the broadcast satellite positions lead to pseudorange and carrier-phase errors. Since the magnitude of ephemeris-induced pseudorange or carrier-phase errors are dependent on the line of sight between the user and the satellite, these errors change with user location. However, the difference in pseudorange or carrier-phase errors as seen by receivers in close proximity is very small, since their respective

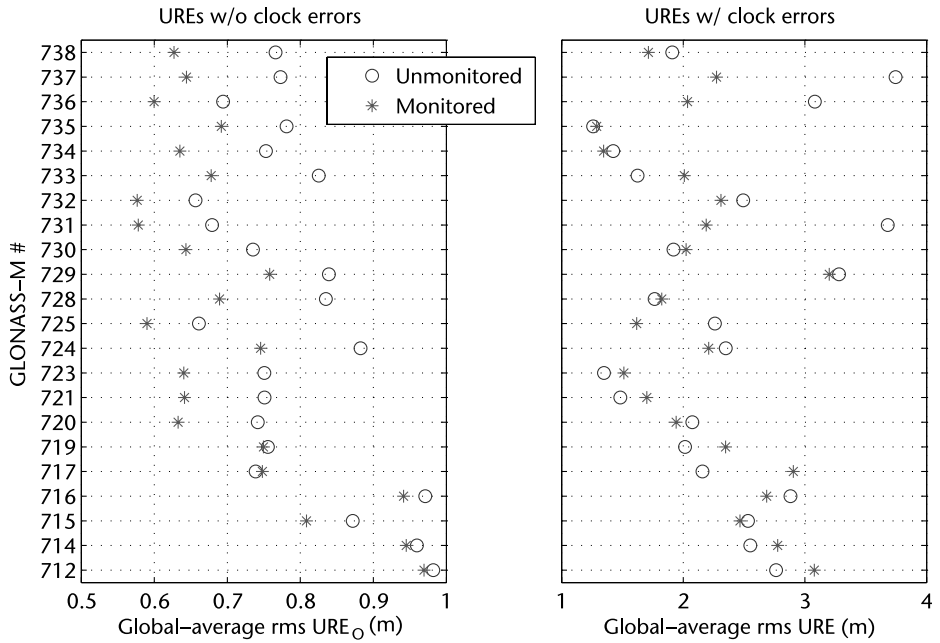


Figure 10.7 Geographic dependency of GLONASS SIS UREs. (From: [9].)

lines of sight to each satellite are very similar. To quantify the amount of change, let the separation between a user U and reference station M be denoted as p (Figure 10.8). We will refer to the actual orbital satellite position as the true position. The error in the estimated satellite position (i.e., the broadcast ephemeris) is represented as ε_s. Let d_{m} and d'_m be the true and estimated distances, respectively, of the reference station to the satellite, and let d_u and d'_u be the corresponding distances of the user to the satellite. Let ϕ_m be the angle formed by the directions of the reference station to

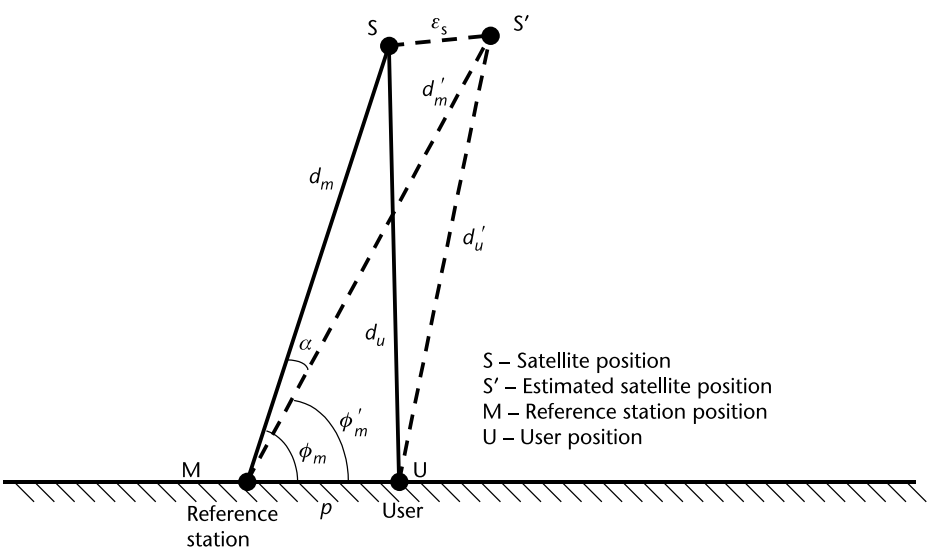


Figure 10.8 Variation of broadcast ephemeris errors with viewing angle.

the user and to the actual satellite position. Let α be the angle formed by the directions of the reference station to the actual and estimated positions of the satellite, S and S′, respectively. The law of cosines gives us the following two relationships:

$$d_u'^2 = d_m'^2 + p^2 - 2pd_m \cos(\phi_m - \alpha')$$
$$d_u^2 = d_m^2 + p^2 - 2pd_m \cos\phi_m$$

where α' is the difference $\phi_m - \phi_m'$ in elevation angles between the actual and estimated satellite positions from the monitor station. (The absolute value of α' is less than or equal to the absolute value of α and the two are equal when the two triangles lie in the same plane.)

Solving the first equation for $d_m' - d_u'$ and the second for $d_u - d_m$, and neglecting the higher-order terms in the binomial expansion of the square root in each of these equations, we obtain

$$d_m' - d_u' \approx -\frac{1}{2}\cdot\left(\frac{p}{d_m'}\right)\cdot p + p\cdot\cos\phi_m + \alpha'\cdot p\cdot\sin\phi_m + \frac{1}{2}\cdot\alpha'^2\cdot p\cdot\cos\phi_m$$
$$d_u - d_m \approx +\frac{1}{2}\cdot\left(\frac{p}{d_m'}\right)\cdot p - p\cdot\cos\phi_m$$

Adding these two equations, we find that the difference between the errors, $\varepsilon_u = d_u' - d_u$ and $\varepsilon_m = d_m' - d_m$, is

$$\varepsilon_m - \varepsilon_u = (d_u' - d_m') + (d_m - d_u) = \alpha'\cdot p\cdot\sin\phi_m + \frac{1}{2}\cdot\alpha'^2\cdot p\cdot\cos\phi_m$$

or

$$|\varepsilon_m - \varepsilon_u| = |(d_u' - d_m') + (d_m - d_u)| \le \alpha\cdot p\cdot\sin\phi_m + \frac{1}{2}\cdot\alpha^2\cdot p\cdot\cos\phi_m$$

where the equality holds if the estimated satellite position lies in the plane defined by the user position, reference station position, and true satellite position.

The difference $\varepsilon_m - \varepsilon_u$ is the error introduced by the pseudorange correction at the user. To simplify the expression, assume that the angle ϕ_m is greater than 5°, that the separation between the user and reference station is less than 1,000 km, and that the direction $\overline{SS'}$ is parallel to the direction $\overline{MU}$. Then

$$\varepsilon_m - \varepsilon_u \le \alpha\cdot p\cdot\sin\phi_m \approx \left(\frac{\varepsilon_s\cdot\sin\phi_m}{d_m}\right)\cdot p\cdot\sin\phi_m = \left(\frac{\varepsilon_s}{d_m}\right)\cdot p\cdot\sin^2\phi_m \qquad (10.4)$$

where ε_s is the error in the satellite's estimated position.

Equation (10.4) implies that the error increases directly with the separation between the reference station measuring the error and the user receiver employing the correction. Suppose, for example, that the error in the satellite's estimated position

is 5m and suppose the user is 100 km from the reference station. Then the error in the correction due to that separation is less than

$$\left(\frac{5\,\text{m}}{2\times10^4\,\text{km}}\right)\times100\,\text{km}=2.5\,\text{cm}$$

for elevation angles >5°.

10.2.2.2 Temporal Correlation

GNSS satellite ephemeris typically varies very slowly with time. For example, GPS ephemeris errors for 1 day in November 2001 were examined in [11] and it was observed that radial, along-track, cross-track, and overall three-dimensional (3-D) ephemeris errors for all satellites changed by no more than 1 cm over 10 seconds. The accuracy of the GPS broadcast ephemeris data has since increased significantly, and even lower rates of change are currently experienced. Similar low rates of change (corresponding to high levels of temporal correlation) have been observed for the ephemeris data broadcast by other GNSS satellites.

10.2.3 Relativistic Effects

Both Einstein's general and special theories of relativity are factors in the pseudorange and carrier-phase measurement process [12, 13]. The need for special relativity (SR) relativistic corrections arises any time the signal source (in this case, a GNSS satellite) or the signal receiver (GNSS receiver) is moving with respect to the chosen isotropic light speed frame, which in a GNSS system is the ECI frame. The need for general relativity (GR) relativistic corrections arises any time the signal source and signal receiver are located at different gravitational potentials.

The satellite clock is affected by both special and general relativity. In order to compensate for both of these effects, the satellite clock frequency needs to be adjusted prior to launch. Examples are:

- In the case of a highly inclined elliptical orbit (denoted as HEO) QZSS SV with reference frequency f_0 = 10.23MHz, the satellite clock is offset by the nominal $\Delta f/f_0 = -5.399E{-}10$ to compensate for the frequency difference between the ground surface and satellite orbit. For this reason, the center frequency in the satellite orbit is not exactly precise. For example, the L5 band signal is offset by –0.6352 Hz (nominal) [5]. The frequency observed by the user at sea level will be 1,176.45 MHz; hence, the user does not have to correct for this effect.
- In the case of GLONASS which uses FDMA, each satellite's carrier frequencies of L1 and L2 subbands are coherently derived from a common onboard time/frequency standard which to the user at sea level is 5.0 MHz. To compensate for relativistic effects, the nominal value of this reference frequency, as observed at the satellite, is biased from 5.0 MHz by the relative value $\Delta f/f_0$ = –4.36E–10 or $f = -2.18E{-}3$ Hz that is equal to 4.99999999782 MHz [7].

The user or SATNAV CS does have to make a correction for another relativistic periodic effect that arises because of eccentricities of the satellite orbits. This periodic effect is caused by the periodic change in the speed of the satellite relative to the ECI frame and also by the satellite's periodic change in its gravitational potential.

When the satellite is at perigee, the satellite velocity is higher and the gravitational potential is lower; both cause the satellite clock to run slower. When the satellite is at apogee, the satellite velocity is lower and the gravitational potential is higher; both cause the satellite clock to run faster [12, 13]. This effect can be compensated for by [2]

$$\Delta t_r = Fe\sqrt{A}\sin E_k \tag{10.5}$$

where

F = -4.442807633 × 10^{-10} s/$m^{1/2}$;

e = satellite orbital eccentricity;

A = semimajor axis of the satellite orbit;

E_k = eccentric anomaly of the satellite orbit.

The signal specifications for GPS, Galileo, BeiDou, QZSS, and NAVIC all prescribe application of the correction in (10.5) within the user receiver. In the case of GPS, [14] stated that this relativistic effect can reach a maximum of 70 ns (21m in range). Correcting the satellite clock for this relativistic effect will result in a more accurate estimation of the time of transmission by the user. For the GLONASS satellites, the broadcast satellite clock correction parameters already include the relativistic correction in (10.3) so the user equipment need not (and should not) apply this correction.

Due to rotation of the Earth during the time of signal transmission, a relativistic error is introduced, known as the *Sagnac effect*, when computations for the satellite positions are made in an Earth-centered Earth-fixed (ECEF) coordinate system (see Section 2.2.2). During the propagation time of the SV signal transmission, a clock on the surface of the Earth will experience a finite rotation with respect to an Earth-centered inertial (ECI) coordinate system (see Section 2.2.1). Figure 10.9 illustrates this phenomenon known as the Sagnac effect. Clearly, if the user experiences a net rotation away from the SV, the propagation time will increase and vice versa. If left uncorrected, the Sagnac effect can lead to maximum position errors on the order of 40m [15] for GLONASS and GPS and 46m for Galileo [15]. Corrections for the Sagnac effect are often referred to as *Earth rotation corrections*.

There are a number of approaches for correcting for the Sagnac effect. One common approach is to avoid the Sagnac effect entirely by working within an ECI coordinate system for satellite and user position computations. An ECI frame can be conveniently obtained by freezing an ECEF frame at the instant of time when pseudorange measurements are made to the set of visible satellites. The Sagnac effect does not arise in an ECI frame. Importantly, the satellite positions that are used in the standard GNSS user position solution (Section 2.5) must correspond to the times of transmission, which are generally not the same. The time of transmission

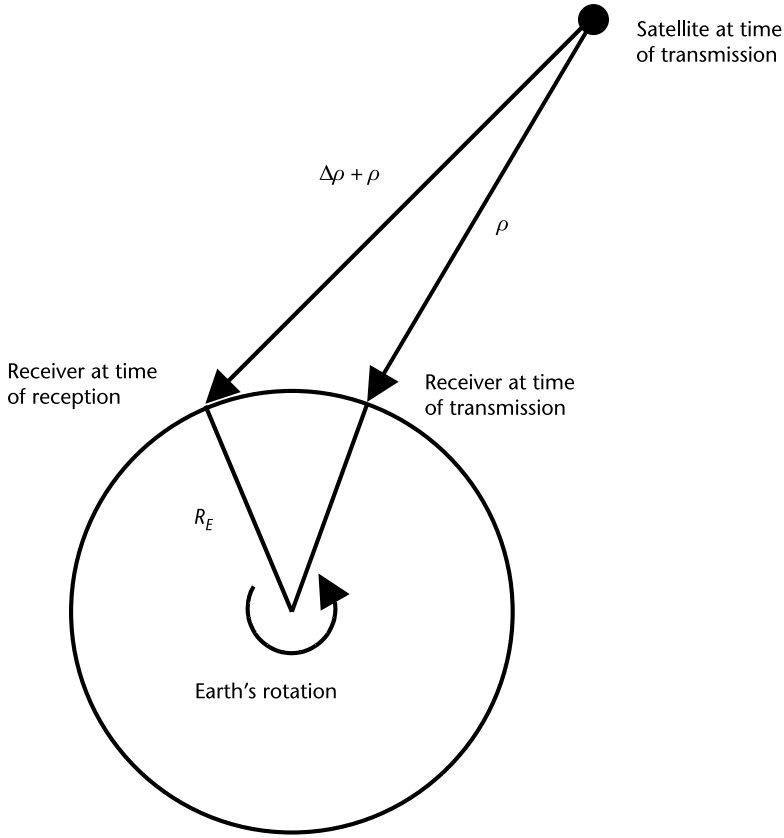


Figure 10.9 The Sagnac effect.

for each satellite, T_s, is a natural measurement of a GNSS receiver as discussed in Section 8.10. Users of commercial equipment can access time of transmission for each satellite by simply subtracting the pseudorange measurement (after applying the clock corrections discussed in Section 10.2.1) divided by the speed of light from the receiver's time tag for the measurement. Next, each satellite position can be computed in terms of its ECEF coordinates (x_s, y_s, z_s) at its time of transmission using the broadcast ephemeris data. (Tables 3.1 and 3.2 provide a GPS example). Then each satellite position can be transformed into the common ECI frame using the rotation:

$$\begin{bmatrix} x_{eci} \\ y_{eci} \\ z_{eci} \end{bmatrix} = \begin{bmatrix} \cos\dot{\Omega}(T_u - T_s) & \sin\dot{\Omega}(T_u - T_s) & 0 \\ -\sin\dot{\Omega}(T_u - T_s) & \cos\dot{\Omega}(T_u - T_s) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_s \\ y_s \\ z_s \end{bmatrix}$$

In this formulation, the time of reception, T_u, is initially unknown prior to the position/time estimate. As an example, it may be initially approximated as the

average time of transmission among visible satellites plus a mid-range satellite-to-receiver transit time (e.g., 75 ms for a GPS Earth-based user). Once the position solution is generated using the least-squares technique as described in Section 2.5, the user clock correction can be applied to obtain a much better estimate of T_u and the process can be iterated. The user's position coordinates are the same in both the ECEF and ECI frames at the signal reception time, since by definition these two frames were fixed at that instant. A number of alternative Earth rotation correction formulations, along with numerical examples, are provided in [16].

Finally, a GNSS signal experiences space-time curvature due to the gravitational field of the Earth. As an example, the magnitude of this relativistic effect for GPS can range from 0.001 ppm in relative positioning to about 18.7 mm for point positioning [17].

10.2.4 Atmospheric Effects

The propagation speed of a wave in a medium can be expressed in terms of the index of refraction for the medium. The index of refraction is defined as the ratio of the wave's propagation speed in free space to that in the medium by the formula

$$n = \frac{c}{v} \tag{10.6}$$

where c is the speed of light equal to 299,792,458 m/s as defined within the ITRF. The medium is dispersive if the propagation speed (or equivalently, the index of refraction) is a function of the wave's frequency. In a dispersive medium, the propagation velocity v_p of the signal's carrier phase differs from the velocity v_g associated with the waves carrying the signal information. The information-carrying aspect can be thought of as a group of waves traveling at slightly different frequencies.

To clarify the concepts of group and phase velocities, consider two components, S_1 and S_2, of an electromagnetic wave with frequencies f_1 and f_2 (or ω_1 and ω_2) and phase velocities v_1 and v_2, traveling in the x-direction. The sum S of these signals is

$$S = S_1 + S_2 = \sin\omega_1\left(t - \frac{x}{v_1}\right) + \sin\omega_2\left(t - \frac{x}{v_2}\right)$$

Using the trigonometric identity,

$$\sin\alpha + \sin\beta = 2\cos\frac{1}{2}(\alpha - \beta)\cdot\sin\frac{1}{2}(\alpha + \beta)$$

we find that

$$S = 2\cos\left[\frac{1}{2}(\omega_1 - \omega_2)t - \frac{1}{2}\left(\frac{\omega_1}{v_1} - \frac{\omega_2}{v_2}\right)x\right] \times \sin\left[\frac{1}{2}(\omega_1 + \omega_2)t - \frac{1}{2}\left(\frac{\omega_1}{v_1} + \frac{\omega_2}{v_2}\right)x\right]$$

$$= 2\cos\frac{1}{2}(\omega_1 - \omega_2)\left[t - \frac{x}{\frac{\frac{1}{2}(\omega_1 - \omega_2)}{\frac{1}{2}\left(\frac{\omega_1}{v_1} - \frac{\omega_2}{v_2}\right)}}\right] \times \sin\left[\frac{1}{2}(\omega_1 + \omega_2)t - \frac{1}{2}\left(\frac{\omega_1}{v_1} + \frac{\omega_2}{v_2}\right)x\right]$$

The cosine part is a wave group (the modulation imposed on the sinusoid, that part of the wave that carries the information) that moves with velocity

$$\begin{aligned} v_g &= \frac{\frac{1}{2}(\omega_1 - \omega_2)}{\frac{1}{2}\left(\frac{\omega_1}{v_1} - \frac{\omega_2}{v_2}\right)} = \frac{2\pi(f_1 - f_2)}{2\pi\left(\frac{f_1}{v_1} - \frac{f_2}{v_2}\right)} = \frac{f_1 - f_2}{\frac{1}{\lambda_1} - \frac{1}{\lambda_2}} = \frac{\left(\frac{v_1}{\lambda_1} - \frac{v_2}{\lambda_2}\right)}{\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right)} \\ &= \frac{\left(\frac{v_1}{\lambda_1} - \frac{v_1}{\lambda_2} + \frac{v_1}{\lambda_2} - \frac{v_2}{\lambda_2}\right)}{\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right)} = v_1 - \lambda_1 \frac{v_2 - v_1}{\lambda_2 - \lambda_1} \end{aligned} \tag{10.7}$$

where λ_1 and λ_2 are the corresponding signal wavelengths.

For signals with narrow bandwidths relative to the carrier frequency, such as the GNSS signals, we can replace $v_2 - v_1$ by the differential dv, $\lambda_2 - \lambda_1$ by the differential $d\lambda$, and λ_1 by λ, and add the subscript p to v to denote phase velocity explicitly to get

$$v_g = v_p - \lambda \frac{dv_p}{d\lambda} \tag{10.8}$$

which implies that the difference between the group velocity and phase velocity depends on both the wavelength and the rate of change of phase velocity with wavelength.

The corresponding indices of refraction are related by [17]

$$n_g = n_p + f \frac{dn_p}{df} \tag{10.9}$$

where the indices of refraction are defined by

$$n_p = \frac{c}{v_p} \qquad n_g = \frac{c}{v_g} \tag{10.10}$$

and f denotes the signal frequency. In a nondispersive medium, wave propagation is independent of frequency and the signal phase and signal information propagate at the same speed with $v_g = v_p$ and $n_g = n_p$.

10.2.4.1 Ionospheric Effects

The ionosphere is a dispersive medium located primarily in the region of the atmosphere between about 70 km and 1,000 km above the Earth's surface. Within this region, ultraviolet rays from the Sun ionize a portion of gas molecules and release free electrons. These free electrons influence electromagnetic wave propagation, including GNSS satellite signal broadcasts

The following is based on a similar development in [17]. The index of refraction for the phase propagation in the ionosphere can be approximated as

$$n_p = 1 + \frac{c_2}{f^2} + \frac{c_3}{f^3} + \frac{c_4}{f^4} \cdots \tag{10.11}$$

where the coefficients c_2, c_3, and c_4 are frequency-independent but are a function of the number of electrons (i.e., electron density) along the satellite-to-user signal propagation path. The electron density is denoted as n_e. A similar expression for n_g can be obtained by differentiating (10.11) with respect to frequency and substituting the result along with (10.11) into (10.9). This results in the following:

$$n_g = 1 - \frac{c_2}{f^2} - \frac{2c_3}{f^3} - \frac{3c_4}{f^4} \cdots$$

Neglecting higher-order terms, the following approximations are obtained:

$$n_p = 1 + \frac{c_2}{f^2} \qquad n_g = 1 - \frac{c_2}{f^2} \tag{10.12}$$

The coefficient c_2 is estimated as $c_2 = -40.3\ n_e\ \text{Hz}^2$. Rewriting the above yields

$$n_p = 1 - \frac{40.3 n_e}{f^2} \qquad n_g = 1 + \frac{40.3 n_e}{f^2} \tag{10.13}$$

using (10.9), the phase and group velocity are estimated as

$$v_p = \frac{c}{1 - \frac{40.3 n_e}{f^2}} \qquad v_g = \frac{c}{1 + \frac{40.3 n_e}{f^2}} \tag{10.14}$$

It can be observed that the phase velocity will exceed that of the group velocity. The amount of retardation of the group velocity is equal to the advance of the carrier phase with respect to free-space propagation. In the case of GNSS, this translates to the signal information (e.g., ranging code and navigation data) being delayed and the carrier phase experiencing an advance, a phenomenon referred to as ionospheric divergence. Importantly, the magnitude of the error on the pseudorange measurement and the error on the carrier-phase measurement (both in meters) are equal; only the sign is different. The reduction in the carrier-phase measurement value due to the presence of free electrons in the ionosphere can be intuitively explained as being due to the fact that the distance between crest to crest in the electric field of the signal is lengthened for the portion of the signal path contained within the ionosphere.

The measured range is

$$S = \int_{SV}^{User} n \, ds \tag{10.15}$$

whereas the line-of-sight (i.e., geometric) range is

$$l = \int_{SV}^{User} dl \tag{10.16}$$

The path-length difference due to ionospheric refraction is

$$\Delta S_{iono} = \int_{SV}^{User} n \, ds - \int_{SV}^{User} dl \tag{10.17}$$

and the delay attributed to the phase refractive index is

$$\Delta S_{iono,p} = \int_{SV}^{User} \left(1 - \frac{40.3 n_e}{f^2}\right) ds - \int_{SV}^{User} dl \tag{10.18}$$

Similarly, the delay induced by the group refractive index is

$$\Delta S_{iono,g} = \int_{SV}^{User} \left(1 + \frac{40.3 n_e}{f^2}\right) ds - \int_{SV}^{User} dl \tag{10.19}$$

Since the delay will be small compared to the satellite-to-user distance, we simplify (10.18) and (10.19) by integrating the first term along the line-of-sight path. Thus, *ds* changes to *dl* and we now have

$$\Delta S_{iono,p} = -\frac{40.3}{f^2}\int_{SV}^{User} n_e dl \qquad \Delta S_{iono,g} = \frac{40.3}{f^2}\int_{SV}^{User} n_e dl \tag{10.20}$$

The electron density along the path length is referred to as the *total electron content* (TEC) and is defined as

$$\text{TEC} = \int_{SV}^{User} n_e dl$$

The TEC is expressed in units of electrons/m^2 or occasionally *TEC units* (TECU) where 1 TECU is defined as 10^{16} electrons/m^2. The TEC is a function of time of day, user location, satellite elevation angle, season, ionizing flux, magnetic activity, sunspot cycle, and scintillation. It nominally ranges between 10^{16} and 10^{19} with the two extremes occurring around midnight and mid-afternoon, respectively. We can now rewrite (10.20) in terms of the TEC:

$$\Delta S_{iono,p} = \frac{-40.3\,\text{TEC}}{f^2} \qquad \Delta S_{iono,g} = \frac{40.3\,\text{TEC}}{f^2} \tag{10.21}$$

Since the TEC is generally referenced to the vertical direction through the ionosphere, the above expressions reflect the path delay along the vertical direction with the satellite at an elevation angle of 90° (i.e., zenith). For other elevation angles, we multiply (10.20) by an *obliquity factor*. The obliquity factor, also referred to as a mapping function, accounts for the increased path length that the signal will travel within the ionosphere. Various models exist for the obliquity factor. One example, from [18], is (terms are defined in Figure 10.10)

$$F_{pp} = \left[1 - \left(\frac{R_e \cos\phi}{R_e + h_I}\right)^2\right]^{-\frac{1}{2}} \tag{10.22}$$

The height of the maximum electron density, h_I, in this model is 350 km. With the addition of the *obliquity factor,* the path delay expressions from (10.21) become

$$\Delta S_{iono,p} = -F_{pp}\frac{40.3\,\text{TEC}}{f^2} \qquad \Delta S_{iono,g} = F_{pp}\frac{40.3\,\text{TEC}}{f^2}$$

Since the ionospheric delay is frequency dependent, it can virtually be eliminated by making ranging measurements with a dual-frequency receiver. Differencing pseudorange measurements made on two frequencies [e.g., B1-C/E1/L1 (1,575.42 MHz) and B2a/E5a/L5 (1,176.45 MHz)] enables the estimation of the delays on both frequencies (neglecting multipath and receiver noise errors). These are first-order estimates since they are based on (10.12). An *ionospheric-free* pseudorange may be formed using pseudorange measurements on two frequencies as prescribed in [2] for GPS L1 and L2:

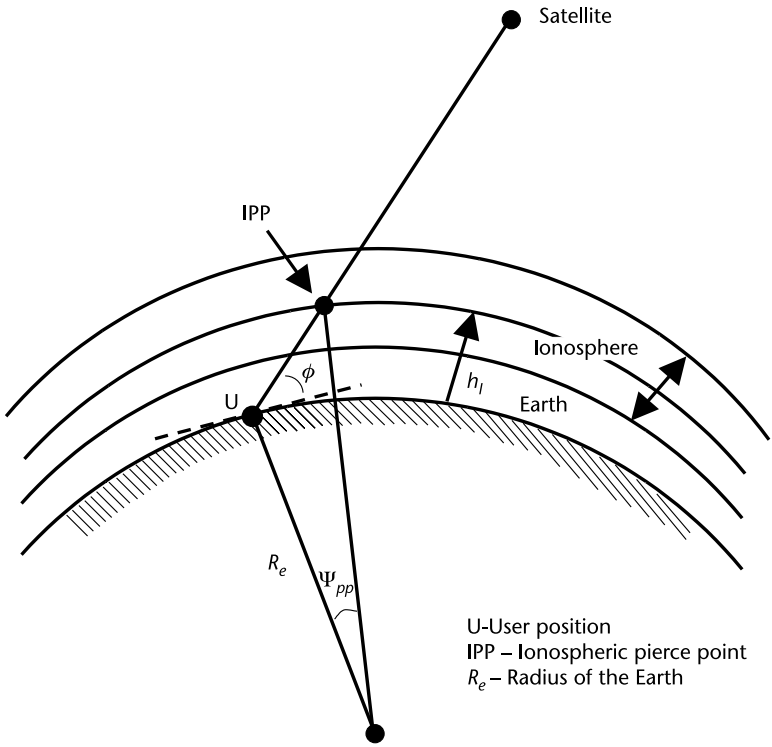


Figure 10.10 Ionospheric modeling geometry.

$$\rho_{ionospheric-free} = \frac{\rho_{L2} - \gamma\rho_{L1}}{1-\gamma} \tag{10.23}$$

where $\gamma = (f_{L1}/f_{L2})^2$. Although ionospheric delay errors are removed, this approach has the drawback that measurement errors are significantly magnified through the combination. A preferred approach is to use the L1 and L2 pseudorange measurements to estimate the ionospheric error on L1 using the following expression:

$$\Delta S_{iono,corr_{L1}} = \left(\frac{f_{L2}^2}{f_{L2}^2 - f_{L1}^2}\right)(\rho_{L1} - \rho_{L2}) \tag{10.24}$$

The path length difference on L2 can be estimated by multiplying $\Delta S_{iono,corr_{L1}}$ by

$$(f_1 / f_2)^2 = (77/60)^2$$

These estimated corrections may be smoothed over time, since ionospheric delay errors typically do not change very rapidly and are subtracted from pseudorange measurements made by each frequency.

It should be noted that the higher-order terms in (10.11) usually account for differences at the millimeter level (rising to centimeter level during extreme ionospheric disturbances) and may be safely neglected for most applications [19].

In the case of a single-frequency receiver, it is obvious that (10.24) cannot be used. Consequently, models of the ionosphere are employed to correct for the ionospheric delay.

Klobuchar Model

One important example is the Klobuchar model used in GPS and other SATNAV systems, which removes (on average) about 50% of the ionospheric delay at mid-latitudes through a set of coefficients included in a GNSS navigation message. This model assumes that the vertical ionospheric delay can be approximated by half a cosine function of the local time during daytime and by a constant level during nighttime [20].

Almost three times as much delay is incurred when viewing satellites at low elevation than at the zenith. For a signal arriving at vertical incidence, the delay ranges from about 10 ns (3m) at night to as much as 50 ns (15m) during the day. At low satellite viewing angles (0° through 10°), the delay can range from 30 ns (9m) at night up to 150 ns (45m) during the day [21]. Reference [22] stated that the value for residual ionospheric delays, averaged over the globe ranges from 9.8m to 19.6m.

NeQuick G

The NeQuick G model has been developed for use in Galileo UE and is based on a 3-D representation of the electron density using an adaptation of the NeQuick ionospheric electron density model for quasi-real-time corrections and driven by three broadcast coefficients in the navigation message. NeQuick G is designed to reach a correction capability of at least 70% of the ionospheric code delay (rms), with a lower slant TEC (STEC) residual error bound of 20 TECU for any location, time of day, season, and solar activity, excluding periods where the ionosphere is largely disturbed due to, for instance, geomagnetic storms. Such performance has been assessed successfully using GPS data only and GPS+GIOVE data during GIOVE Experimentation [23].

Figure 10.11 shows the global daily RMS ionospheric residual error in meters at L1 after correction with the Galileo NeQuick G model and GPS Ionospheric Correction Algorithm from April 2013 to March 2016. It can be observed that the NeQuick G model residual error is less than those obtained by the Klobuchar model (based on [19]).

Spatial Correlation

The following relationship between the delay, ε^{Iono}, expressed in units of length, due to the ionosphere, the frequency, f, of the signal, the elevation angle, ϕ', at the ionospheric pierce point, and the total electron content, TEC, along the path of the signal is:

$$\varepsilon^{Iono} = \frac{1}{\sin\phi'} \cdot \frac{40.3}{f^2} \cdot TEC$$

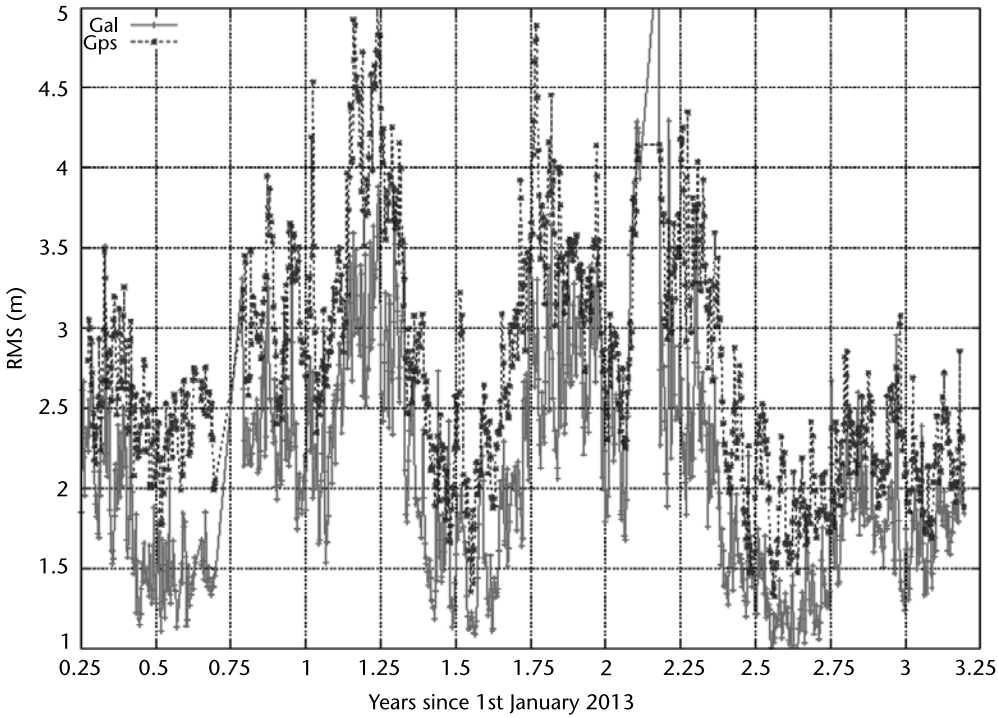


Figure 10.11 Global daily RMS ionospheric residual error in meters at L1 after correction with Galileo NeQuick G and GPS ICA from April 2013 to March 2016. (Courtesy of ESA/Raul Orus.)

The sin ϕ' term accounts for the additional path length in the ionosphere when the direction of the satellite is off the vertical. The *ionospheric pierce point* is that point on the displacement vector from the user position to the satellite position midway through the ionosphere typically taken to be 300 km to 400 km in altitude [17] (see Figure 10.12).

The difference in delay due to the difference in elevation angles for a horizontal separation of user and reference station is

$$\begin{aligned}\left|\varepsilon_u^{Iono} - \varepsilon_m^{Iono}\right| &= \frac{1}{\sin\phi'} \cdot \frac{40.3}{f^2} \cdot TEC - \frac{1}{\sin\phi'_m} \cdot \frac{40.3}{f^2} \cdot TEC \\ &= \left|\frac{1}{\sin\phi'} - \frac{1}{\sin\phi'_m}\right| \cdot \frac{40.3}{f^2} \cdot TEC \\ &\frac{p}{d_m} \cdot \left|\frac{p}{d_m} - \cos\phi'_m\right| \cdot \frac{40.3}{f^2} \cdot TEC\end{aligned} \tag{10.25}$$

where p = distance between the user and the reference station, ϕ_m = elevation angle of the satellite from the reference station, and ϕ'_m = elevation angle at the reference station's ionospheric pierce point.

The *TEC* usually lies in the range 10^{16} to 10^{18} electrons/m^2, with 50×10^{16} electrons/m^2 typical in the temperate zones, so that the difference in delays experienced by the reference station and the user 100 km away due to the difference in elevation angle is typically

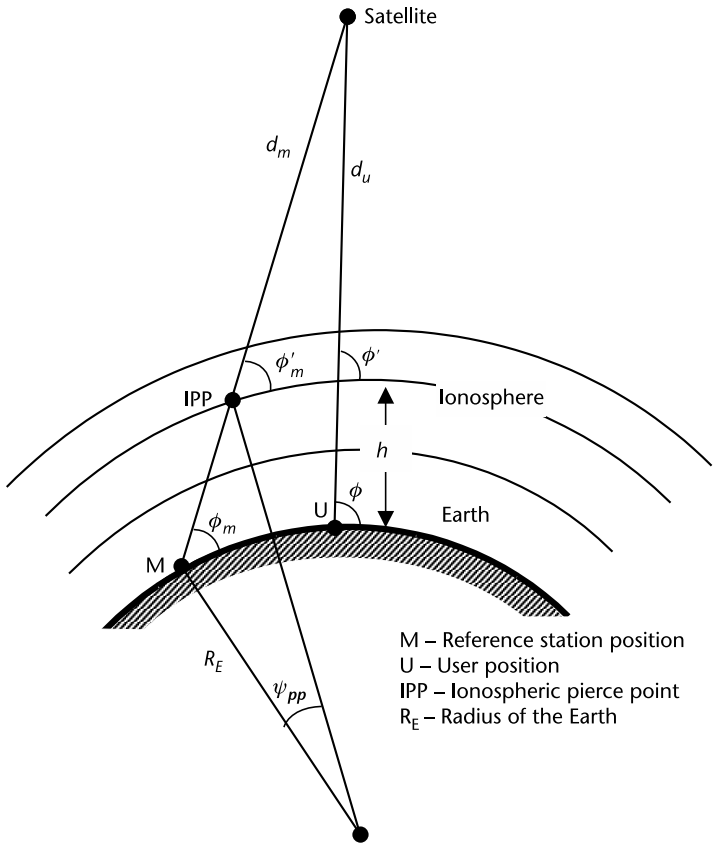


Figure 10.12 Ionospheric delay difference.

$$\left|\varepsilon_u^{Iono} - \varepsilon_m^{Iono}\right| = \left|\frac{p}{d_m}\cdot\left(\frac{p}{d_m} - \cos\phi'_m\right)\cdot\frac{40.3}{f^2}\cdot TEC\right|$$

$$\approx \left|-\frac{100\,\text{km}}{2\times10^4\,\text{km}}\cdot\cos 45^\circ\cdot\frac{40.3}{\left(1.575\times10^9\right)^2}\cdot 50\times10^{16}\right|$$

$$= 0.03\,\text{m}$$

The variation of the ionospheric delay difference due to differences in elevation angle as a function of separation is shown in Figure 10.13 for three values of satellite elevation angle and a TEC of 50×10^{16} electrons/m^2.

Spatial variations in TEC within the ionosphere typically lead to much greater differences in ionospheric delay than those attributable to elevation angle. The difference in vertical ionospheric delays (i.e., delays observed for a satellite that is directly overhead) due to TEC gradients is typically in the range of 0.2 to 0.5m over 100 km when the ionosphere is undisturbed, but can be greater than 4m over 100 km when the ionosphere is disturbed [24, 25]. Slant range delays during daylight hours were evaluated in [26] for a network of GPS receivers over a 1-year timeframe. The conclusions from [26] were that the difference in ionospheric delays seen by two receivers separated by 400 km in a mid-latitude region is expected to be less than 2m in magnitude 95% of the time even during the peak of the 11-year

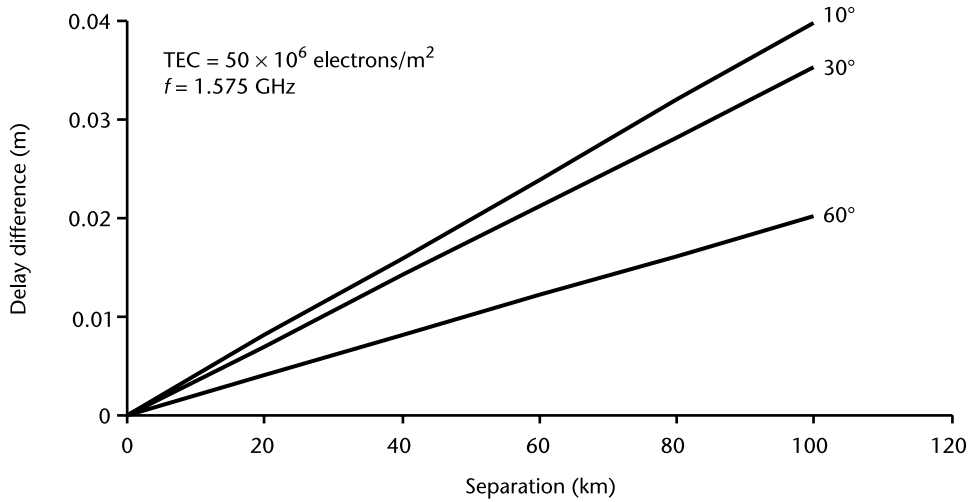


Figure 10.13 Variation of ionospheric delay difference due to elevation angle differences.

solar cycle. There are various physical phenomena, including *traveling ionospheric disturbances* (TIDS), which are small-scale irregularities in the ionosphere, that can cause steep spatial gradients in TEC over distances as short as 10 km.

Temporal Correlation

Ionospheric delays typically change very slowly with time, normally following a daily cycle of very low values at local nighttime, followed by a ramping up to a maximal delay in the early local afternoon, and then a decline back to the steady night value again. In mid-latitude regions, the time rate of change of vertical ionospheric delays rarely exceeds 8 cm/min [27]. In other regions of the world, rates of up to 65 cm/min have been observed [27]. Some recent studies have indicated that rates of over 3 m/min may occur on rare occasions. These observed rates include both the effect of changing elevation angles and TEC.

10.2.4.2 Tropospheric Delay

The troposphere is the lower part of the atmosphere that is nondispersive for frequencies up to 15 GHz [17]. Within this medium, the phase and group velocities associated with the GNSS carrier and signal information (ranging code and navigation data) on the GNSS L-band frequencies are equally delayed with respect to free-space propagation. (Note that S-band carrier and signal information are also equally delayed but this discussion focuses on L-band signals.) This delay is a function of the tropospheric refractive index, which is dependent on the local temperature, pressure, and relative humidity. Left uncompensated, the range equivalent of this delay can vary from about 2.4m for a satellite at the zenith and the user at sea level to about 25m for a satellite at an elevation angle of approximately 5°[17].

From (10.17), we have that the path length difference attributed to the tropospheric delay as

$$\Delta S_{tropo} = \int_{sv}^{user} (n-1)ds$$

where the integration is along the signal path. The path-length difference can also be expressed in terms of refractivity,

$$\Delta S_{tropo} = 10^{-6} \int_{sv}^{user} N ds \tag{10.26}$$

where the refractivity,, is defined by

$$N \equiv 10^6 (n-1)$$

The refractivity is often modeled as including both a dry (hydrostatic) and wet (nonhydrostatic) component [28]. The dry component, which arises from the dry air, gives rise to about 90% of the tropospheric delay and can be predicted very accurately. The wet component, which arises from the water vapor, is more difficult to predict due to uncertainties in the atmospheric distribution. Both components extend to different heights in the troposphere; the dry layer extends to a height of about 40 km while the wet component extends to a height of about 10 km.

We define $N_{d,0}$ and $N_{w,0}$ as the dry and wet component refractivities, respectively, at standard sea level. To express both $N_{d,0}$ and $N_{w,0}$ in pressure and temperature, the formulas of [29] can be used:

$$N_{d,0} \approx a_1 \frac{p_0}{T_0}$$

with p_0 = partial pressure of the dry component at standard sea level (mbar), T_0 = absolute temperature at standard sea-level (K), and a_1 = empirical constant (77.624 K/mbar).

$$N_{w,0} \approx a_2 \frac{e_0}{T_0} + a_3 \frac{e_0}{T_0^2}$$

where a_2 and a_3 are empirical constants (–12.92 K/mbar and 371,900 K^2/mbar, respectively).

Path delay also varies with the user's height, h. Thus, both the dry and wet component refractivities are dependent on the atmospheric conditions at the user's height above the reference ellipsoid. One model that takes the height into account and is successfully demonstrated in [30], combines parts of the works cited in [28, 29, 31, 32]. The dry component as a function of the height is determined by

$$N_d(h) = N_{d,0} \left[\frac{h_d - h}{h_d} \right]^{\mu} \tag{10.27}$$

and h_d, the upper extent of the dry component of the troposphere referenced to sea level, is determined from

$$h_d = 0.011385 \frac{p_0}{N_{d,0} \times 10^{-6}}$$

where μ stems from the underlying use of the ideal gas law. Hopfield [28] found that setting $\mu = 4$ gives the best results for the model.

Similarly, the refractivity, $N_w(h)$, of the wet component of the troposphere is determined from

$$N_w(h) = N_{w,0} \left[\frac{h_w - h}{h_w} \right]^{\mu} \tag{10.28}$$

where h_w is the extent of the wet component of troposphere determined by

$$h_w = 0.0113851 \frac{1}{N_{w,0} \times 10^{-6}} \left[\frac{1,255}{T_0} + 0.05 \right] e_0$$

The path-length difference when the satellite is at zenith and the user is at sea level is from (10.26):

$$\Delta S_{tropo} = 10^{-6} \int_{h=0}^{h_d} N_d(h) dh + 10^{-6} \int_{h=0}^{h_d} N_d(h) dh \tag{10.29}$$

Evaluation of (10.29) using the expressions for $N_d(h)$ and $N_w(h)$ in (10.27) and (10.28) yields

$$\begin{aligned} \Delta S_{tropo} &= \frac{10^{-6}}{5} \left[N_{d,0} h_d + N_{w,0} h_w \right] \\ &= d_{dry} + d_{wet} \end{aligned} \tag{10.30}$$

To compute the tropospheric correction in (10.30), pressure and temperature inputs are required, which can be obtained using meteorological sensors. When the satellite is not at zenith, a mapping function model is needed to determine how much greater of a delay can be anticipated due to the larger path length of the signal through the troposphere. It is common to refer to the delay for a satellite at zenith as a *vertical delay* or *zenith delay* and the delay for satellites at any other arbitrary elevation angle as a *slant delay*. Mapping functions that relate slant and vertical delays will be discussed later in this section.

One accurate method for modeling the troposphere's dry and wet components at zenith without meteorological sensors was developed at the University of New Brunswick (UNB). In this model [33], referred to as UNB3, the dry and wet

components are considered functions of height, h, in meters above mean sea level and of five meteorological parameters: pressure, p, in millibars, temperature, T, in Kelvin, water vapor pressure, e, in millibars, temperature lapse rate, β, in K/m, and water vapor lapse rate, λ (unitless). Each of the meteorological parameters is calculated by interpolating values from Tables 10.1 and 10.2. Using pressure as an example, the average pressure, $p_0(\phi)$, at latitude ϕ ($15° < \phi < 75°$) is calculated by using the two values in the p_0 column of Table 10.1 corresponding to those two values of latitude, ϕ_i and ϕ_{i+1}, that are closest to ϕ, as follows:

$$p_0(\phi) = p_0(\phi_i) + \left[p_0(\phi_{i+1}) - p_0(\phi_i)\right] \cdot \frac{(\phi - \phi_i)}{(\phi_{i+1} - \phi_i)}$$

Similarly, the seasonal variation, $\Delta p(\phi)$, is found in the same way from Table 10.2, as follows:

$$\Delta p(\phi) = \Delta p(\phi_i) + \left[\Delta p(\phi_{i+1}) - \Delta p(\phi_i)\right] \cdot \frac{(\phi - \phi_i)}{(\phi_{i+1} - \phi_i)}$$

For latitudes less than 15° simply use the values of parameters in the first row without interpolation; for latitudes greater than 75°, use the values of parameters in the last row. Finally, the pressure, p, is determined, taking into account the day of the year, D, with the first day being January 1, as follows:

$$p = p_0(\phi) - \Delta p(\phi) \cdot \cos\left[\frac{2\pi(D - D_{\min})}{365.25}\right]$$

Table 10.1 Average Meteorological Parameters for Tropospheric Delay

Parameter Averages					
Latitude (°)	*p_0 (mbar)*	*T_0 (K)*	*e_0 (mbar)*	*β_0 (K/m)*	*λ_0*
15° or less	1,013.25	299.65	26.31	6.30×10^{-3}	2.77
30	1,017.25	294.15	21.79	6.05×10^{-3}	3.15
45	1,015.75	283.15	11.66	5.58×10^{-3}	2.57
60	1,011.75	272.15	6.78	5.39×10^{-3}	1.81
75° or greater	1,013.00	263.65	4.11	4.53×10^{-3}	1.55

Table 10.2 Seasonal Meteorological Parameters for Tropospheric Delay

Seasonal Variation of Parameters					
Latitude (°)	*Δp (mbar)*	*ΔT (K)*	*Δe (mbar)*	*$\Delta\beta$ (K/m)*	*$\Delta\lambda$*
15° or less	0.00	0.00	0.00	0.00×10^{-3}	0.00
30	–3.75	7.00	8.85	0.25×10^{-3}	0.33
45	–2.25	11.00	7.24	0.32×10^{-3}	0.46
60	–1.75	15.00	5.36	0.81×10^{-3}	0.74
75 or greater	–0.50	14.50	3.39	0.62×10^{-3}	0.30

where

$$D_{\min} = \begin{cases} 28 & \text{in northern latitudes} \\ 211 & \text{in southern latitudes} \end{cases}$$

The difference in the values of the parameter D_{min} in the Northern and Southern hemispheres accounts for the difference (183 days) in seasons in these hemispheres. Once all five meteorological parameters have been calculated in exactly the same way that the pressure is calculated, the wet and dry components of the delay can be determined from the following equations for d_{dry} and d_{wet} in (10.27):

$$d_{dry} = \left(1 - \frac{\beta \cdot h}{T}\right)^{\frac{g}{R_d \beta}} \cdot \left(\frac{10^{-6} k_1 R_d p}{g_m}\right)$$

$$d_{wet} = \left(1 - \frac{\beta \cdot h}{T}\right)^{\frac{(\lambda+1)g}{R_d \beta} - 1} \cdot \left(\frac{10^{-6} k_2 R_d}{g_m (\lambda + 1) - \beta R_d} \cdot \frac{e}{T}\right)$$

where $k_1 = 77.604$ K/mbar, $k_2 = 382000$ K^2/mbar, $R_d = 287.054$ J/kg/K, $g_m = 9.784$ m/s^2, and g = 9.80665 m/s^2.

For elevation angles other than 90°, the model in (10.30) does, in general, not apply. To account for, for example, the elevation angle of the satellite, so-called mapping functions may be introduced in:

$$\begin{aligned} &\Delta S_{tropo} = m_d \cdot d_{dry} + m_w \cdot d_{wet} \\ &\text{or} \\ &\Delta S_{tropo} = m \cdot \left(d_{dry} + d_{wet}\right) \end{aligned} \tag{10.31}$$

where m_d = dry-component mapping function, m_w = wet-component mapping function, and m = general mapping function.

Existing mapping functions can be divided into two groups: the geodetic-survey oriented applications and the navigation oriented applications [34]. An example of the geodetic-survey oriented group are the Niell mapping functions as described in [35], with separate mapping functions for dry and wet components of the tropospheric delay. Navigation-oriented mapping functions include both analytical models and more complex forms such as the fractional form introduced by [36]. The advantage of the analytical forms is that it is not computationally intensive to determine the mapping function values. An example of analytical models is Black and Eisner's mapping function which is a function of the satellite's elevation angle, E:

$$m(E) = \frac{1.001}{\sqrt{0.002001 + \sin^2(E)}}$$

A more accurate, but more complex model that may be used for the mapping function has the following continued fractional form [36]:

$$m_i(E) = \cfrac{1 + \cfrac{a_i}{1 + \cfrac{b_i}{1 + \cfrac{c_i}{1 + \ldots}}}}{\sin E + \cfrac{a_i}{\sin E + \cfrac{b_i}{\sin E + \cfrac{c_i}{\sin E + \ldots}}}}$$

where E is the elevation angle, a_i, b_i, and c_i are the mapping function parameters, and i represents either the dry or wet component. Note that the term in the numerator normalizes the mapping function with respect to zenith. The parameters a_i, b_i, and c_i can be estimated from ray-tracing delay values at various elevation angles. Examples of mapping functions that describe the troposphere delay accurately down to a satellite elevation angle of 2° are described in [34]. The models in [34] are a function of satellite elevation angle and height. Note that these models require more computation time than the analytical models. One example of a three parameter continued fractional form is the UNBabc model. The a, b, and c parameters for the dry component mapping function are given by:

$$\begin{aligned} a_d &= \left(1.18972 - 26.855h + 0.10664\cos\phi\right)/1000 \\ b_d &= 0.0035716 \\ c_d &= 0.082456 \end{aligned}$$

The a, b, and c parameters for the wet component mapping function are given by:

$$\begin{aligned} a_w &= \left(0.61120 - 35.348h - 0.01526\cos\phi\right)/1000 \\ b_w &= 0.0018576 \\ c_w &= 0.062741 \end{aligned}$$

Spatial Correlation

As discussed above, the speed of electromagnetic radiation varies, depending on temperature, pressure, and relative humidity, as it passes through the troposphere. In this section, we obtain an estimate of the kind of delay difference we can expect from the signal traveling through the troposphere and choose a model described in [37], which expresses the tropospheric delay of a signal from a GNSS satellite to a user at the Earth's surface, as follows:

$$\begin{aligned} \varepsilon_u^{Tropo} &= \csc\phi \cdot \left(1.4588 + 0.0029611 \cdot N_s\right) \\ &\quad -0.3048 \cdot \left[0.00586 \cdot \left(N_s - 360\right)^2 + 294\right] \cdot \phi^{-2.30} \end{aligned} \tag{10.32}$$

where ε_u^{Tropo} = tropospheric delay experienced by the user in meters, ϕ = elevation angle from the user to the satellite in degrees, and N_s = surface refractivity.

If we denote the elevation angle of the satellite from the reference station by ϕ_m, then from Figure 10.14, we can determine the difference $\csc\phi - \csc\phi_m$ in terms of the horizontal distance p between the user and reference station and the height d_s of the satellite, as follows:

$$\left|\csc\phi - \csc\phi_m\right| = \left|\frac{d_u}{d_s} - \frac{d_m}{d_s}\right| = \left|\frac{d_u - d_m}{d_s}\right| \le p \cdot \frac{\cos\phi_m}{d_s}$$

where d_m is the distance from the monitoring station to the satellite and d_u is the distance from the user receiver to the satellite. (The inequality sign may be dropped if the triangle lies in a vertical plane.) This yields the following equation for the delay difference where, for the moment, we hold N_s constant:

$$\begin{aligned}\left|\varepsilon_u^{Tropo} - \varepsilon_m^{Tropo}\right| \le\ & p \cdot \frac{\csc\phi_m}{d_s} \cdot \left(1.4588 + 0.0029611 \cdot N_s\right) \\ & -0.3048 \cdot \left[0.00586 \cdot \left(N_s - 360\right)^2 + 294\right] \cdot \left(\phi^{-2.30} - \phi_m^{-2.30}\right)\end{aligned} \tag{10.33}$$

The second term of the right member in (10.33) was added to fit data at low elevation angles—about 10° or less—and is negligible for higher GNSS elevation angles (i.e., greater than 10°). For higher elevation angles, the difference in tropospheric delay error is proportional to the separation between the user and reference station.

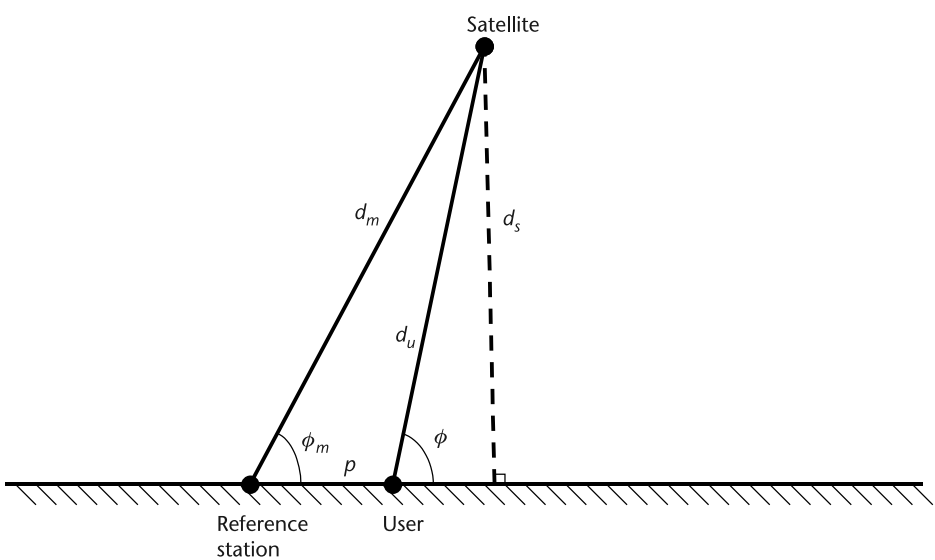


Figure 10.14 Horizontal tropospheric delay difference.

Suppose, for example, that the elevation angle is 45° and $p = 100$ km. Then, if we use a midrange value for N_s of 360, we find from the model that the deviation of the tropospheric correction at the user position differs from that at the reference station position by an amount

$$\begin{aligned}\left|\varepsilon_u^{Tropo} - \varepsilon_m^{Tropo}\right| &\le p \cdot \frac{\csc \phi_m}{d_s} \cdot \left(1.4588 + 0.0029611 \cdot N_s\right) \\ &= (100\text{km}) \cdot \frac{\csc 45^\circ}{2 \times 10^4 \text{km}} \times \left(1.4588 + 0.0029611 \times 360\right) \\ &\approx 0.02 \text{ m}\end{aligned}$$

Thus, the error is in the order of 2 cm. The variation of the deviation as a function of separation due to elevation angle differences is shown in Figure 10.15. Note that over the entire 100-km separation, the variation of delay difference due to variation in the surface refractivity is less than 1.5 mm for this tropospheric model, an order of magnitude smaller than that due to a variation in elevation angle from 10° to 90°. Thus, allowing N_s to vary in the derivation of (10.33) would have produced a small, negligible additional term in (10.33). However, the total delay difference is also small. Even for extreme values of refractivity (400) and low angles (10°), the differences in delays are not much more than 2 cm.

Real-world data suggests that tropospheric delays vary more rapidly with distance than can be attributed solely to differences in viewing angle. Much larger differences in tropospheric delay from location to location arise in reality because the troposphere often differs significantly from the model, especially at an interface between land and water or where the user and reference station are separated by a weather front. In a study described in [38], differences in tropospheric delays as

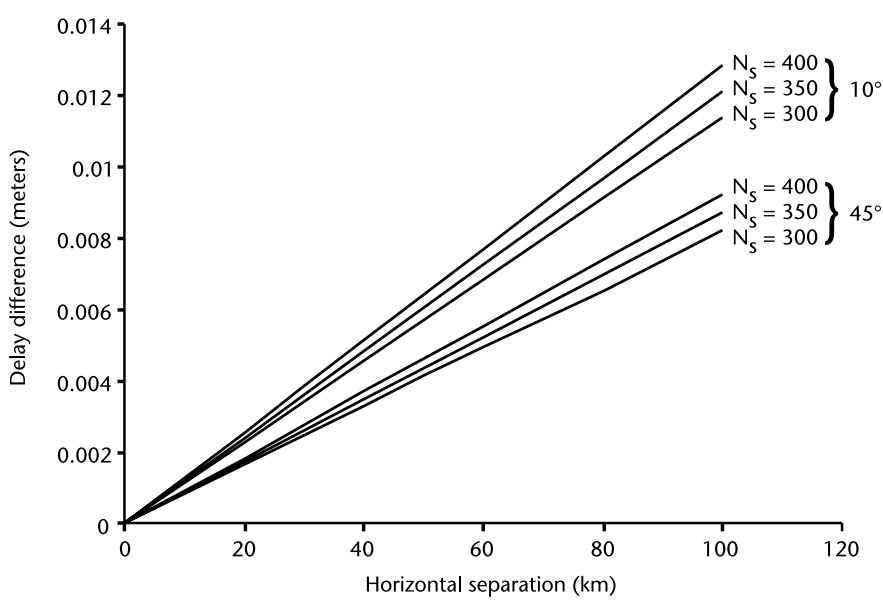


Figure 10.15 Variation in tropospheric delay difference due to elevation angle.

large as 40 cm were observed over a 25-km baseline for satellites above 5°. This suggests a vertical tropospheric delay difference on the order of 4 cm over this baseline; a much larger value than could be attributed to differences in viewing angle alone. The smallest difference in tropospheric error observed in [38] over a 25-km baseline was 10 cm.

A difference in heights between the user receiver and the reference station has a greater effect than a horizontal displacement. Reference [37] develops the following relationship between the tropospheric delay, ε_m^{Tropo} meters, experienced by the reference station and delay, ε_h^{Tropo} meters, experienced by a user at a height h kilometers above the station (Figure 10.16):

$$\varepsilon_h^{Tropo} = \varepsilon_m^{Tropo} \cdot e^{-\left[(0.0002N_s+0.07)\cdot h+\left(\frac{0.83}{N_s}-0.0017\right)\cdot h^2\right]}$$

At an altitude of 1 km above the reference station, the user experiences a delay of

$$\begin{aligned}\varepsilon_h^{Tropo} &= \varepsilon_m^{Tropo} \cdot e^{-\left[(0.0002\times 360+0.07)\cdot 1+\left(\frac{0.83}{360}-0.0017\right)\cdot 1^2\right]} \\ &= 0.45\times\varepsilon_m^{Tropo} = 0.45\times 3.6\text{ m} = 1.6\text{ m}\end{aligned}$$

and the difference in delays is

$$\varepsilon_h^{Tropo} - \varepsilon_m^{Tropo} = 3.6\,\text{m} - 1.6\,\text{m} = 2\,\text{m}$$

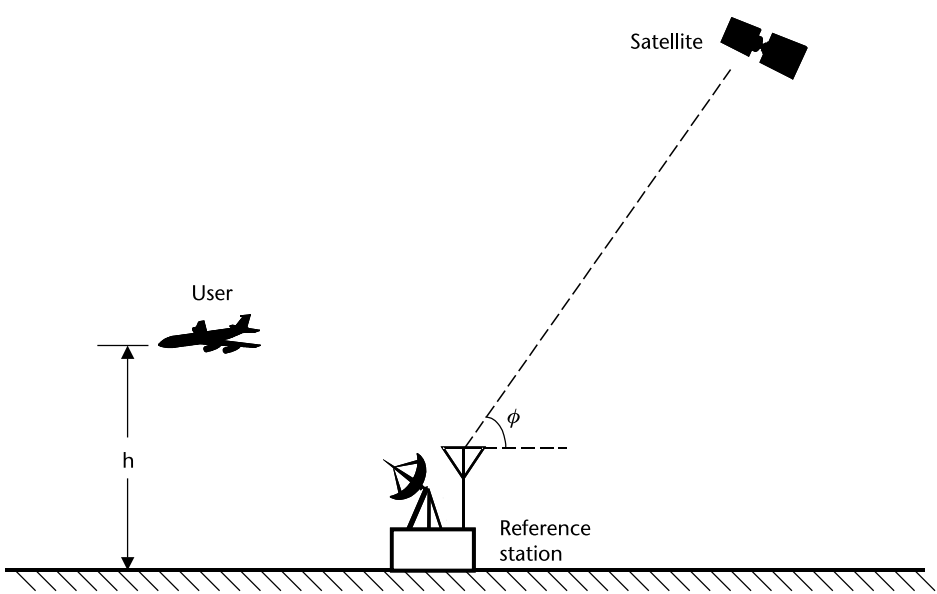


Figure 10.16 Vertical tropospheric delay difference.

That is, assuming $N_S = 360$ and that the elevation angle is 45°, the delay at a height of 1 km is only 45% of the 3.6-m delay, calculated from (10.32), at the reference station, or 1.6m. The difference is 2m.

The variation in the difference in tropospheric delays between a signal reaching the ground having a refractivity of N_S and the signal at an altitude h above the ground is shown in Figure 10.17 for two different elevation angles of the satellite.

Temporal Correlation

Although vertical tropospheric delays do not change very rapidly with time for a stationary receiver, slant tropospheric delays can due to the rate of change of elevation angle. For stationary users, the elevation angle to a GNSS satellite can vary at a rate up to 0.5°/min, just due to the motion of the satellite. For a satellite at 5°, this can lead to tropospheric delay changing at a rate of up to 2 m/min. For satellites above 10°, the maximum rate of change is around 0.64 m/min. A receiver on a moving platform that is rapidly changing altitude can experience an even higher rate of change in tropospheric error due to the altitude dependence discussed above.

10.2.5 Receiver Noise and Resolution

Measurement errors are also induced by the receiver tracking loops. In terms of the DLL, dominant sources of pseudorange measurement error (excluding multipath, which will be discussed in Section 10.2.6) are thermal noise jitter and the effects of interference. For example, the composite receiver noise and resolution error contribution for a BPSK-R(1) signal will be slightly larger than that for a BPSK-R(10) because the BPSK-R(1) signal has a smaller root-mean-square bandwidth than the BPSK-R(10). Typical modern receiver 1σ values for the noise and resolution error are on the order of a decimeter or less in nominal conditions (i.e., without external

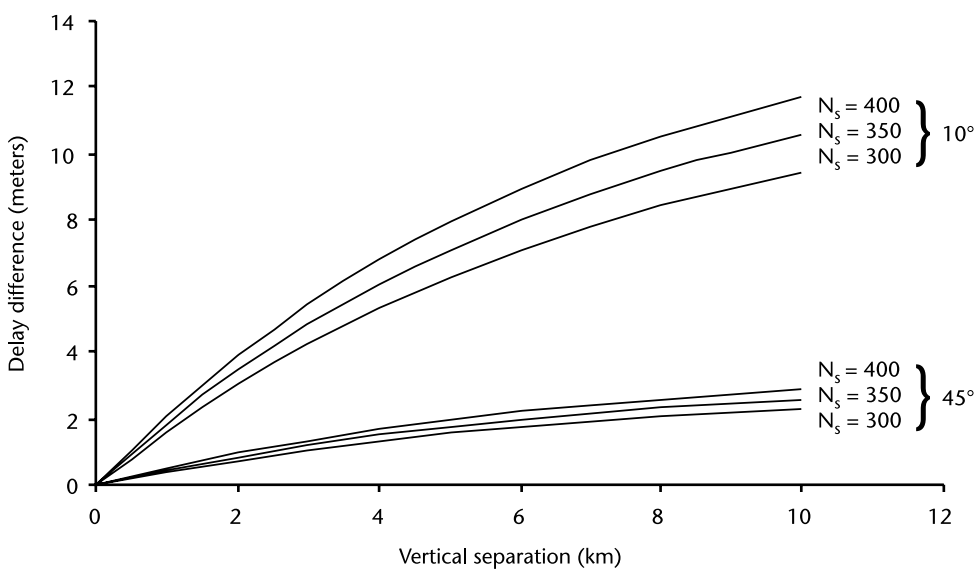


Figure 10.17 Variation in the vertical delay difference with refractivity and elevation angle.

interference) and negligible compared to errors induced by multipath. Receiver noise and resolution errors affect carrier phase measurements made by a PLL. For the BPSK-R(1) and BPSK-R(10) signals mentioned above, PLL measurements errors in nominal conditions are on the order of 1.2 mm (1 σ) when tracking the BPSK-R(1) signal and 1.6 mm (1 σ) when tracking the BPSK-R(10) signal. Extensive treatment of DLL and PLL errors is provided in Section 8.9. The effects of interference on DLLs and PLLs are discussed in Section 9.2.

10.2.6 Multipath and Shadowing Effects

One of the most significant errors incurred in the receiver measurement process is multipath. Multipath errors on both pseudorange and carrier-phase measurements were discussed in detail in Section 9.5. As described in that section, multipath errors vary significantly in magnitude depending on the environment the receiver is located within, satellite elevation angle, receiver signal processing, antenna gain pattern, and signal characteristics. Within this chapter, as an example, we will use typical one-sigma multipath levels in a relatively benign environment of 20 cm and 2 cm, respectively, for a wide bandwidth BPSK-R(1) signal receiver's pseudorange and carrier-phase measurements.

Receiver Noise and Multipath in DGNSS Systems

Unlike the other error sources considered thus far, receiver noise and multipath result in pseudorange and carrier-phase errors that are uncorrelated between receivers separated by even very short baselines. Multipath, in particular, often dominates error budgets for short-baseline code- and carrier-based DGNSS systems for two reasons. First, it causes pseudorange and carrier-phase errors that are generally statistically larger than those caused by receiver noise. Second, the fact that multipath errors are uncorrelated from receiver to receiver means that the difference in measurement error caused by multipath between two receivers has a variance described as the sum of the multipath error variance attributable to each alone. As discussed in Section 9.5, the magnitude of multipath errors varies significantly depending on the type of receiver and environment.

Both receiver noise and multipath errors can change very rapidly. Since these errors are not common between the user and reference station in a DGNSS scenario, the rates of change of these errors are only important in that averaging of some form within the user equipment can often be employed to reduce their consequence.

10.2.7 Hardware Bias Errors

10.2.7.1 Satellite Biases

Each of the ranging codes generated onboard the SV experiences a different delay from signal generation to output from the antenna phase center. This delay is due to the different analog and digital signal paths corresponding to each signal. This delay is defined as the *equipment group delay* and consists of a bias and an uncertainty

term [2, 4, 6, 7]. Thus, all SV generated signals have a *unique offset* from GNSS system time and are transmitted at a different time. This can be observed for the GPS example in Figure 10.18(a). Please note that while a GPS construct is used to discuss this topic other SATNAV systems may also employ this construct.

Equation (10.20) shows that the ionosphere delays each code inversely proportional to the square of its frequency. Keeping this in mind and staying with the GPS construct as an example, if both the L1 P(Y) and L2 P(Y) signals experienced identical equipment group delays and were transmitted at exactly the same time, the ionospheric-free equation (10.22) would generate the ionospheric-free pseudorange, which would have the same effective transmission time as the L1 P(Y) and L2 P(Y) signals and only the ionospheric delay would be removed. This is depicted in Figure 10.18(b).

However, the L1P(Y) and L2P(Y) signals are not transmitted at the same time. The ionospheric-free equation (10.22) will remove the ionospheric delay. The effective transmission time of this new composite ionospheric-free pseudorange will be a different time from either the L1P(Y) or the L2P(Y) transmission times but mathematically related to the difference between the two. This is depicted in Figure 10.18(c).

Different codes on different frequencies can be mathematically combined to remove the effects of the ionosphere. Each ionospheric-free ranging code pair has a different effective transmission time. However, each ionospheric-free ranging code pair effective transmission time is offset from GNSS system time. Thus, all ranging code pairs are offset from GNSS system time by some amount.

In keeping with the GPS example, the GPS CS uses the L1 P(Y) and L2 P(Y) code pair exclusively to compute the SV clock offset and ephemeris parameters (Section 3.3.1.4). Within GPS, the L1 P(Y) - L2 P(Y) ionospheric-free ranging code pair is currently the reference for all the SV clock and ephemeris calculations. The satellite clock bias (10.3) which is calculated from the a_{f0}, a_{f1} and a_{f2} terms is the best estimate of the difference between GPS system time and the L1 P(Y) to L2 P(Y) ionospheric-free ranging code pair effective transmission time.

The group differential delay, T_{GD}, is the difference between the L1 P(Y) to L2 P(Y) ionospheric-free ranging code pair effective transmission time and the L1 P(Y) transmission time. T_{GD} is mathematically related to the difference between the L2 P(Y) transmission time and the L1 P(Y) transmission time because of both ranging codes are used to generate the L1 P(Y) to L2 P(Y) ionospheric-free ranging code.

The intersignal correction (ISC) for any ranging code is the difference between the L1 P(Y) transmission time and the ranging code transmission time. By combining the clock bias terms (a_{f0}, a_{f1}, and a_{f2}), T_{GD}, and the respective ISC, the transmission time of any code or code pair relative to GPS time can be calculated. This is shown in Figure 10.18(d).

GNSS satellite antennas also can cause bias errors. The apparent antenna phase and group delay centers move slightly as a function of off-boresight angle. This effect can result in pseudorange biases of up to ±0.5m and carrier-phase biases of up to ±2 cm. Estimates of these biases are produced by the International GNSS Service (IGS) and freely available on the Internet in a file format referred to as the Antenna Exchange Format (ANTEX).

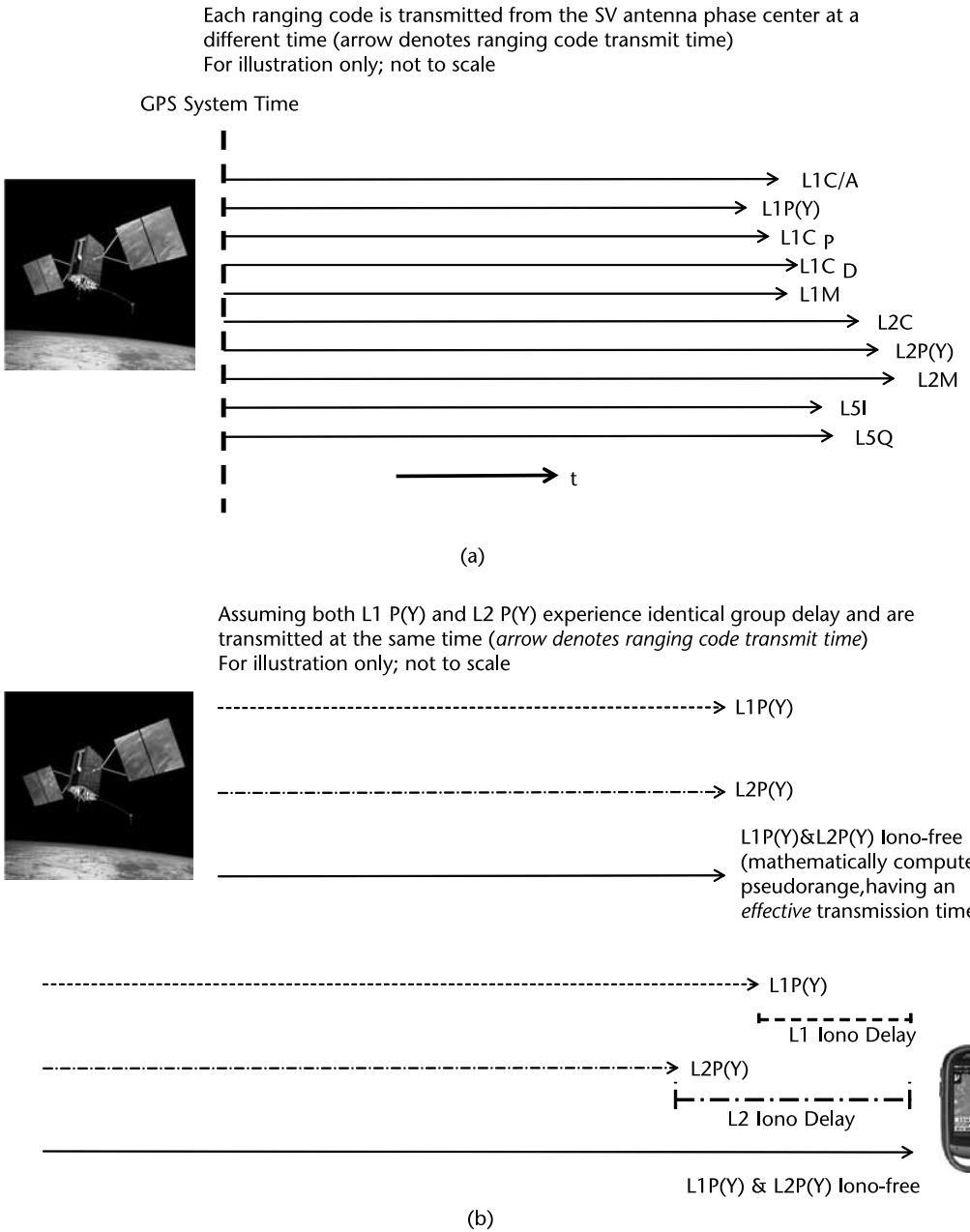
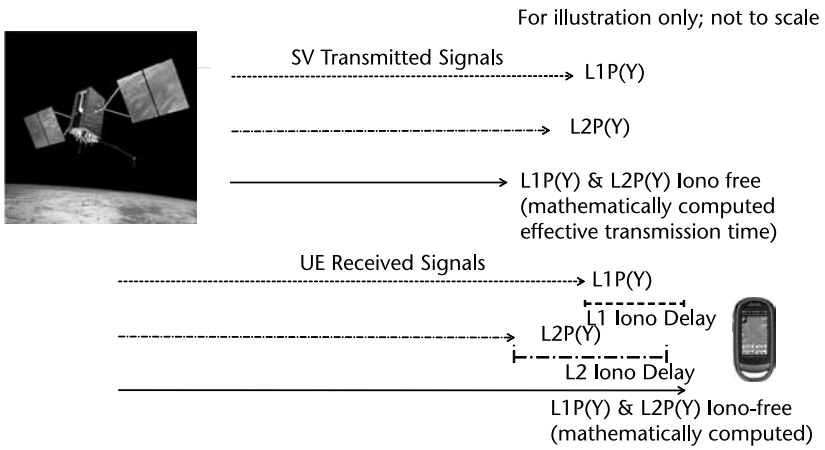


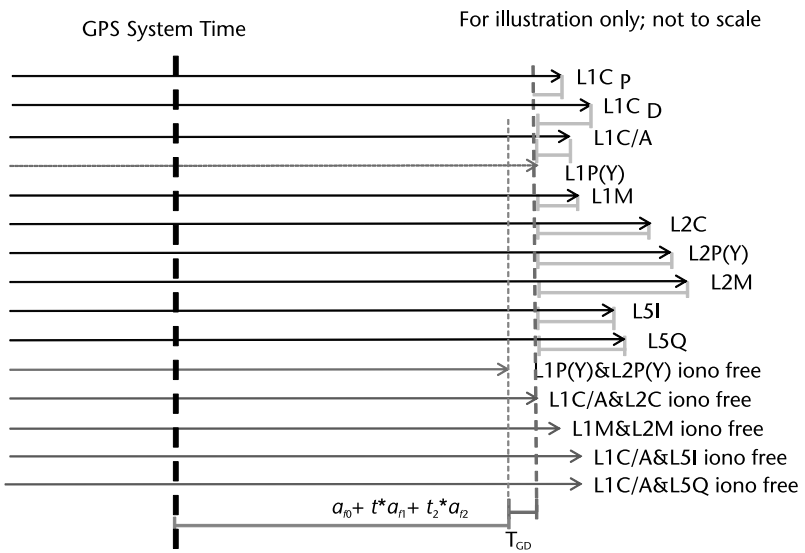
Figure 10.18 GPS example. (a) Different SV ranging code transmission times (Courtesy of Gary Okerson.) (b) Ionospheric delay with ideal case, identical group delay on L1 P(Y) and L2 P(Y). (Courtesy of The MITRE Corporation/Gary Okerson.) (c) Ionospheric delay practical case, different group delays on L1 P(Y) and L2 P(Y). (Courtesy of The MITRE Corporation/Gary Okerson.) (d) Ranging code offset from GPS system time as a function of SV clock offset, TGD, and ISC. (Courtesy of The MITRE Corporation/Gary Okerson.)

10.2.7.2 User Equipment Biases

User equipment bias errors introduced by the receiver hardware are often ignored because they are relatively small in comparison to other error sources, especially when cancellation is considered. GNSS signals are delayed as they travel through



(c)



(d)

Figure 10.18 (continued)

the antenna, analog hardware (e.g., RF and IF filters, low-noise amplifiers, mixers) and digital processing until the point where pseudorange and carrier-phase measurements are physically made within the digital receiver channels (see Chapter 8). Although the absolute delay values for propagation from the antenna phase center until the digital channels may be quite large [over 1 μs with long antenna-receiver cable runs or when surface acoustic wave (SAW) filters are employed], for similar signals on the same carrier frequency the delays experienced for the set of visible signals are nearly exactly equal. The absolute delay is important for timing applications, and must be calibrated out. However, for many applications, the common delay does not affect performance since it does not influence positioning accuracy, but rather directly appears only in the least-squares estimate of receiver

clock bias. BPSK-R(1) signals have measurably different power spectra due to their short-ranging codes. Since GNSS receiver front ends, in general, do not have constant group delay throughout the passband, very small intersatellite biases can be observed upon BPSK-R(1) pseudoranges as long as the signals are on the same frequency. In this case, these intersatellite biases are typically on the order of a few millimeters for carrier phase measurements and on the centimeter level for pseudorange measurements.

Hardware biases between spectrally different signals upon one frequency, or among signals on different carrier frequencies are larger in magnitude. In [39], differential group delay biases between L1 GPS C/A and Galileo OS signals were analyzed within a representative receiver to be on the order of several nanoseconds (~1m in range). These biases are not common to all measurements and thus would influence positioning performance if not calibrated or estimated.

Within multifrequency receivers, a portion of electrical paths followed by the signals on different frequencies may be physically different resulting in sizeable differential range errors. For the positioning user, bias between multifrequency signals may often be ignored since it results in a common error for every ionospheric-free pseudorange code (see Section 10.2.4.1), which will drop out in the estimated receiver clock bias.

Another error that can be attributed to the receiver hardware is the hardware-induced multipath [40]. This error is caused by reflections of the GNSS signal that occur within the receiver hardware due to the presence of an impedance mismatch between RF components. This error can be removed or reduced by careful design of the receiver front end.

Lastly, the receiver antenna can produce both pseudorange and carrier phase biases that are a function of the elevation angle and azimuth to each tracked satellite. The pseudorange biases can be greater than a meter in magnitude for inexpensive antennas, but are controlled by design to no more than tens of centimeters for high precision antennas. The carrier phase biases can be as large as a few centimeters in magnitude. Many high-precision users use calibration data to remove antenna-induced biases. Such data is widely available on the Internet in ANTEX format. Alternatively, for differential systems where the reference station(s) and end user are using the same model antenna oriented in the same direction (e.g., using the North mark on such equipment), the biases will tend to be common-mode and cancel out.

10.3 Pseudorange Error Budgets

Based on the above discussion regarding error constituents, we can develop pseudorange error budgets to aid our understanding of standalone GNSS accuracy. These budgets are intended to serve as guidelines for position error analyses. As indicated in (10.1), position error is a function of both the pseudorange error (UERE) and user/satellite geometry (DOP). The geometry factor will be discussed in Section 11.2.1.

The total system UERE is composed of components from each system segment: the space segment, the CS, and the user segment. This budget can be made based on either the use of single-frequency measurements or the use of dual-frequency measurements to determine the ionospheric delay. The error components are root-sum-squared (RSS) to form the total system UERE, which is assumed to be Gaussian distributed. The use of RSS addition of UERE components is justified under the assumption that the errors can be treated as independent random variables such that the variances add, or equivalently the one sigma total error is the RSS of the individual one sigma values.

Tables 10.3 and 10.4 show estimates of representative contemporary UERE budgets based on the data presented in Sections 10.2.1 to 10.2.7. *It is important to note that the actual values of the parameters in Tables 10.3 and 10.4 will vary as a function of the SATNAV system in use as well as the type of user equipment.* Table 10.3 describes a typical UERE budget for a dual-frequency receiver, whereas Table 10.4 shows a representative UERE budget for a single-frequency receiver. For a single-frequency user, the dominant pseudorange error source is the residual ionospheric delay after applying the broadcast ionospheric delay corrections. Dual-frequency users can use the technique described in Section 10.2.4.1 to nearly completely remove the error due to ionospheric delays.

Table 10.3 Typical GNSS UERE Budget for Dual-Frequency Receiver

Segment Source	*Error Source*	*1σ Error (m)*
Space/control	Broadcast clock	0.4
	Broadcast ephemeris	0.3
User	Residual ionospheric delay	0.1
	Residual tropospheric delay	0.2
	Receiver noise and resolution	0.1
	Multipath	0.2
System UERE	Total (RSS)	0.6

Table 10.4 Typical GNSS UERE Budget for Single-Frequency Receiver

Segment Source	*Error Source*	*1σ Error (m)*
Space/control	Broadcast clock	0.4
	Differential group delay	0.15
	Broadcast ephemeris	0.3
User	Residual ionospheric delay	7.0*
	Tropospheric delay	0.2
	Receiver noise and resolution	0.1
	Multipath	0.2
System UERE	Total (RSS)	7.03*

*Note that residual ionospheric errors tend to be highly correlated amongst satellites resulting in position errors being far less than predicted using DOP · UERE (see discussion in Section 11.2.2.)

References

[1] Ward, P., "An Inside View of Pseudorange and Delta Pseudorange Measurements in a Digital NAVSTAR GPS Receiver," *International Telemetering Conference, GPS-Military and Civil Applications*, San Diego, CA, October 14, 1981, pp. 63–69.

[2] GPS Directorate, IS-GPS-200H, NAVSTAR *GPS Space Segment/Navigation User Interfaces*, U.S. Air Force GPS Directorate, El Segundo, CA, September 24, 2013.

[3] European Union, *European GNSS (Galileo) Open Service Signal in Space Interface Control Document*, Issue 1.2, November 2015.

[4] China Satellite Navigation Office, *BeiDou Satellite Navigation System Signal in Space Interface Control Document Open Service Signal (Version 2.0)*, December 2013.

[5] Japan Aerospace Exploration Agency, *Interface Specification for QZSS (IS-QZSS)* Ver. 1.7, July 14, 2016

[6] ISRO Satellite Centre Indian Space Research Organization, ISRO-IRNSS-ICD-SPS-1.0, *Indian Regional Navigation Satellite System, Signal In Space ICD for Standard Positioning Service Version 1.0*, Bangalore, June 2014.

[7] Russian Institute of Space Device Engineering, *GLONASS Interface Control Document,* Navigational Radiosignal in Bands L1, L2 (Edition 5.1), Moscow, 2008.

[8] Walter, T., and J. Blanch, "Characterization of GNSS Clock and Ephemeris Errors to Support ARAIM," *Proc. of the ION 2015 Pacific PNT Meeting*, Marriott Waikiki Beach Resort & Spa, Honolulu, HI, April 20–23, 2015, pp. 920–931.

[9] Heng, L., et al., "Statistical Characterization of GLONASS Broadcast Clock Errors and Signal-In-Space Errors," *Proc. of the 2012 International Technical Meeting of The Institute of Navigation*, Marriott Newport Beach Hotel & Spa, Newport Beach, CA, January 30–February 1, 2012, pp. 1697–1707.

[10] Montenbruck, O., and P. Steigenberger, "IGS-MGEX: Preparing for a Multi-GNSS World," *U.S. PNT Advisory Board*, Annapolis, MD, June 2015. http://www.gps.gov/governance/advisory/meetings/2015-06/montenbruck.pdf. Accessed on January 10, 2017.

[11] Olynik, M., et al., "Temporal Variability of GPS Error Sources and Their Effect on Relative Positioning Accuracy," *Proc. of The Institute of Navigation National Technical Meeting*, January 2002.

[12] Hatch, R., "Relativity and GPS-I," *Galilean Electrodynamics*, Vol. 6, No. 3, May/June 1995, pp. 52–57.

[13] Ashby, N., and J. J. Spilker, Jr., "Introduction to Relativity Effects on the Global Positioning System," in *Global Positioning System: Theory and Applications, Vol. II*, B. Parkinson and J. J. Spilker, Jr., (eds.), Washington, D.C.: American Institute of Aeronautics and Astronautics, 1996.

[14] Seeber, G., *Satellite Geodesy*, Berlin, Germany: Walter de Gruyter, 1993.

[15] Nelson, R. A., *Relativistic Time Transfer in the Solar System*, Bethesda, MD: Satellite Engineering Research Corporation, May 29, 2007.

[16] Ashby, N., and M. Weiss, *Global Positioning System Receivers and Relativity*, National Institute of Standards and Technology (NIST) Technical Note 1385, Boulder, CO, March 1999.

[17] Hofmann-Wellenhof, B., H. Lichtenegger, and J. Collins, *GPS Theory and Practice*, New York: Springer-Verlag, 1993.

[18] Special Committee 159, *Minimum Operational Performance Standards for Global Positioning System/Wide Area Augmentation System Airborne Equipment*, RTCA, Inc., Document DO-229E, Washington, D.C., December 14, 2016.

[19] Prieto-Cerdeira, R., et al., "Performance of the Galileo Single-Frequency Ionospheric Correction During In-Orbit Validation," *GPS World Magazine*, June 2014.

[20] Klobuchar, J, "A First Order, Worldwide, Ionospheric, Time-Delay Algorithm," AFCRL-TR-75-0502, *Air Force Surveys in Geophysics*, No. 324, September 25, 1975.

[21] Jorgensen, P. S., "An Assessment of Ionospheric Effects on the User," *NAVIGATION: Journal of the Institute of Navigation*, Vol. 36, No. 2, Summer 1989.

[22] U.S. Department of Defense, *Global Positioning System Standard Positioning Service Performance Standard*, September 2008.

[23] European Commission, *European GNSS (Galileo) Open Service Ionospheric Correction Algorithm for Galileo Single Frequency Users*, Version 1.2, September 2016.

[24] Komjathy, A., et al., "The Ionospheric Impact of the October 2003 Storm Event on WAAS," *Proc. of The Institute of Navigation ION GNSS 2004*, Long Beach, CA, September 2004.

[25] Wanninger, L., "Effects of the Equatorial Ionosphere on GPS," *GPS World*, July 1993.

[26] Klobuchar, J., P. Doherty, and M. B. El-Arini, "Potential Ionospheric Limitations to GPS Wide-Area Augmentation System," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 42, No. 2, Summer 1995.

[27] Doherty, P., et al., "Statistics of Time Rate of Change of Ionospheric Range Delay," *Proceedings of The Institute of Navigation ION GPS-94*, Salt Lake City, UT, September 1994.

[28] Hopfield, H., "Two-Quartic Tropospheric Refractivity Profile for Correcting Satellite Data," *Journal of Geophysical Research*, Vol. 74, No. 18, 1969.

[29] Smith, E., Jr., and S. Weintraub, "The Constants in the Equation for Atmospheric Refractive Index at Radio Frequencies," *Proc. of the Institute of Radio Engineers*, No. 41, 1953.

[30] Remondi, B., "Using the Global Positioning System (GPS) Phase Observable for Relative Geodesy: Modeling, Processing, and Results," Ph.D. Dissertation, Center for Space research, University of Austin, Austin, TX, 1984.

[31] Goad, C., and L. Goodman, "A Modified Hopfield Tropospheric Refraction Correction Model," *Proc. of the Fall Annual Meeting of the American Geophysical Union*, San Francisco, CA, 1974.

[32] Saastomoinen, J., "Atmospheric Correction for the Troposphere and Stratosphere in Radio Ranging of Satellites," *Use of Artificial Satellites for Geodesy*, Geophysical Monograph 15, Washington, D.C.: American Geophysical Union, 1972.

[33] Collins, P., R. Langley, and J. LaMance, "Limiting Factors in Tropospheric Propagation Delay Error Modelling for GPS Airborne Navigation," *Proc. of The Institute of Navigation Annual Meeting*, Cambridge, MA, June 1996.

[34] Guo, J., and Langley R. B., "A New Tropospheric Propagation Delay Mapping Function for Elevation Angles Down to 2°," *Proc. of The Institute of Navigation ION GPS/GNSS 2003*, Portland, OR, September 9–12, 2003.

[35] Niell, A. E., "Global Mapping Functions for the Atmosphere Delay at Radio Wavelengths," *Journal of Geophysical Research*, Vol. 101, No. B2, 1996, pp. 3227–3246.

[36] Marini, J. W., "Correction of Satellite Tracking Data for an Arbitrary Tropospheric Profile," *Radio Science*, Vol. 7, No. 2, 1972, pp. 223–231.

[37] Altshuler, E. E., *Corrections for Tropospheric Range Error*, Report AFCRL-71-0419, Air Force Cambridge Research Laboratory, Hanscom Field, Bedford, MA, July 27, 1971.

[38] Coster, A. J., et al., "Characterization of Atmospheric Propagation Errors for DGPS," *Proc. of The Institute of Navigation's Annual Meeting*, Denver, CO, June 1998.

[39] Hegarty, C., E. Powers, and B. Fonville, "Accounting for Timing Biases Between GPS, Modernized GPS, And Galileo Signals," *Proc. of The 36th Annual Precise Time and Time Interval (PTTI) Meeting*, Washington, D.C., December 2004.

[40] Keith, J. P., "Multipath Errors Induced by Electronic Components in Receiver Hardware," M.S.E.E. thesis, Ohio University, Athens, OH, November 2002.

CHAPTER 11

Performance of Stand-Alone GNSS

Chris Hegarty, Joe Leva, Karen Van Dyke, and Todd Walter

11.1 Introduction

The accuracy with which a user receiver can determine its position or velocity, or synchronize to GNSS system timescales, depends on a complicated interaction of various factors. In general, GNSS accuracy performance depends on the quality of the pseudorange and carrier-phase measurements as well as the broadcast navigation data. In addition, the fidelity of the underlying physical model that relates these parameters is relevant. For example, the accuracy to which the satellite clock offsets relative to a chosen common timescale are known to the user, or the accuracy to which satellite-to-user propagation errors are compensated, are important. Relevant errors are induced by the control, space, and user segments.

To analyze the effect of errors on accuracy, a fundamental assumption is usually made that the error sources can be allocated to individual satellite pseudoranges and can be viewed as effectively resulting in an equivalent error in the pseudorange values. The effective accuracy of the pseudorange value is termed the user-equivalent range error (UERE). The UERE for a given satellite is considered to be the (statistical) sum of the contributions from each of the error sources associated with the satellite. Usually, the error components are considered independent and the composite UERE for a satellite is approximated as a zero mean Gaussian random variable where its variance is determined as the sum of the variance of each of its components. UERE is often assumed to be independent and identically distributed from satellite to satellite. However, for certain cases of interest, it is sometimes appropriate for these assumptions to be modified. For example, if one is considering the processing of GNSS satellites from two core constellations, the UERE associated with one system might be modeled with a different variance than the other. In other situations, it might be appropriate to model certain components of UERE with variances that monotonically increase with decreasing elevation angle and a smaller subset as being correlated among the satellites.

The accuracy of the position/time solution determined by GNSS is ultimately expressed as the product of a geometry factor and a pseudorange error factor.

Loosely speaking, error in the GNSS position solution that is based upon only pseudorange measurements is estimated by the formula

$$(\text{error in GNSS solution}) = (\text{geometry factor}) \times (\text{pseudorange error factor}) \quad (11.1)$$

Under appropriate assumptions, the pseudorange error factor is the satellite UERE. The geometry factor expresses the composite effect of the relative satellite/user geometry on the GNSS solution error. It is generically called the dilution of precision (DOP) associated with the satellite/user geometry.

Section 11.2 presents algorithms for estimating PVT for one or more GNSS constellations and provides a derivation of (11.1). 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. Sections 11.3 through 11.5 discuss, respectively, the three other important performance metrics of availability, integrity, and continuity.

11.2 Position, Velocity, and Time Estimation Concepts

Chapter 2 described some basic techniques for estimating the position, velocity, and time (PVT) of a possibly mobile GNSS receiver. This section discusses a variety of additional concepts regarding PVT estimation, beginning with an expanded description of the role of geometry in GNSS PVT accuracy and a number of accuracy metrics that are commonly used. This section also describes a number of advanced PVT estimation techniques including the use of weighted-least-squares (WLS), additional estimated parameters (beyond the user x, y, z position coordinates and clock offset), and Kalman filtering.

11.2.1 Satellite Geometry and Dilution of Precision in GNSS

As motivation for the concept of DOP as it applies to GNSS, consider once again the foghorn example introduced in Section 2.1.1. In this example, a user locates his or her position from ranging measurements from two foghorns. The assumptions are that the user has a synchronized time base relative to the foghorns and has knowledge of the location of the foghorns and their transmission times. The user measures the time of arrival of each of the foghorn signals and computes a propagation time, which determines the user's range from each foghorn. The user locates his or her position from the intersection of the range rings determined from the time-of-arrival measurements.

In the presence of measurement errors, the range rings used to compute the user's location will be in error and result in error in the computed position. The concept of DOP is the idea that the position error that results from measurement errors depends on the user/foghorn relative geometry. Graphically, these ideas are illustrated in Figure 11.1. Two geometries are indicated. In Figure 11.1(a), the foghorns are located approximately at right angles with respect to the user location. In Figure 11.1(b), the angle between the foghorns as viewed from the user is much smaller. In both cases, portions of the error-free range rings are indicated and intersect at the user's location. Additional ring segments are included that illustrate the

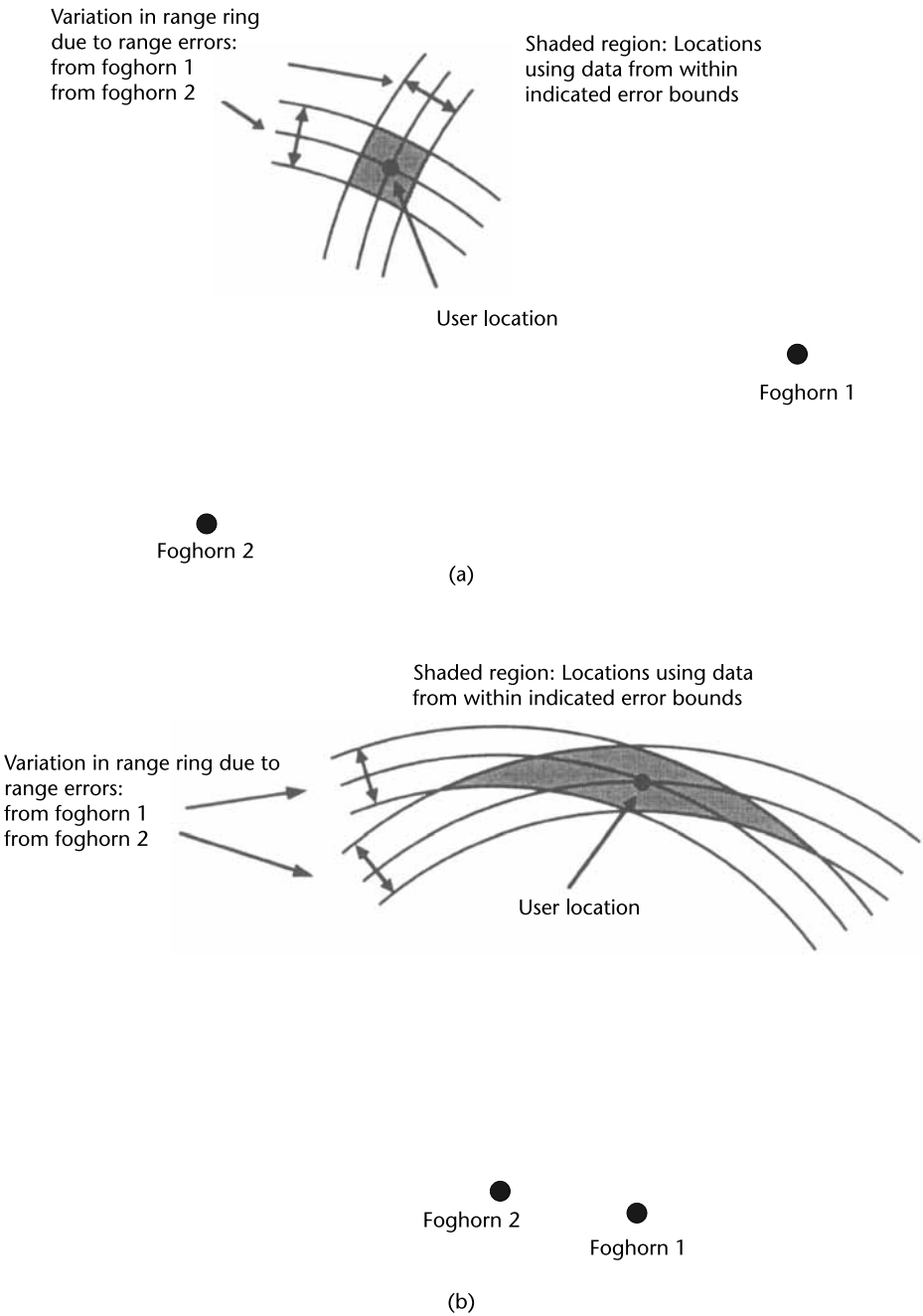


Figure 11.1 Relative geometry and dilution of precision: (a) geometry with low DOP, and (b) geometry with high DOP.

variation in range ring position resulting from ranging errors to the foghorns. The error range illustrated in both figures is the same. The shaded regions indicate the set of locations that can be obtained if one uses ranging measurements within the illustrated error bounds. The accuracy of the computed location is very different for the two cases. With the same measurement error variation, geometry (b) gives considerably more error in the computed user's location than in (a), as is evident

from comparison of the shaded regions. Geometry (b) is said to have a larger dilution of precision then geometry (a). For comparable measurement errors, geometry (b) results in larger errors in the computed location.

A formal derivation of the DOP relations in GNSS begins with the linearization of the pseudorange equations given in Section 2.5.2. The linearization is the Jacobian relating changes in the user position and time bias to changes in the pseudorange values. This relationship is inverted in accordance with the solution algorithm and is used to relate the covariance of the user position and time bias to the covariance of the pseudorange errors. The DOP parameters are defined as geometry factors that relate parameters of the user position and time bias errors to those of the pseudorange errors.

The offset $\Delta\mathbf{x}$ in the user's position and time bias relative to the linearization point is related to the offset in the error-free pseudorange values $\Delta\boldsymbol{\rho}$ by the relation

$$\mathbf{H}\Delta\mathbf{x} = \Delta\boldsymbol{\rho} \tag{11.2}$$

The vector $\Delta\mathbf{x}$ has four components. The first three are the position offset of the user from the linearization point; the fourth is the offset of the user time bias from the bias assumed in the linearization point. $\Delta\boldsymbol{\rho}$ is the vector offset of the error-free pseudorange values corresponding to the user's actual position and the pseudorange values that correspond to the linearization point $\mathbf{H}$ is the $n \times 4$ matrix

$$\mathbf{H} = \begin{bmatrix} a_{x1} & a_{y1} & a_{z1} & 1 \\ a_{x2} & a_{y2} & a_{z2} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ a_{xn} & a_{yn} & a_{zn} & 1 \end{bmatrix} \tag{11.3}$$

and the $\mathbf{a}_i = (a_{xi}, a_{yi}, a_{zi})$ are the unit vectors pointing from the linearization point to the location of the ith satellite. If $n = 4$ and data from just four satellites are being used, and if the linearization point is close to the user's location, the user's location and time offset are obtained by solving (11.2) for $\Delta\mathbf{x}$ (i.e., if the linearization point is close enough to the user position, iteration is not required). One obtains

$$\Delta\mathbf{x} = \mathbf{H}^{-1}\Delta\boldsymbol{\rho} \tag{11.4}$$

and the offset of the user's position from the linearization point is expressed as a linear function of $\Delta\boldsymbol{\rho}$. In the case of $n > 4$, the method of least squares can be used to solve (11.2) for $\Delta\mathbf{x}$ (see Appendix A). The least-squares result can be obtained formally by multiplying both sides of (11.2) on the left by the matrix transpose of $\mathbf{H}$ obtaining $\mathbf{H}^T\mathbf{H}\Delta\mathbf{x} = \mathbf{H}^T\Delta\boldsymbol{\rho}$. The matrix combination $\mathbf{H}^T\mathbf{H}$ is a square 4×4 matrix, and one can solve for $\Delta\mathbf{x}$ by multiplying both sides by the inverse, $(\mathbf{H}^T\mathbf{H})^{-1}$. (The matrix will be invertible provided the tips of the unit vectors $\mathbf{a}_i$ do not all lie in a plane.) One obtains

$$\Delta\mathbf{x} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\Delta\boldsymbol{\rho} \tag{11.5}$$

which is the least-squares formulation for $\Delta\mathbf{x}$ as a function of $\Delta\boldsymbol{\rho}$. We observe that if $n = 4$, $(\mathbf{H}^T\mathbf{H})^{-1} = \mathbf{H}^{-1}(\mathbf{H}^T)^{-1}$ and (11.5) reduces to (11.4).

The pseudorange measurements are not error-free and can be viewed as a linear combination of three terms,

$$\Delta\boldsymbol{\rho} = \boldsymbol{\rho}_T - \boldsymbol{\rho}_L + d\boldsymbol{\rho} \tag{11.6}$$

where $\boldsymbol{\rho}_T$ is the vector of error-free (true) pseudorange values, $\boldsymbol{\rho}_L$ is the vector of pseudorange values computed at the linearization point, and $d\boldsymbol{\rho}$ represents the net error in the pseudorange values. Similarly, $\Delta\mathbf{x}$ can be expressed as

$$\Delta\mathbf{x} = \mathbf{x}_T - \mathbf{x}_L + d\mathbf{x} \tag{11.7}$$

where $\mathbf{x}_T$ is the error-free (true) position and time, $\mathbf{x}_L$ is the position and time defined as the linearization point, and $d\mathbf{x}$ is the error in the position and time estimate. Substituting (11.6) and (11.7) into (11.5) and using the relation $\mathbf{x}_T - \mathbf{x}_L = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T(\boldsymbol{\rho}_T - \boldsymbol{\rho}_L)$ [this follows from the relation $\mathbf{H}(\mathbf{x}_T - \mathbf{x}_L) = (\boldsymbol{\rho}_T - \boldsymbol{\rho}_L)$, which is a restatement of (11.2)], one obtains

$$d\mathbf{x} = \left[\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\right]d\boldsymbol{\rho} = \mathbf{K}d\boldsymbol{\rho} \tag{11.8}$$

The matrix $\mathbf{K}$ is defined by the expression in brackets. Equation (11.8) gives the functional relationship between the errors in the pseudorange values and the induced errors in the computed position and time bias. It is valid provided that the linearization point is sufficiently close to the user's location and that the pseudorange errors are sufficiently small so that the error in performing the linearization can be ignored.

Equation (11.8) is the fundamental relationship between pseudorange errors and computed position and time bias errors. The matrix$(\mathbf{H}^T\mathbf{H})^{-1}\ \mathbf{H}^T$, which is sometimes called the *least-squares solution matrix* or *pseudoinverse* of $\mathbf{H}$, is a $4 \times n$ matrix and depends only on the relative geometry of the user and the satellites participating in the least square solution computation. In many applications, the user/satellite geometry can be considered fixed and (11.8) yields a linear relationship between the pseudorange errors and the induced position and time bias errors.

The pseudorange errors are considered to be random variables and (11.8) expresses $d\mathbf{x}$ as a random variable functionally related to $d\boldsymbol{\rho}$. The error vector $d\boldsymbol{\rho}$ is usually assumed to have components that are jointly Gaussian and to be zero mean. With the geometry considered fixed, it follows that $d\mathbf{x}$ is also Gaussian and zero mean. The covariance of $d\mathbf{x}$ is obtained by forming the product $d\mathbf{x}d\mathbf{x}^T$and computing an expected value. By definition, one obtains

$$\operatorname{cov}(d\mathbf{x}) = E\left[d\mathbf{x}d\mathbf{x}^T\right] \tag{11.9}$$

where $\operatorname{cov}(d\mathbf{x}) = E[d\mathbf{x}d\mathbf{x}^T]$ denotes the covariance of $d\mathbf{x}$ and E represents the expectation operator. Substituting from (11.8) and viewing the geometry as fixed, one obtains

$$\begin{aligned}\operatorname{cov}(d\mathbf{x}) &= E\left[\mathbf{K}d\boldsymbol{\rho}d\boldsymbol{\rho}^T\mathbf{K}^T\right] = E\left[\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T d\boldsymbol{\rho}d\boldsymbol{\rho}^T\mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\right] \\ &= \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\operatorname{cov}(d\boldsymbol{\rho})\mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\end{aligned} \tag{11.10}$$

Note that in this computation, $(\mathbf{H}^T\mathbf{H})^{-1}$ is symmetric. [This follows from an application of the general matrix relations $(\mathbf{AB})^T\,\mathbf{B}^T\mathbf{A}^T$ and $(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$, which are valid whenever the indicated operations are defined.] A commonly used simplifying assumption is that the components of $d\boldsymbol{\rho}$ are identically distributed and independent and have a variance equal to the square of the satellite UERE. With these assumptions, the covariance of $d\boldsymbol{\rho}$ is a scalar multiple of the identity

$$\operatorname{cov}(d\boldsymbol{\rho}) = \mathbf{I}_{n\times n}\sigma_{UERE}^2 \tag{11.11}$$

where $\mathbf{I}_{n\times n}$ is the $n \times n$ identity matrix. Substitution into (11.10) yields

$$\operatorname{cov}(d\mathbf{x}) = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\sigma_{UERE}^2 \tag{11.12}$$

Under the stated assumptions, the covariance of the errors in the computed position and time bias is just a scalar multiple of the matrix $(\mathbf{H}^T\mathbf{H})^{-1}$. The vector $d\mathbf{x}$ has four components, which represent the error in the computed value for the vector $\mathbf{x}_T = (x_u, y_u, z_u, ct_b)$. The covariance of $d\mathbf{x}$ is a 4×4 matrix and has an expanded representation

$$\operatorname{cov}(d\mathbf{x}) = \begin{bmatrix} \sigma_{x_u}^2 & \sigma_{x_u y_u}^2 & \sigma_{x_u z_u}^2 & \sigma_{x_u ct_b}^2 \\ \sigma_{x_u y_u}^2 & \sigma_{y_u}^2 & \sigma_{y_u z_u}^2 & \sigma_{y_u ct_b}^2 \\ \sigma_{x_u z_u}^2 & \sigma_{y_u z_u}^2 & \sigma_{z_u}^2 & \sigma_{z_u ct_b}^2 \\ \sigma_{x_u ct_b}^2 & \sigma_{y_u ct_b}^2 & \sigma_{z_u ct_b}^2 & \sigma_{ct_b}^2 \end{bmatrix} \tag{11.13}$$

The components of the matrix $(\mathbf{H}^T\mathbf{H})^{-1}$ quantify how pseudorange errors translate into components of the covariance of $d\mathbf{x}$.

Dilution of precision parameters in GNSS are defined in terms of the ratio of combinations of the components of $\operatorname{cov}(d\mathbf{x})$ and σ_{UERE}. (It is implicitly assumed in the DOP definitions that the user/satellite geometry is considered fixed. It is also assumed that local east, north, up (ENU) user coordinates are being used in the specification of $\operatorname{cov}(d\mathbf{x})$ and $d\mathbf{x}$. The positive x-axis points east, the y-axis points north, and the z-axis points up; see Section 2.2.3.) The most general parameter is termed the geometric dilution of precision (GDOP) and is defined by the formula

$$\text{GDOP} = \frac{\sqrt{\sigma_{x_u}^2 + \sigma_{y_u}^2 + \sigma_{z_u}^2 + \sigma_{ct_b}^2}}{\sigma_{UERE}} \tag{11.14}$$

A relationship for GDOP is obtained in terms of the components of $(\mathbf{H}^T\mathbf{H})^{-1}$ by expressing $(\mathbf{H}^T\mathbf{H})^{-1}$ in component form

$$\left(\mathbf{H}^T\mathbf{H}\right)^{-1} = \begin{bmatrix} D_{11} & D_{12} & D_{13} & D_{14} \\ D_{21} & D_{22} & D_{23} & D_{24} \\ D_{31} & D_{32} & D_{33} & D_{34} \\ D_{41} & D_{42} & D_{43} & D_{44} \end{bmatrix} \tag{11.15}$$

and then substituting (11.15) and (11.13) into (11.12). A trace operation on (11.13) followed by a square root shows that GDOP can be computed as the square root of the trace of the $(\mathbf{H}^T\mathbf{H})^{-1}$ matrix:

$$\text{GDOP} = \sqrt{D_{11} + D_{22} + D_{33} + D_{44}} \tag{11.16}$$

Equation (11.14) can be rearranged to obtain

$$\sqrt{\sigma_{x_u}^2 + \sigma_{y_u}^2 + \sigma_{z_u}^2 + \sigma_{ct_b}^2} = \text{GDOP} \times \sigma_{UERE} \tag{11.17}$$

which has the form given in (11.1). The square-root term on the left side gives an overall characterization of the error in the GNSS solution. GDOP is the geometry factor. It represents the amplification of the standard deviation of the measurement errors onto the solution. From (11.16), GDOP is seen to be a function solely of the satellite/user geometry. The value σ_{UERE} is the pseudorange error factor.

Several other DOP parameters are in common use that are useful to characterize the accuracy of various components of the position/time solution. These are termed position dilution of precision (PDOP), horizontal dilution of precision (HDOP), vertical dilution of precision (VDOP), and time dilution of precision (TDOP). These DOP parameters are defined in terms of the satellite UERE and elements of the covariance matrix for the position/time solution as follows:

$$\sqrt{\sigma_{x_u}^2 + \sigma_{y_u}^2 + \sigma_{z_u}^2} = \text{PDOP} \times \sigma_{UERE} \tag{11.18}$$

$$\sqrt{\sigma_{x_u}^2 + \sigma_{y_u}^2} = \text{HDOP} \times \sigma_{UERE} \tag{11.19}$$

$$\sigma_{z_u} = \text{VDOP} \times \sigma_{UERE} \tag{11.20}$$

$$\sigma_{ct_b} = \text{TDOP} \times \sigma_{UERE} \tag{11.21}$$

The DOP values can be expressed in terms of the components of $(\mathbf{H}^T\mathbf{H})^{-1}$as follows:

$$\text{PDOP}=\sqrt{D_{11} + D_{22} + D_{33}} \tag{11.22}$$

$$\text{HDOP}=\sqrt{D_{11} + D_{22}} \tag{11.23}$$

$$\text{VDOP}=\sqrt{D_{33}} \tag{11.24}$$

$$\text{TDOP}=\sqrt{D_{44}} \tag{11.25}$$

Note in the TDOP expression that the variable ct_b represents a range equivalent of the time bias error and σ_{ct_b} is its standard deviation.

11.2.2 DOP Characteristics of GNSS Constellations

It is important to understand that GNSS satellite geometry and thus the DOP parameters vary with time. Figure 11.2 shows a typical variation of DOPs over a 1-day interval. The figure shows VDOP, HDOP, and TDOP predicted for a user in Bedford, Massachusetts (42.4906N, 71.260W) for the GPS expanded 27-satellite constellation (see Section 3.2.1). The results assume that the user can only track satellites with elevation angles that are at or above 5°. Over the 24-hour period, the average HDOP, VDOP, and TDOP values were 1.0, 1.4, and 0.9, respectively. The range of HDOP, VDOP, and TDOP were 0.7 to 1.7, 1.0 to 2.1, and 0.5 to 1.6, respectively. Correlation between the DOPs and the number of visible satellites is apparent (see Figure 11.3). DOPs often (but not always) become lower when more satellites are visible.

DOPs also vary with user location. For a user situated on the Earth, the statistics of the DOPs do not typically vary significantly with longitude, but do vary significantly with latitude. Figures 11.4 through 11.7 provide average DOPs (over 24 hours in time and 360° in longitude) as a function of latitude for the four core GNSS constellations (GPS, GLONASS, Galileo, and BeiDou). These results are based upon the reference constellation designs described in Chapters 3 to 6, and assume that satellites are only visible to the user above a 5° elevation angle. For

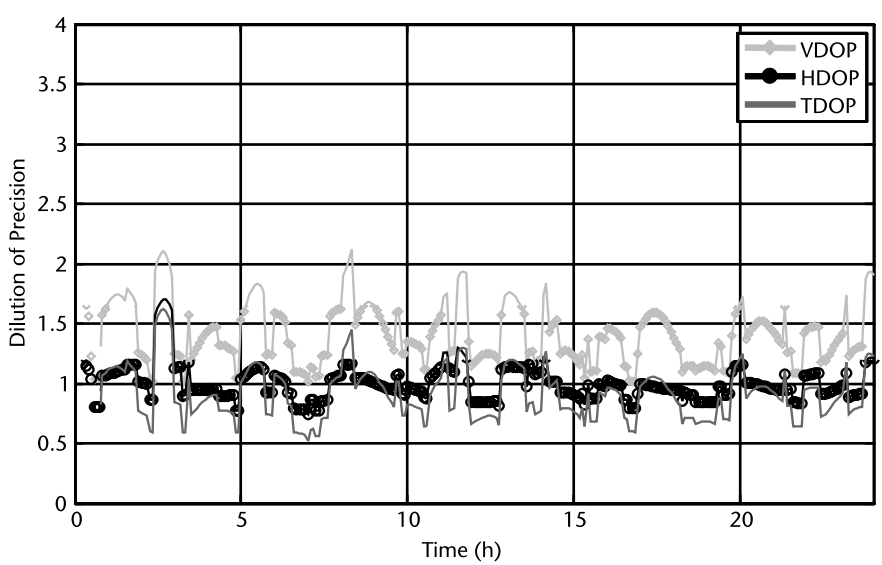


Figure 11.2 Predicted DOPs for a user in Bedford, Massachusetts (42.4906N, 71.260W) over a 1-day interval for the GPS Expanded 27-satellite constellation.

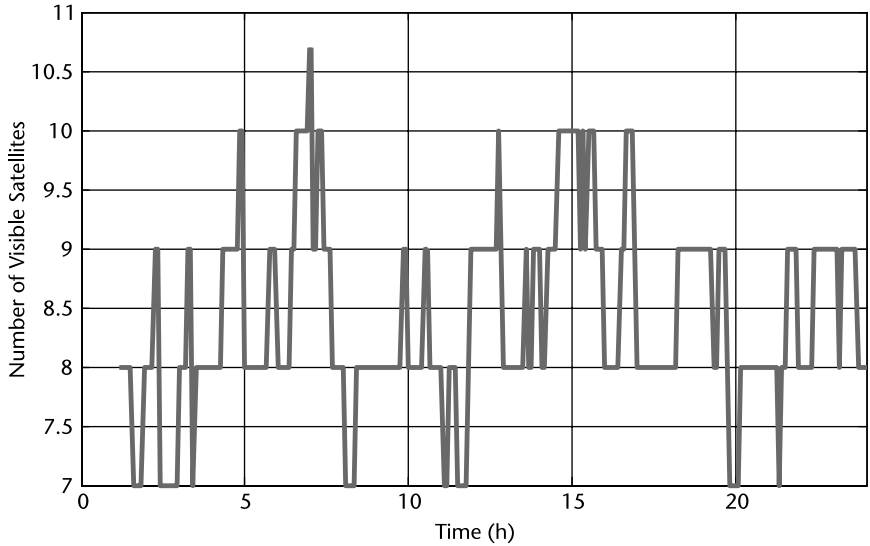


Figure 11.3 Number of visible satellites for a user in Bedford, Massachusetts (42.4906N, 71.260W) over a 1-day interval for the GPS Expanded 27-satellite constellation.

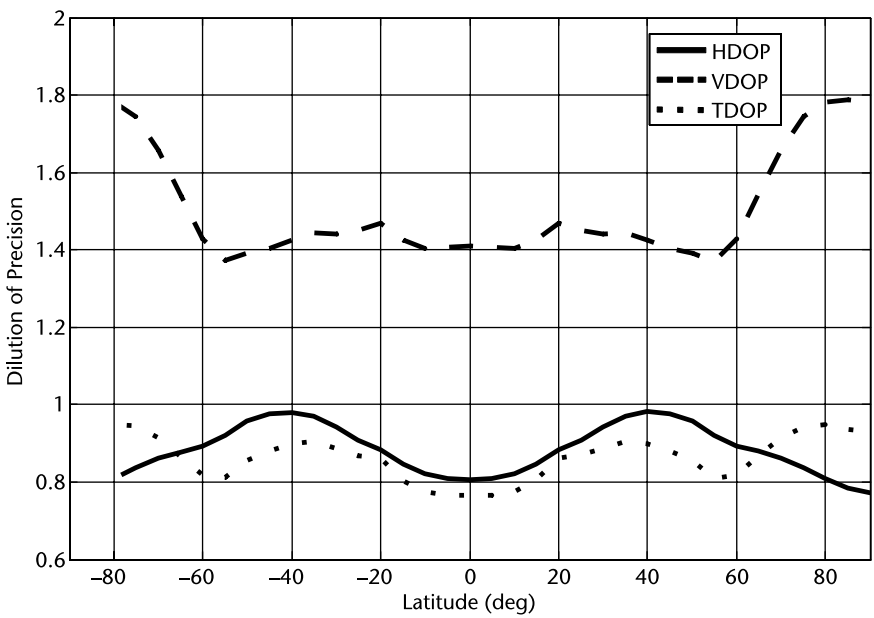


Figure 11.4 Average predicted DOPs as a function of user latitude for the GPS Expanded 27-satellite constellation.

BeiDou, only the 27-satellite MEO subconstellation was included for the results shown in Figure 11.7. It should also be noted that the Galileo constellation is anticipated to eventually include operational spare satellites that will result in better DOP values than shown in Figure 11.6, which only assumes 24 operational satellites.

For three of the four GNSS core constellations (GPS, Galileo, and BeiDou), VDOP values are highest at the North and South Poles, and conversely HDOP

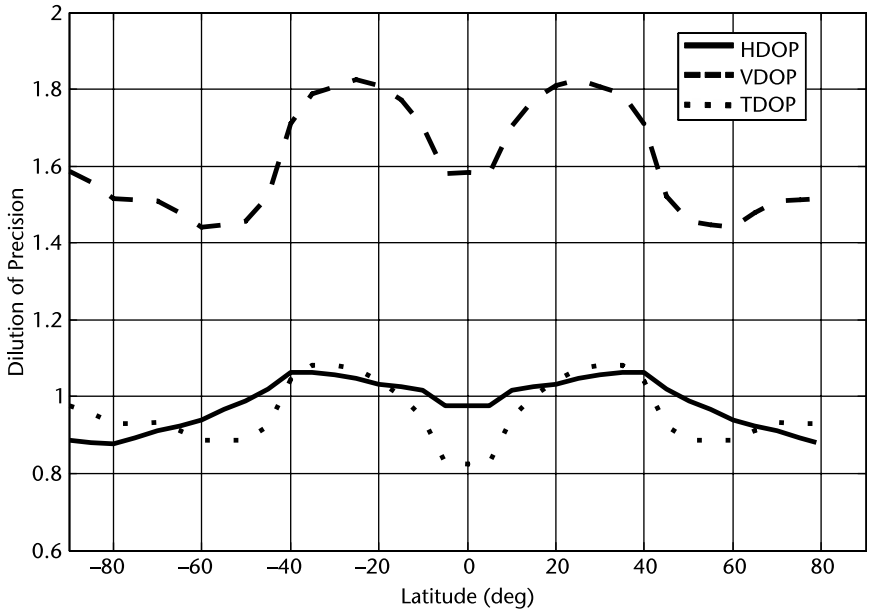


Figure 11.5 Average predicted DOPs as a function of user latitude for the GLONASS 24-satellite constellation.

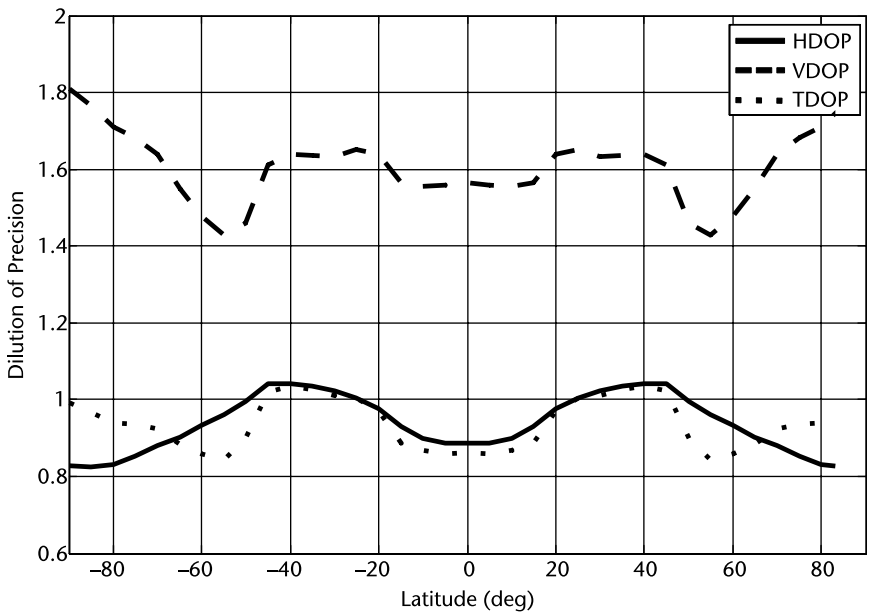


Figure 11.6 Average predicted DOPs as a function of user latitude for the Galileo 24-satellite constellation.

values are lowest at these same locations. This characteristic can be explained through the use of a *sky plot*. A sky plot shows the location of each visible satellite from the perspective of the user viewing the sky directly above his or her location. Figure 11.8 shows a sky plot for a user at the North Pole viewing only the GPS 27-satellite constellation over 24 hours. The user is at the center of the concentric

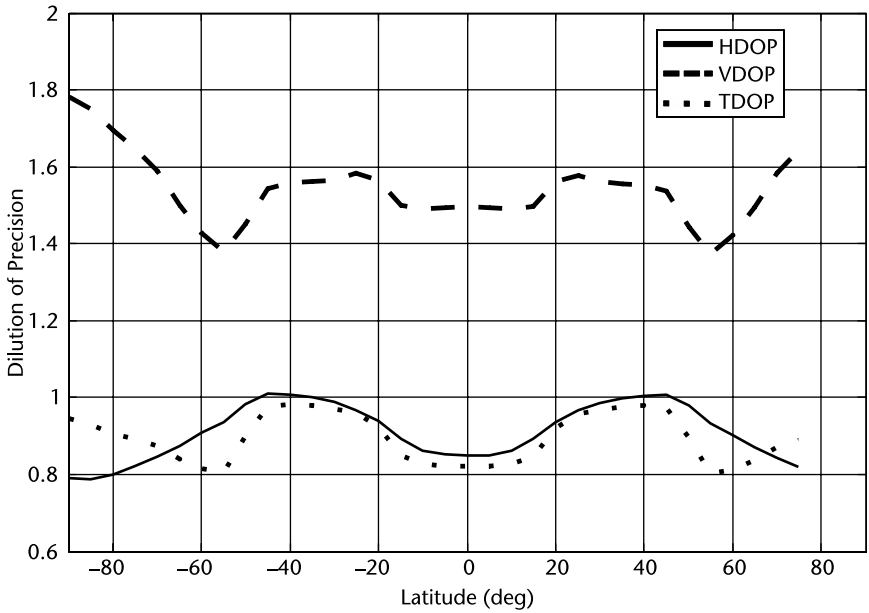


Figure 11.7 Average Predicted DOPs as a function of user latitude for the BeiDou 27-satellite MEO constellation.

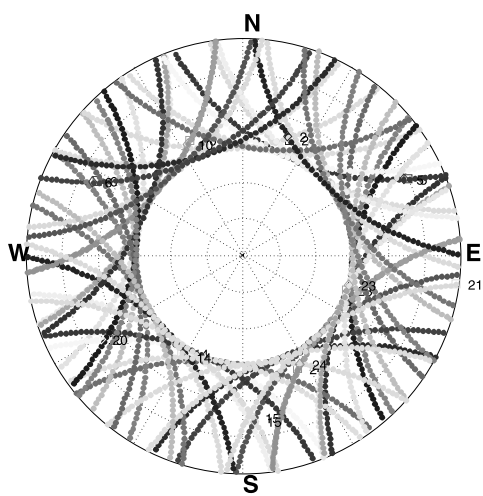


Figure 11.8 Sky plot of visible GPS satellites for user at North Pole.

circles, with the outermost circle representing 0° elevation. Each smaller circle delineates a 15° increase in elevation angle. The azimuth is 0° at North (N) and increases in the clockwise direction (azimuth directions are ill-defined at the Earth's poles, but this convention is the norm for sky plots). Note that, for the GPS 27-satellite constellation, a user at the North Pole will never see a satellite above 45.3° elevation angle. This visibility hole is due to the 55° inclination angle of the GPS orbital planes. VDOP suffers from the lack of overhead satellites, whereas HDOP is low because there are often many satellites well-distributed around the user in azimuth. GLONASS provides lower VDOPs at very high latitudes due to the high

inclination angle of its orbital planes (64.8° as compared with GPS, Galileo, and BeiDou MEO orbital planes, which are at inclinations of 55° or 56°).

A common feature in modern GNSS user equipment (see Chapter 8) is the ability to track satellites from multiple constellations. The DOPs applicable for a multiconstellation receiver are far better (lower) than the results for each single constellation at a time shown above. For instance, GPS and Galileo DOPs are assessed in [1]. The results for each constellation alone very closely match the results shown in Figures 11.4 and 11.6. When combined, the daily average VDOP for the worst-case location on Earth dropped from ~1.8 (each constellation alone) to below 1.15. The daily average HDOP dropped for the worst-case location from ~1 to below 0.65.

11.2.3 Accuracy Metrics

The formulae derived in Section 11.2.1 allow one to compute one-sigma horizontal, vertical, or three-dimensional position errors as a function of satellite geometry and the one-sigma range error. They also allow one to compute one-sigma user clock errors. It is important to recall that these formulae were derived under the assumptions that pseudorange errors are zero-mean with a Gaussian distribution and that pseudorange errors are independent from satellite to satellite. Oftentimes, other metrics besides one-sigma position errors are used to characterize GNSS accuracy performance. Some common metrics are derived and discussed in this section.

If pseudorange errors are Gaussian-distributed, (11.8) tells us that vertical position errors also have a Gaussian distribution:

$$dz = \sum_{m=1}^{N} K_{3,m} d\rho_m \tag{11.26}$$

where dz is the error in the vertical component of the computed position. This result is obtained by noting that a linear function of Gaussian random variables is itself a Gaussian random variable. One common measure of vertical positioning accuracy is the error magnitude that 95% of the measurements fall within, which is approximately equal to the two-sigma value for a Gaussian random variable. Thus:

$$95\% \text{ vertical position accuracy} \approx 2\sigma_{dz} = 2 \cdot \text{VDOP} \cdot \sigma_{UERE} \tag{11.27}$$

assuming that pseudorange errors are additionally zero-mean and independent among satellites.

As an example of the application of (11.27), consider a GPS PPS user at the North Pole. From Figure 11.4, the average value of VDOP at this location is 1.8. The UERE for the dual-frequency user is approximately 0.6 (see Table 10.3). Using (11.27), the predicted 95% vertical position accuracy is thus approximately $2 \times 1.8 \times 0.6 = 2.2$m at this location. The global average VDOP for the 27-satellite GPS constellation is 1.45, yielding a predicted 95% vertical position accuracy of 1.7m for an average location.

With regard to horizontal position errors, (11.8) can be specialized to the horizontal plane yielding:

$$d\mathbf{R} = \mathbf{K}_{2\times n} d\boldsymbol{\rho} \tag{11.28}$$

where $d\mathbf{R} = (dx, dy)^T$ is the vector component of the position error in the horizontal plane, $d\boldsymbol{\rho} = (d\boldsymbol{\rho}_1, \ldots, d\boldsymbol{\rho}_n)^T$ is the pseudorange measurement errors, and n is the number of satellites being used in the position calculation. $\mathbf{K}_{2\times n}$ is the upper $2 \times$ n submatrix of $\mathbf{K}$ and consists of its first two rows. For the standard least square solution technique, $\mathbf{K} = (\mathbf{H}^T\mathbf{H})^{-1}\,\mathbf{H}^T$.

For a fixed satellite geometry, (11.28) expresses the horizontal position errors as a linear function of the pseudorange measurement errors. If the pseudorange errors are zero mean and jointly Gaussian, $d\mathbf{R}$ also has these properties. If the pseudorange errors are also uncorrelated and identically distributed with variance σ_{UERE}^2, the covariance of the horizontal errors is given as

$$\operatorname{cov}(d\mathbf{R}) = \left((\mathbf{H}^T\mathbf{H})^{-1}\right)_{2\times 2} \sigma_{UERE}^{\ 2} \tag{11.29}$$

where the subscript notation denotes the upper left 2×2 submatrix of $(\mathbf{H}^T\mathbf{H})^{-1}$. The density function for $d\mathbf{R}$ is

$$f_{d\mathbf{R}}(x,y) = \frac{1}{2\pi\left[\det(\operatorname{cov}(d\mathbf{R}))\right]^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{u}^T\left[\operatorname{cov}(d\mathbf{R})\right]^{-1}\mathbf{u}\right) \tag{11.30}$$

where $\mathbf{u} = (x,y)^T$ and where det represents the determinant of a matrix.

The density function defines a two-dimensional bell-shaped surface. Contours of constant density are obtained by setting the exponent in parenthesis to a constant. One obtains equations of the form

$$\mathbf{u}^T\left[\operatorname{cov}(d\mathbf{R})\right]^{-1}\mathbf{u} = m^2 \tag{11.31}$$

where the parameter m ranges over positive values. The contour curves that result form a collection of concentric ellipses when plotted in the plane. The ellipse obtained when m equals 1 is termed the 1σ ellipse and has the equation

$$\mathbf{u}^T\left[\operatorname{cov}(d\mathbf{R})\right]^{-1}\mathbf{u} = 1 \tag{11.32}$$

(The 1σ ellipse is defined here as a specific cut through the probability density function and is not to be confused with 1σ containment. The latter curve is the locus of points, one point on each ray from the origin, where the points are at a distance of 1σ for the ray's direction. In general, the 1σ containment curve is a figure-eight-shaped curve that encloses the 1σ ellipse.) If the major and minor axis of the ellipse are aligned with the x and y axes, the equation for the ellipse reduces to $x^2 / \sigma_x^{\ 2} + y^2 / \sigma_y^{\ 2} = 1$. In general, however, the off-diagonal terms in are nonzero, and the elliptical contours for the density function are rotated relative to the x and y axes. Denote the major and minor axes of the 1σ ellipse by σ_L and σ_S, respectively (long and short). In general, the 1σ ellipse is contained in a box of width σ_x and height σ_y centered on the ellipse. Figure 11.9 illustrates graphically the relationship between the ellipse and the parameters σ_x, σ_y, σ_L, and σ_S.

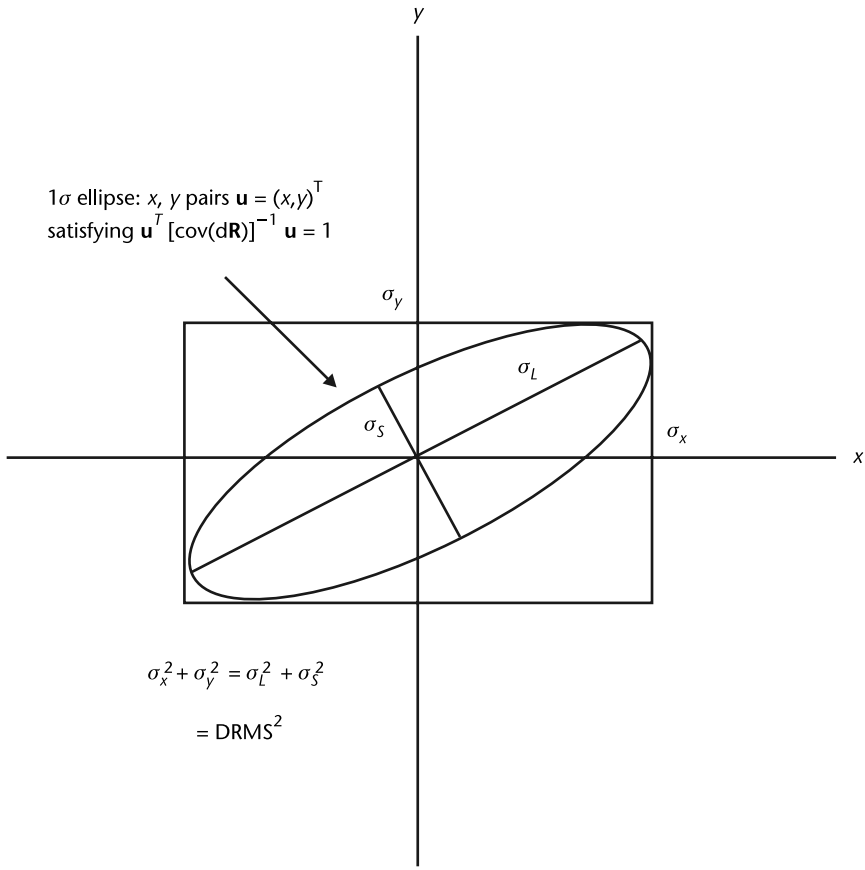


Figure 11.9 Relationship between 1σ ellipse and distribution parameters.

The probability that the error lies within the elliptical contour defined for a specific value of m is $1 - e^{-m^2/2}$. In particular, the probability of being in the 1σ ellipse ($m = 1$) is 0.39; the probability of being in the 2σ ellipse ($m = 2$) is 0.86. (These values are in contrast to the one-dimensional Gaussian result that the probability of being within $\pm 1\sigma$ of the mean is 0.68.)

Several parameters are in common use that characterize the magnitude of the horizontal error. The *distance root mean square* (drms) is defined by the formula

$$\text{drms} = \sqrt{\sigma_x^{\,2} + \sigma_y^{\,2}} \tag{11.33}$$

For a zero-mean random variable such as $d\mathbf{R}$, one has $\text{drms} = \sqrt{E\left(|d\mathbf{R}|^2\right)}$ and the drms corresponds to the square root of the mean value of the squared error (hence, its name). From (11.19), one immediately has

$$\text{drms} = \text{HDOP} \cdot \sigma_{UERE} \tag{11.34}$$

and the drms can be computed from the values of HDOP and σ_{UERE}. The probability that the computed location is within a circle of radius drms from the true location depends on the ratio σ_S/σ_L for the 1σ ellipse. If the two-dimensional error

distribution is close to being circular ($\sigma_S/\sigma_L \approx 1$), the probability is about 0.63; for a very elongated distribution ($\sigma_S/\sigma_L \approx 0$), the probability approaches 0.69. Two times the drms is given by

$$2\text{drms} = 2 \cdot \text{HDOP} \cdot \sigma_{UERE} \tag{11.35}$$

and the probability that the horizontal error is within a circle of radius 2drms ranges between 0.95 and 0.98 depending on the ratio σ_S/σ_L. The 2-drms value is commonly taken as the 95% limit for the magnitude of the horizontal error.

Another common metric for horizontal errors is *circular error probable* (CEP). The CEP is defined as the radius of a circle that contains 50% of the error distributions when centered at the correct (i.e., error-free) location. Thus, the probability that the magnitude of the error is less than the CEP is precisely 1/2. The CEP for a two-dimensional Gaussian random variable can be approximated by the formula

$$\text{CEP} \approx 0.59\left(\sigma_L + \sigma_S\right) \tag{11.36}$$

assuming it is zero mean. For a derivation of this and other approximations, see [2].

The CEP can also be estimated in terms of drms and, using (11.34), in terms of HDOP and σ_{UERE}. This is convenient since HDOP is widely computed in GNSS applications. Figure 11.10 presents curves giving the probability that the magnitude of the error satisfies $|d\mathbf{R}| \leq k$ drms as a function of k for different values of the ratio σ_S/σ_L. (The horizontal error is assumed to have a zero-mean two-dimensional Gaussian distribution.) For k equal to 0.75, one obtains a probability in the range 0.43 to 0.54. Hence, one has the approximate relation

$$\text{CEP} \approx 0.75 \text{ drms} = 0.75 \cdot \text{HDOP} \cdot \sigma_{UERE} \tag{11.37}$$

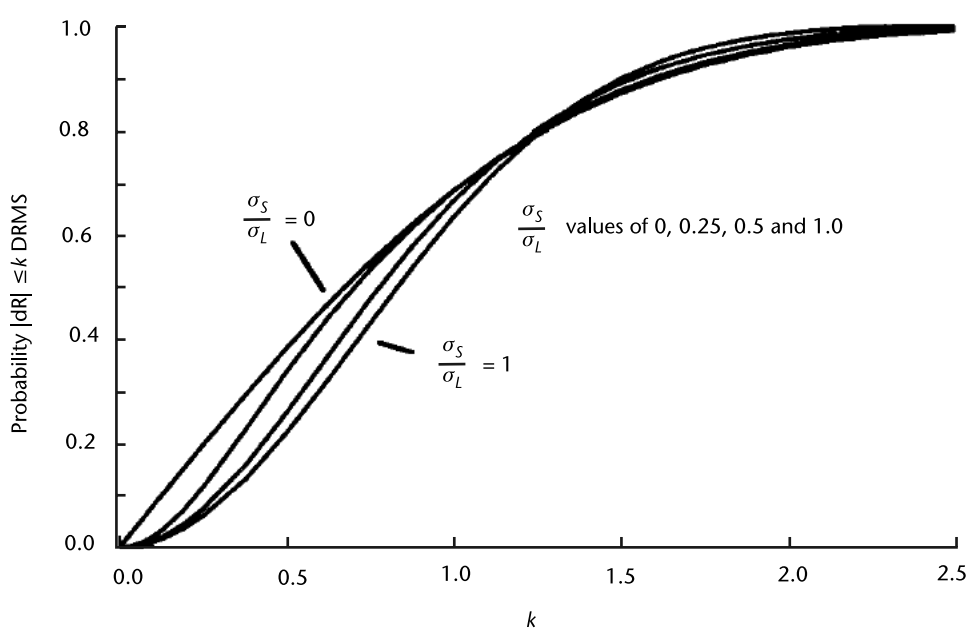


Figure 11.10 Cumulative distribution of radial error for various values of σ_S/σ_L for a two-dimensional Gaussian random variable.

It is interesting to note that for $k = 1.23$, the probability that $|d\mathbf{R}| \leq k$ drms is roughly 0.78, almost independent of σ_S/σ_L. The probabilities associated with several other values of k are summarized in Table 11.1.

As an application of these formulations, for an average global HDOP of 0.9 corresponding to the GPS 27-satellite constellation and with $\sigma_{UERE} = 0.6$m, estimates for the CEP, the 80% point, and the 95% point for the magnitude of the horizontal error are given as follows:

$$\begin{aligned} \text{CEP}_{50} &\approx 0.75 \cdot \text{HDOP} \cdot \sigma_{UERE} = 0.75 \times 0.9 \times 0.6 = 0.4 \text{ m} \\ \text{CEP}_{80} &\approx 1.28 \cdot \text{HDOP} \cdot \sigma_{UERE} = 1.28 \times 0.9 \times 0.6 = 0.7 \text{ m} \\ \text{CEP}_{95} &\approx 2.0 \cdot \text{HDOP} \cdot \sigma_{UERE} = 2.0 \times 0.9 \times 0.6 = 1.1 \text{ m} \end{aligned} \tag{11.38}$$

For applications where three-dimensional error distributions are of interest, one final commonly used metric is *spherical error probable* (SEP), which is defined as the radius of a sphere centered at the true position that contains 50% of the measured positions.

11.2.4 Weighted Least Squares

Oftentimes, the UEREs among the visible satellites are not well described as being independent and identically distributed. In such circumstances, the least-squares position estimate is not optimal. As derived in Appendix A, if the pseudorange errors are Gaussian and the covariance of UEREs for the visible satellites is given by the matrix **R**, then the optimal solution for user position is given by the *weighted least squares* (WLS) estimate

$$\Delta\mathbf{x} = \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1}\Delta\boldsymbol{\rho} \tag{11.39}$$

(Note that, as with the ordinary least squares solution, we are truly solving for a correction to an initial estimate of the user position and clock error.) Equation (11.39) collapses to (11.5) in the case when $\mathbf{R} = \sigma_{UERE}^2\mathbf{I}$ with **I** equal to the $n \times n$ identity matrix, as expected since this case corresponds to our original independent and identically distributed assumption. For a general matrix **R,** (11.39) can be

Table 11.1 Approximate Formulas for the Magnitude of the Horizontal Error

*Approximation Formula**	*Probability Range*
$\text{CEP}_{50} \approx 0.75$ HDOP σ_{UERE}	0.43 to 0.54
$\text{CEP}_{80} \approx 1.28$ HDOP σ_{UERE}	0.80 to 0.81
$\text{CEP}_{90} \approx 1.6$ HDOP σ_{UERE}	0.89 to 0.92
$\text{CEP}_{95} \approx 2.0$ HDOP σ_{UERE}	0.95 to 0.98

*CEP_{xx} is defined as the radius of the circle that when centered at the error-free location includes xx% of the error distribution. Hence, CEP_{50} = CEP.

thought of as implementing an optimal weighting of pseudorange measurements based upon their relative noise levels and relative importance for each estimated quantity.

As one example of an error covariance matrix, consider the single-frequency GNSS user whose pseudorange measurement errors are dominated by residual ionospheric delays. When models such as those discussed in Section 10.2.4.1 are used to correct for ionospheric delays, residual errors for single-frequency users are highly correlated. This correlation results because when the model overestimates or underestimates the vertical ionospheric delay, it tends to similarly overestimate or underestimate all of the slant ionospheric delays, introducing a positive correlation between residual errors between satellites. The covariance matrix of residual ionospheric errors can be approximated as

$$\mathbf{R} = \sigma_{iv}^{2} \begin{bmatrix} m^2(el_1) & m(el_1)m(el_2) & \cdots & m(el_1)m(el_n) \\ m(el_1)m(el_2) & m^2(el_2) & & \\ \vdots & & \ddots & \\ m(el_1)m(el_n) & & & m^2(el_n) \end{bmatrix} \tag{11.40}$$

where σ_{iv}^2 is the residual vertical ionospheric delay variance, which could be approximated as some fraction of the vertical delay model estimate. The *ij*th element of the matrix in (11.40) are the products of two ionospheric mapping functions, $m(el)$ [e.g., (10.22) could be used], corresponding to the elevation angles (el) for satellite i and j.

Another typical example of a covariance matrix is a diagonal matrix whose diagonal elements are obtained using an approximation for pseudorange error variance versus elevation angle, usually a monotonically increasing function as elevation angle decreases (see, e.g., [3]). The use of such a covariance matrix in a WLS solution deweights low-elevation angle satellites that are expected to be noisier due to typical characteristics of multipath and residual tropospheric errors.

Another diagonal matrix useful for multiconstellation receivers is one that models the UERE for each satellite differently depending on the constellation to which it belongs. This weighting would be appropriate if the observed signal-in-space error characteristics of satellites from each constellation are significantly different from each other.

11.2.5 Additional State Variables

Thus far, we have focused on estimation of the user's (x, y, z) position coordinates and clock bias. The complete set of parameters that are estimated within a GNSS receiver, often referred to as the *state* or *state vector*, may include a number of other variables. For instance, if in addition to pseudorange measurements, Doppler measurements (from a frequency locked loop or phase locked loop) are available, or differenced carrier-phase measurements, then velocity in each of the three coordinates ($\dot{x}, \dot{y}, \dot{z}$) and clock drift, $\dot{t}_u$, may also be estimated. The same least-squares or WLS techniques used for position estimation may be used and the same DOPs apply. The only difference is that in the linearization process, satellite velocities and initial

estimates of user velocity and clock drift are employed. Also, for precise velocity estimation it is important to account for the fact that satellite geometry is slowly changing with time; see, for example, [4].

Additional state variables can also be helpful to address system time offsets when using measurements from multiple GNSS constellations [5]. If one 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. There are several available methods. First and easiest, some GNSS satellites transmit within their broadcast navigation data certain GNSS system time offsets. Such data (e.g., from Message Type 35 within the GPS CNAV data, see Section 3.7.3.2) can be utilized to correct the pseudoranges from some other GNSS constellations so that no additional state variables are required within the receiver estimator.

A second method for accounting for GNSS system time offsets is for the user equipment to directly estimate these parameters from the pseudorange measurements. For each additional GNSS constellation, this requires one additional state variable (e.g., the difference in system time between the primary GNSS constellation and the secondary constellation). The connection matrix **H** (11.3) needs to be modified as well. Let n be the number of satellites visible from the primary GNSS constellation and m be the number of satellites visible from the secondary GNSS constellation. Then the appropriate modification to **H** would be to add a column:

$$\mathbf{H} = \begin{bmatrix} a_{x1} & a_{y1} & a_{z1} & 1 & 0 \\ a_{x2} & a_{y2} & a_{z2} & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{xn} & a_{yn} & a_{zn} & 1 & 0 \\ a_{x,n+1} & a_{y,n+1} & a_{z,n+1} & 1 & 1 \\ a_{x,n+2} & a_{y,n+2} & a_{z,n+2} & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{x,n+m} & a_{x,n+m} & a_{x,n+m} & 1 & 1 \end{bmatrix} \tag{11.41}$$

Extensions to three or more constellations are straightforward. For each new constellation, an additional state needs to be added, and another column to **H**. This second method can often produce more accurate estimates of the system time offset than provided by the broadcast data and more accurate estimates of position. There is unfortunately a downside as well. For two constellations, a minimum of 5 satellites need to be visible to estimate user position, clock error, and the one GNSS system time offset, whereas only 4 satellites would be required using broadcast time offset data. For each additional constellation, one additional visible satellite is required.

A third method for accounting for GNSS system time offsets is to use a hybrid of the first two methods. One simple hybrid method is to use the second approach of directly estimating the system time offsets when there are sufficient satellites visible, but then fixing the estimated system time offsets if the number of visible satellites drops below the minimum required [6]. A higher-performance hybrid method is to estimate the system time offsets, but utilize not just the pseudorange

measurements but also (when available) appropriately weighted broadcast system time offset data [7].

There are numerous other state variables that may be encountered in practice. These can include vertical tropospheric delays [8] and many state variables associated with other sensors. The latter are discussed in detail in Chapter 13.

11.2.6 Kalman Filtering

The least-squares and WLS solutions that were described in Chapter 2 and previously in this chapter have utilized a set of pseudorange measurements at one snapshot in time, along with initial estimates of the user position and clock, to derive an improved estimate of the user's position and clock error at that instant. In practice, the user frequently has access to an entire sequence of measurements over time. Past measurements may often be useful towards obtaining a more accurate PVT estimation. For instance, a stationary user can average least-square position estimates over an hour, a day, or longer, to obtain a more accurate estimate of their position than would be possible using just the latest set of measurements. In principle, even the most agile user can obtain some benefit from incorporating past measurements into their position estimator, provided that it is possible to accurately model the motion of the platform over time and also to model the progression of user clock errors with time. The most common algorithm used to incorporate past measurements in GNSS PVT applications is referred to as a *Kalman filter*. Kalman filters also facilitate the blending of GNSS measurements with measurements from other sensors, and are discussed in detail in Chapter 13.

11.3 GNSS Availability

Availability of a navigation system is the percentage of time that the services of the system are usable. Availability is an indication of the ability of the system to provide a usable navigation service within a specified coverage area. Availability is a function of both the physical characteristics of the environment and the technical capabilities of the transmitter facilities [9]. In this section, GNSS availability is discussed under the assumption that usable navigation service can be equated to GNSS accuracy meeting a threshold requirement. It should be noted that some applications include additional criteria, for example, the provision of integrity (see Section 11.4), that must be met for the system to be considered available.

As discussed in Section 11.2.1, GNSS accuracy is generally expressed by

$$\sigma_p = \mathrm{DOP} \cdot \sigma_{UERE}$$

where σ_p is the standard deviation of the positioning accuracy and σ_{UERE} is the standard deviation of the satellite pseudorange measurement error. Representative σ_{UERE} values are provided in Section 10.3 . The DOP factor could be HDOP, VDOP, PDOP, and so forth, depending on the dimension for which GNSS accuracy is to be determined. The availability of the GNSS navigation function to provide a given accuracy level is therefore dependent on the geometry of the satellites for a specific location and time of day.

In order to determine the availability of GNSS for a specific location and time, the number of visible satellites, as well as the geometry of those satellites, must first be determined. GNSS almanac data, which contains the positions of all satellites in the constellation at a reference epoch, can easily be obtained from various sources on the Internet or as an output from some GNSS receivers. Since the orbits of the GNSS satellites are well known, the position of the satellites at any given point in time can be predicted. However, the process of determining the satellite positions at a particular point in time is not intuitive, and software is needed to perform the calculations. The remainder of Section 11.3 details the availability determination process using the nominal 24-satellite GPS constellation (see Section 3.2.1) as an example.

11.3.1 Predicted GPS Availability Using the Nominal 24-Satellite GPS Constellation

This section examines the availability of the nominal 24-satellite GPS constellation. The nominal 24-satellite constellation is defined in Section 3.2.1. Worldwide GPS coverage is evaluated from 90° N to 90° S latitude with sample points spaced every 5° (in latitude) and for a band in longitude circling the globe spaced every 5°. This grid is sampled every 5 minutes in time over a 12-hour period.

Since the GPS constellation has approximately a 12-hour orbit, the satellite coverage will then repeat itself on the opposite side of the world during the next 12 hours. (The Earth rotates 180° in the 12-hour period and the satellite coverage areas will be interchanged.) A total of 386,280 space/time points are evaluated in this analysis.

GPS availability also is dependent on the mask angle used by the receiver. By lowering the mask angle, more satellites are visible; hence, a higher availability can be obtained. However, there may be problems with reducing the mask angle to include very low elevation angles, which are discussed later in this section. The availability obtained by applying the following mask angles is examined: 7.5°, 5°, 2.5°, and 0°.

Figure 11.11 demonstrates GPS availability based on HDOP using an all-in-view solution. This figure provides the cumulative distribution of HDOP for each of the mask angles considered. The maximum value of HDOP is 2.55 for a mask angle less than or equal to 5°.

Figure 11.12 provides the availability of GPS based on PDOP for the same mask angles. This availability is lower than that for HDOP since unavailability in the vertical dimension is taken into consideration in the calculation of PDOP. The maximum value of PDOP for a 5° mask angle is 5.15, at 2.5° it is 4.7, and for a 0° mask angle the maximum value is 3.1.

Although these graphs demonstrate the improvement in availability that can be obtained when the mask angle is lowered, there is a danger in lowering it too far. During the mission planning process, signal blockage from buildings or other objects that extend higher than the set mask angle must be taken into consideration. There also is a greater potential for atmospheric delay and multipath problems at a lower mask angle.

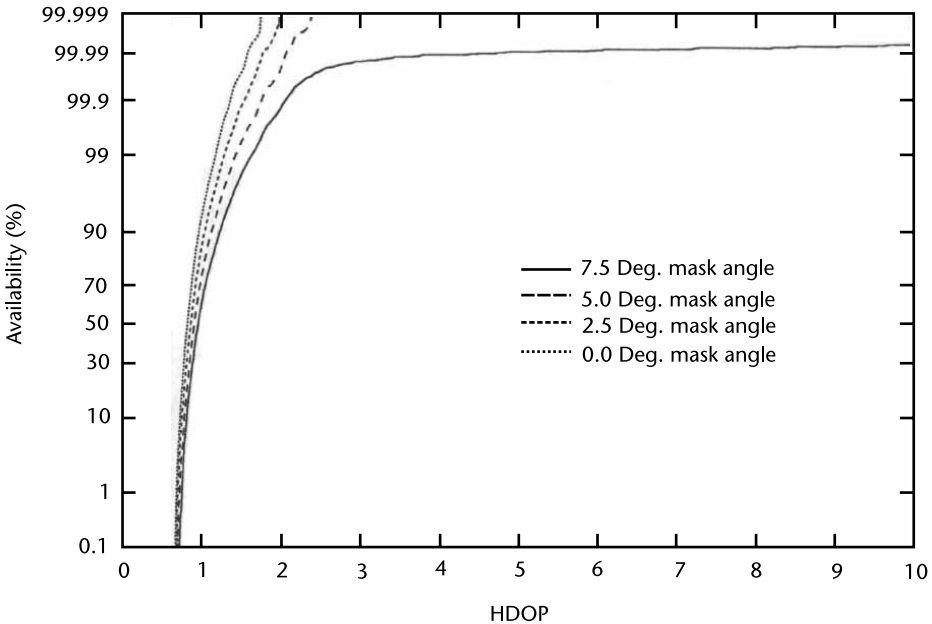


Figure 11.11 Cumulative distribution of HDOP with 7.5°, 5°, 2.5°, and 0° mask angles.

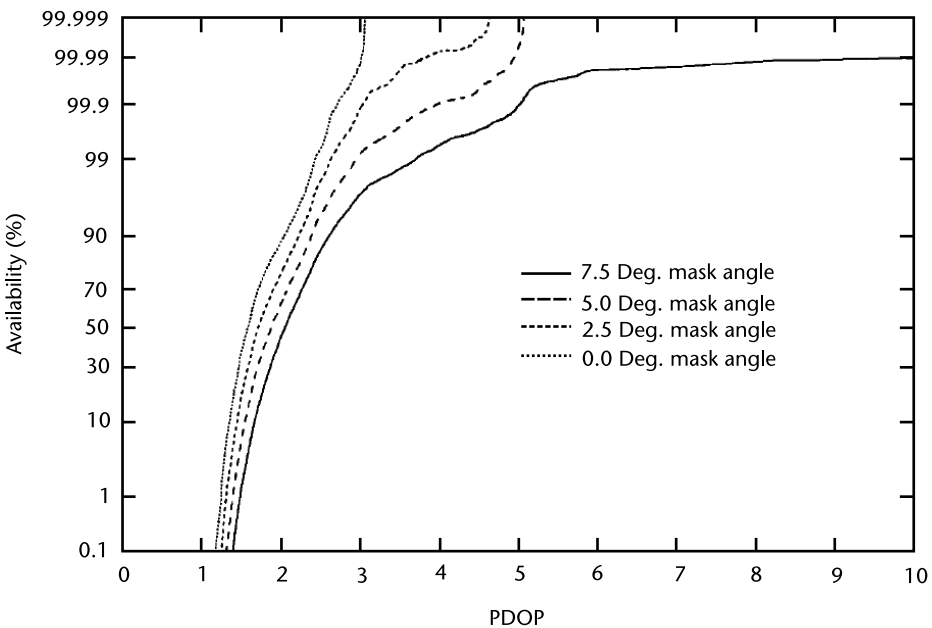


Figure 11.12 Cumulative distribution of PDOP with 7.5°, 5°, 2.5°, and 0° mask angles.

The threshold for the maximum acceptable DOP value is dependent on the desired accuracy level. The availability of GPS, therefore, will depend on the stringency of the accuracy requirement. For this analysis, availability of GPS is chosen to be defined as PDOP ≤ 6 which is commonly used as a service availability threshold in the GPS performance standards [10].

As shown in Figure 11.12, with all 24 GPS satellites operational, the value of PDOP is less than 6.0 for every location and time point analyzed at 0°, 2.5°, and 5° mask angles. Since the analysis grid is sampled every 5 minutes, there could be occurrences where PDOP is greater than 6.0 for a period of less than 5 minutes that would not be detected. Only with a 7.5° mask angle (or higher) does the GPS constellation have outages based on PDOP exceeding 6.0.

At a 7.5° mask angle, the GPS constellation provides an availability of 99.98%. Figure 11.13 displays the locations and duration of the outages that occur. The maximum outage duration is 10 minutes. The GPS constellation is designed to provide optimal worldwide coverage. As a result, when outages do occur, they are concentrated in very high and very low latitudes (above 60°N and below 60°S).

11.3.2 Effects of Satellite Outages on GPS Availability

The previous figures have demonstrated the availability of GPS when all 24 satellites are operational. However, satellites need to be taken out of service for maintenance, and unscheduled outages occur from time to time. In fact, 24 satellites may only be available 72% of the time, while 21 or more satellites are expected to be operational at least 98% of the time [10].

To examine the effect that a reduced constellation of satellites has on the availability of GPS, the analysis is now repeated using the same worldwide grid, but removing one, two, and three satellites from the nominal 24-satellite constella-

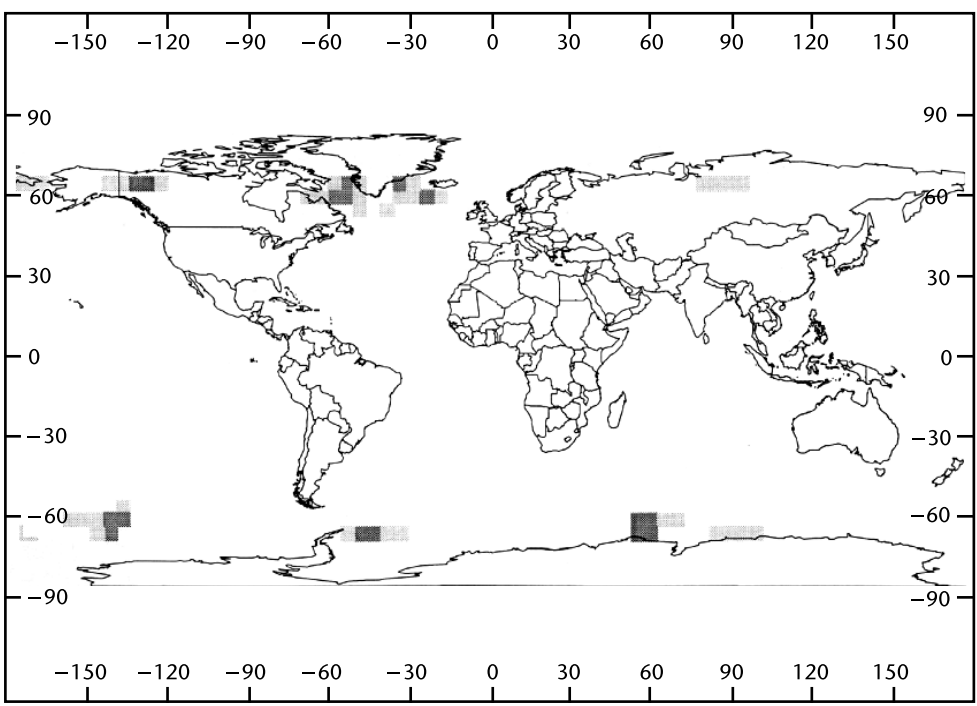


Figure 11.13 Availability of the GPS constellation (PDOP ≤ 6) with a 7.5° mask angle.

tion. Since a 5° mask angle is commonly used, it is the only one considered for this portion of the analysis.

The availability of GPS when satellites are removed from the constellation is very much dependent on which satellites, or combinations of satellites, are taken out of service. The Aerospace Corporation has performed a study that determined cases of one, two, and three satellite failures that resulted in the least, average, and greatest impact on availability [11]. The choices for satellites to be removed in this analysis were based on those satellites that caused an average impact on GPS availability.

The orbital positions of the GPS satellites removed from the constellation are given in the following list:

- Average one satellite—SV A3;
- Average two satellites—SVs A1 and F3;
- Average three satellites—SVs A2, E3, and F2.

(See Section 3.2.1 for satellite identification and orbital location information.)

Figures 11.14 and 11.15 display the cumulative distribution of HDOP and PDOP with up to three satellites removed from the constellation and applying a 5° mask angle. These plots demonstrate the increasing degradation in system performance as more satellites are removed from the constellation.

The availability of GPS, based on PDOP ≤ 6 and a 5° mask angle, is 99.969% with one satellite out of service. The location and duration of the resulting outages are displayed in Figure 11.16. The maximum outage duration that occurs is 15 minutes.

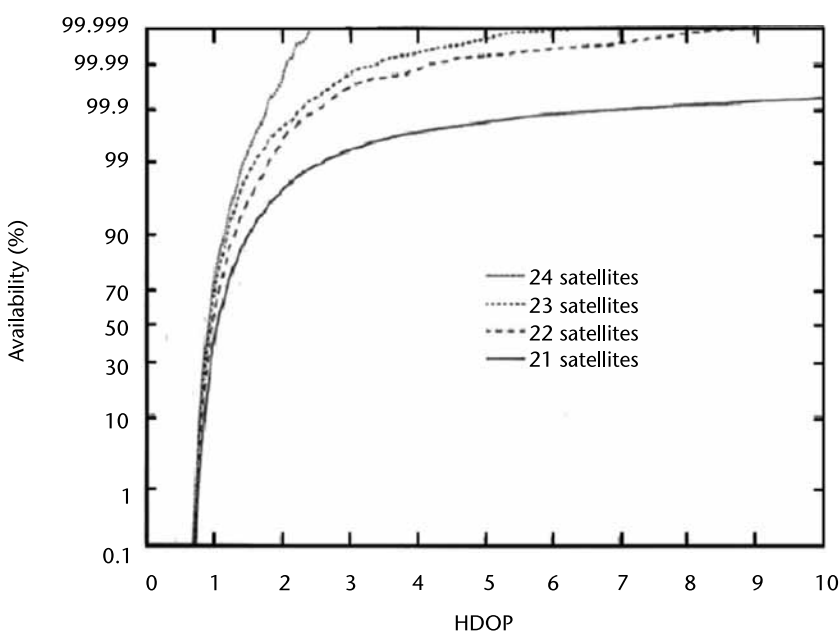


Figure 11.14 Cumulative distribution of HDOP with 5° mask angle cases of 24, 23, 22, and 21 satellites.

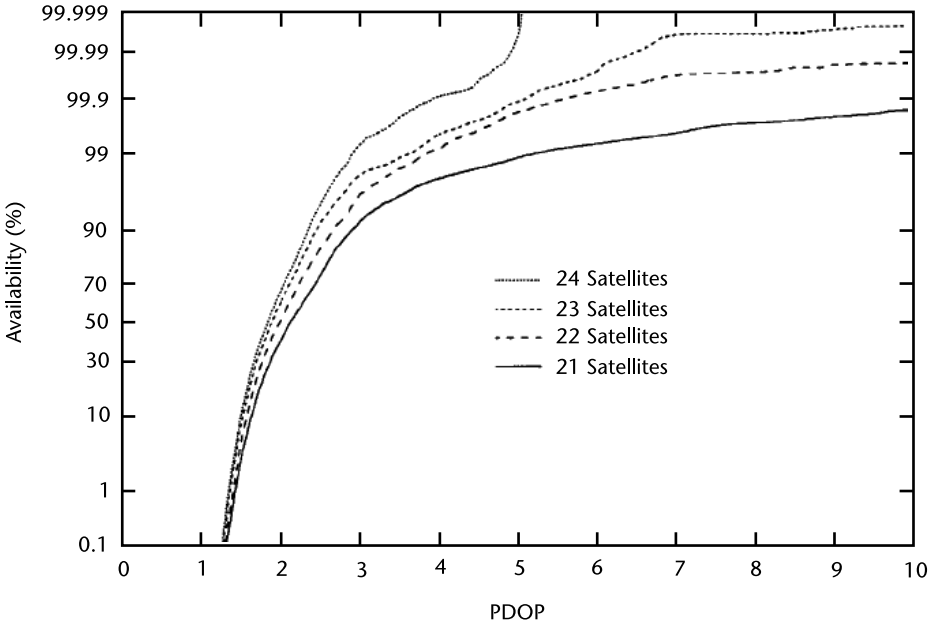


Figure 11.15 Cumulative distribution of PDOP with 5° mask angle cases of 24, 23, 22, and 21 satellites.

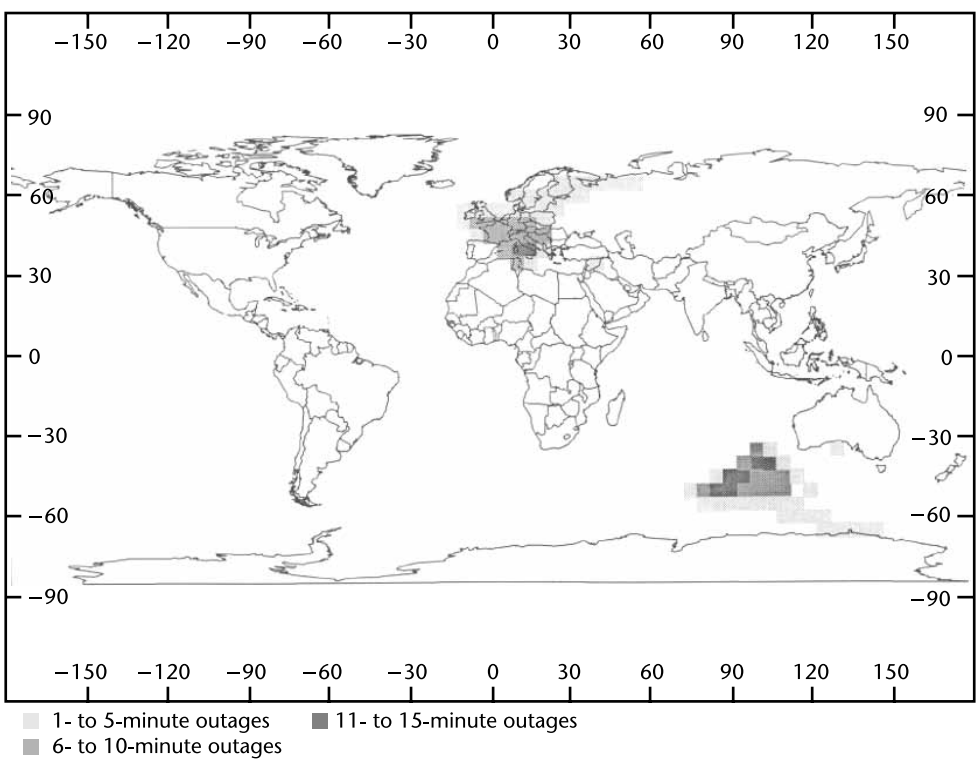


Figure 11.16 Availability of the GPS constellation with a 5° mask angle with one satellite removed from the constellation.

The effects of two satellites out of service are shown in Figure 11.17. Outages now last up to 25 minutes in several locations, but there are only a couple of occurrences of these during the day. The majority of the outages are 10 minutes or less. This constellation provides an availability of 99.903%.

With three satellites out of service, the overall availability of the GPS constellation drops to 99.197%. The number of outage occurrences increases dramatically and outages now last up to 65 minutes. The locations and corresponding duration of these outages are shown in Figure 11.18.

The scenario of having three satellites out of service at the same time should be a very rare occurrence. However, if it were to happen, the user could examine the predicted availability over the course of the day and plan the use of GPS accordingly.

As mentioned previously, the determination of satellite positions and the resulting GNSS availability for any location and point in time is not intuitive and requires software to perform the calculations. GNSS prediction software is commercially available that allows a user to determine GNSS coverage for a single location or for multiple locations. Some GNSS receiver manufacturers also include prediction software with the purchase of a receiver. The typical input parameters used to perform GNSS availability predictions are as follows:

- *GNSS almanac data:* The position of the satellites at a reference epoch may be obtained from several different sources: various Web sites or a GNSS receiver that outputs almanac data.

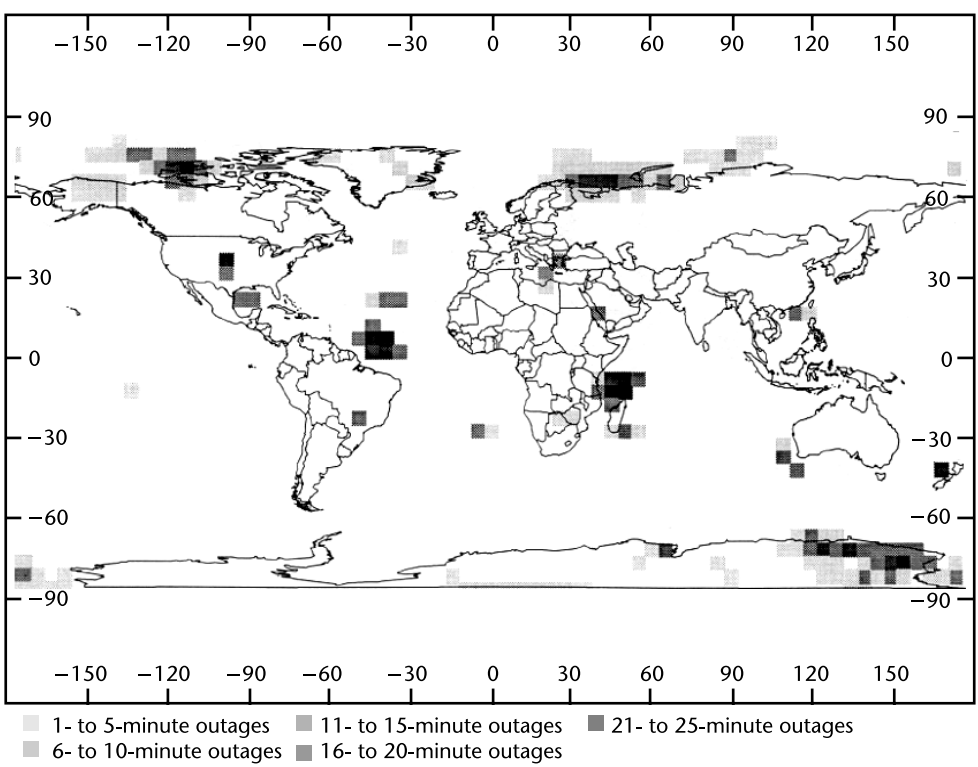


Figure 11.17 Availability of the GPS constellation with a 5° mask angle with two satellites removed from the constellation.

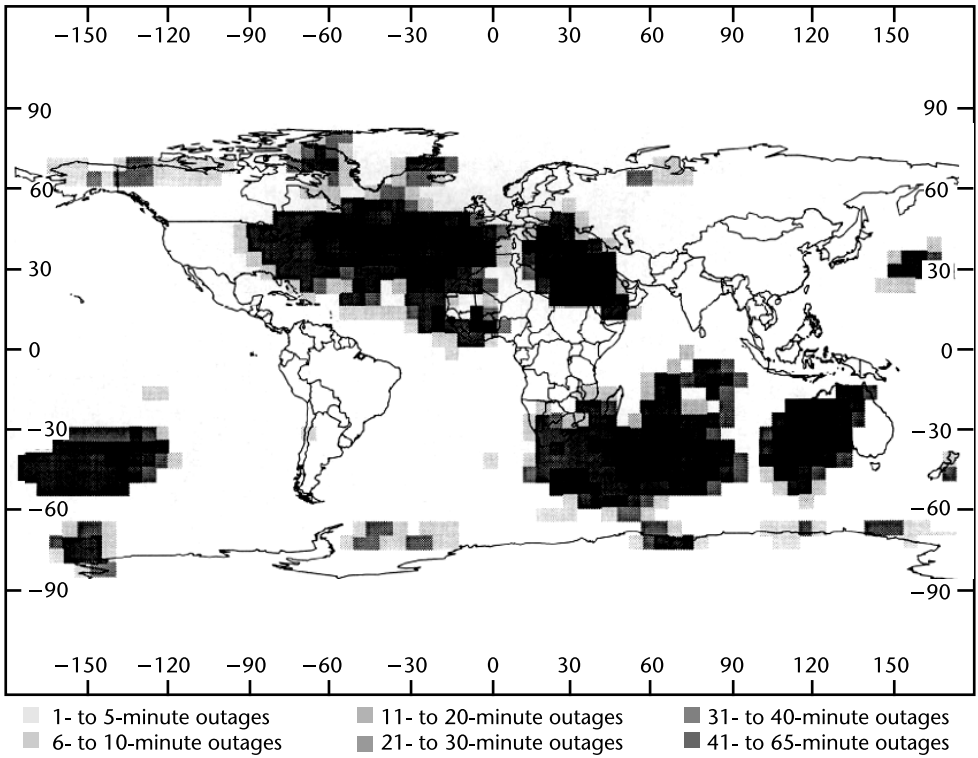


Figure 11.18 Availability of the GPS constellation with a 5° mask angle with three satellites removed from the constellation.

- *Location:* Latitude, longitude, and altitude of the location(s) for which the prediction is to be performed.
- *Date of prediction:* The date for which the prediction is to be performed. GNSS almanacs can typically be used to accurately predict a week or more in the future.
- *Mask angle:* The elevation angle above the horizon at which satellites are considered visible by the GNSS receiver.
- *Terrain mask:* The azimuth and elevation of terrain (buildings, mountains, and so forth) that may block the satellite signal can be entered into the program to ensure an accurate prediction.
- *Satellite outages:* If any satellites are currently out of service, their status will be reflected in the almanac data. However, if satellites are scheduled for maintenance for a prediction date in the future, the software allows the user to mark those satellites unusable. This data can be obtained from various Web sites.
- *Maximum DOP:* As discussed previously, in order to determine availability, a maximum DOP threshold must be set (e.g., PDOP = 6). If the DOP exceeds that value, the software will declare GNSS to be unavailable. Other applications may use criteria other than DOP as the availability threshold. This will be discussed further in Section 11.4 for aviation applications.

Once these parameters have been input into the software, the prediction can be performed. A prediction was performed for a user in Boston (42.3586°N 71.0638°W) on January 3, 2017. Figure 11.19 shows the location of the GPS satellites for the selected location at a snapshot in time (12:30 UTC), as well as the ground track for PRN 1.

Figure 11.20 is a sky plot (see Section 11.2.2). As usual, the outermost circle represents 0° elevation, or the horizon. The second circle is at 15° elevation. The third is at 30°, and each circle increases by 15°. A mask angle of 5° was used (i.e., satellites that were below 5° elevation angle were considered to be not visible due to local terrain, foliage, or man-made structures).

Figure 11.21 displays the rise and set time for the 31 operational GPS satellites at the selected location over a 24-hour period using a 5° mask angle. This type of graph can be very useful for a researcher who wants to plan an experiment with a particular set of satellites and does not want the satellite geometry to change significantly due to a rising or setting satellite. Figure 11.21 also displays the number of visible satellites and PDOP over the 24-hour period. The number of visible satellites ranged from 8 to 12, and the PDOP varied between 1.2 and 2.4.

Figures 11.22 and 11.23 demonstrate the impact that sky blockage or scintillation in one part of the sky can have on availability. (These disruptions are discussed in Chapter 9.) For these figures, it is assumed that the receiver cannot track PRNs 2, 5, 6, 9, and 29, perhaps due to a building or severe scintillation obscuring the sky towards the north and northeast. As shown in the sky plot in Figure 11.22, at 12:30 UTC in Boston, the user can only see 5 satellites due to a partial-sky blockage as compared to the 10 satellites that were visible at the same location when the sky was assumed to be clear as in Figure 11.20. When fewer satellites are visible, performance obviously suffers. The performance degradation due to the loss of 5 satellites over 24 hours is shown in Figure 11.23. (Note that sky blockage would not be expected to affect the same set of satellites over a day, but nonetheless these results are illustrative of the severe performance degradation possible due to

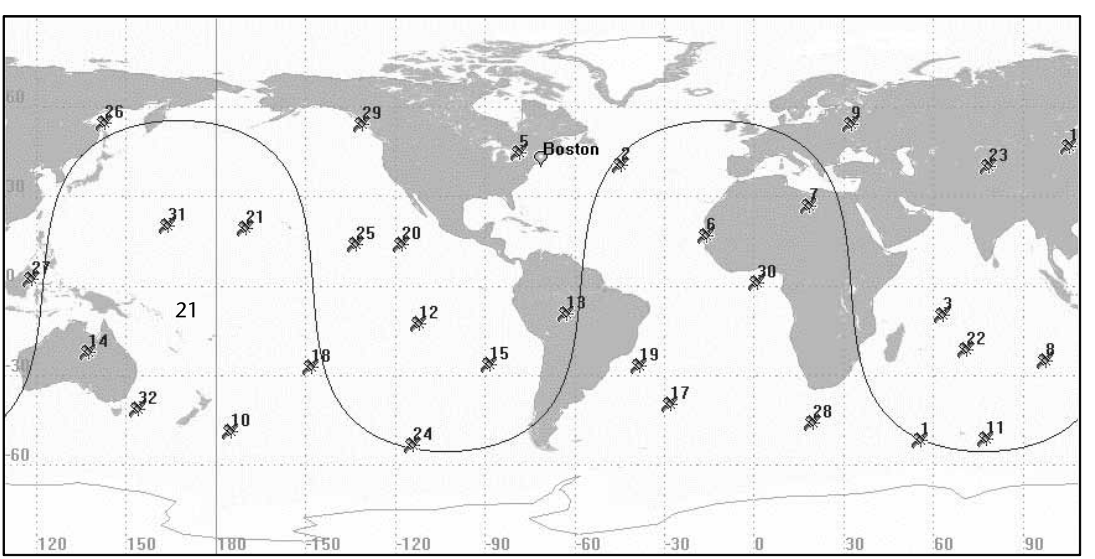


Figure 11.19 Locations of GPS satellites worldwide. (Courtesy of Evan Lewis.)

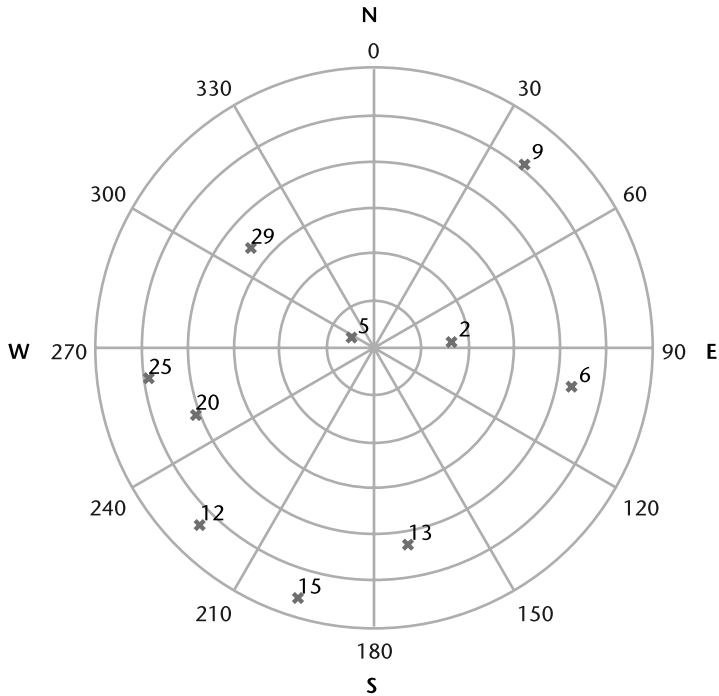


Figure 11.20 Sky plot of GPS satellite visibility. (Courtesy of Evan Lewis.)

partial-sky blockage.) For a significant portion of the day, PDOP spikes up to over 10 and even becomes infinite for a short while due to only 3 satellites being visible at around 11:25 UTC.

11.4 GNSS Integrity

In addition to providing a position, navigation, and timing function, some navigation systems must have the ability to provide timely warnings to users when the system should not be used. This capability is known as the *integrity* of the system. Integrity is a measure of the trust which can be placed in the correctness of the information supplied by the total system. Integrity includes the ability of a system to provide valid and timely warnings to the user, known as alerts, when the system must not be used for the intended operation [9].

11.4.1 Discussion of Criticality

Anomalies can occur in GNSS, caused by either failures in the satellite or the ground control networks, which result in unpredictable range errors above the operational tolerance. These errors are different from the predictable degraded accuracy resulting from poor satellite geometry, which was discussed in the previous section. Integrity anomalies should be rare, occurring only a few times per year [10, 12, 13], but can be critical, especially for air navigation.

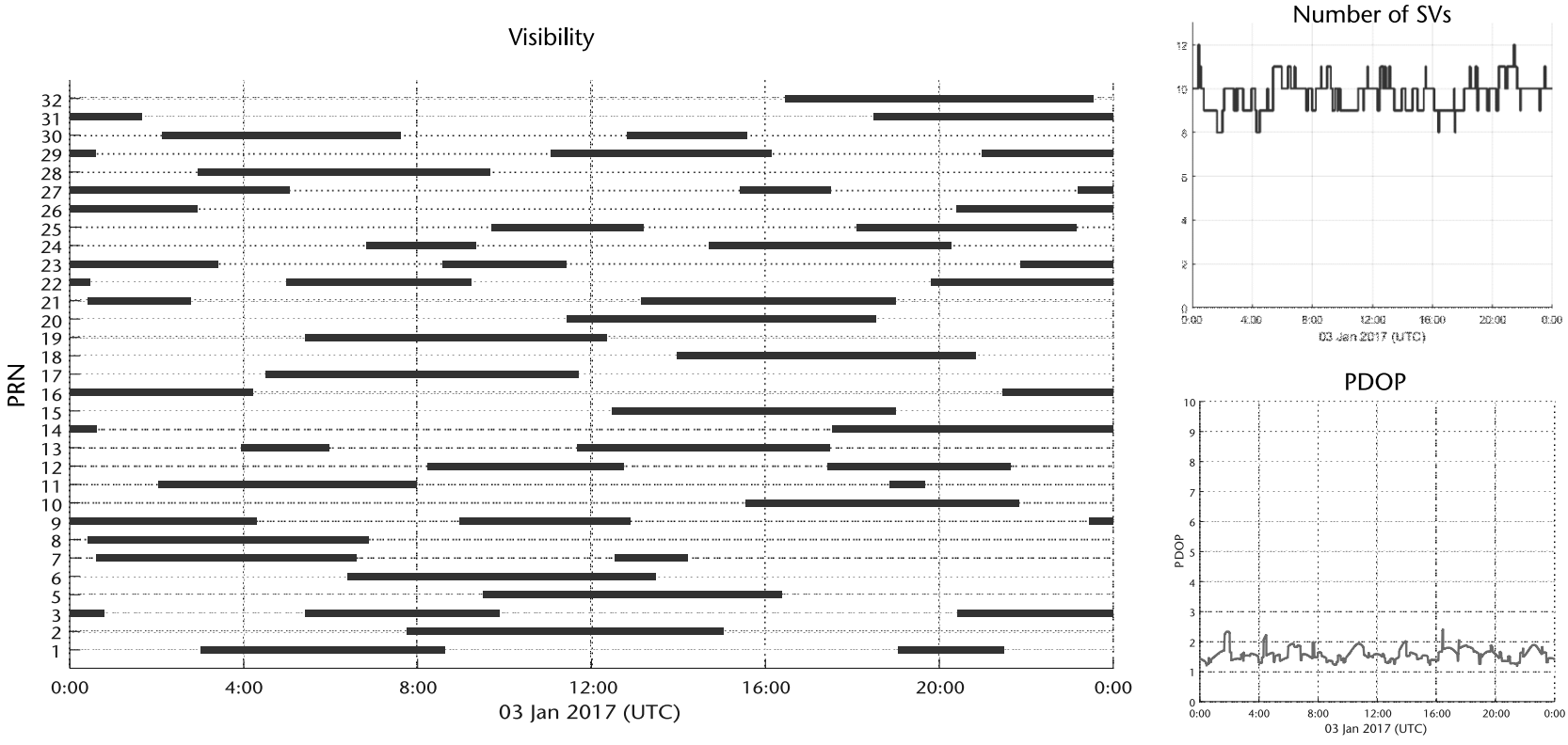


Figure 11.21 Satellite visibility and PDOP over a 24-hour period. (Courtesy of Evan Lewis.)

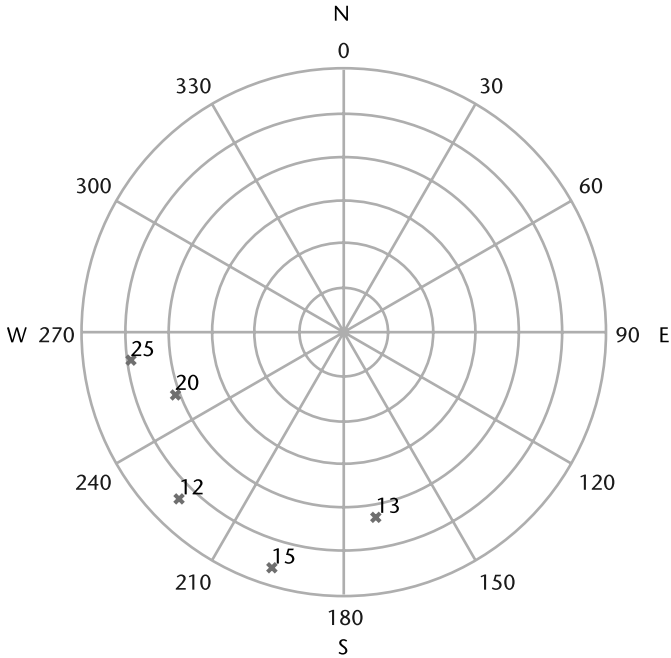


Figure 11.22 Sky plot of GPS satellite visibility with satellites 2, 5, 6, 9, and 29 removed. (Courtesy of Evan Lewis.)

11.4.2 Sources of Integrity Anomalies

There are four main sources of integrity anomalies: system allocated signal-in-space (SIS) aberrations, space segment allocated SIS aberrations, control segment allocated SIS aberrations, and user segment SIS aberrations [14]. Satellite clock anomalies are due to frequency standard problems such as random phase run-off, a large frequency jump, or a combination of both. Jumps in the GPS satellite clocks have been observed when the beam current or temperature of the frequency standard has varied greatly. Clock jumps and other clock anomalies have also been observed for the GLONASS satellites [15]. Neither Galileo nor BeiDou are yet fully operational globally, so anomalies within these GNSS core constellations have not yet received the same level of scrutiny as within GPS and GLONASS. Overall, clock anomalies are the most prevalent source of GNSS space segment anomalies and the most common source of major service anomalies. These anomalies can result in thousands of meters of range error.

For GPS, the first-generation Block I satellites experienced many more clock anomalies than the Block II generation of satellites [12] and did not have the radiation hardening against the space environment that has been built into the Block II satellites. Consequently, Block I satellites were subject to bit hits, which affect the navigation message, as well as C-field-tuning word hits. The C-field-tuning register that aligns the cesium beam is affected by solar radiation. Changing the bits that account for the alignment/direction of the cesium beam has in some instances resulted in ranging errors of thousands of meters in only a few minutes.

For GLONASS, the number of observed clock and other anomalies is also diminishing with system maturity. GLONASS signal-in-space anomalies were analyzed for the period from January 2009 to August 2012 in [15], and 192 potential

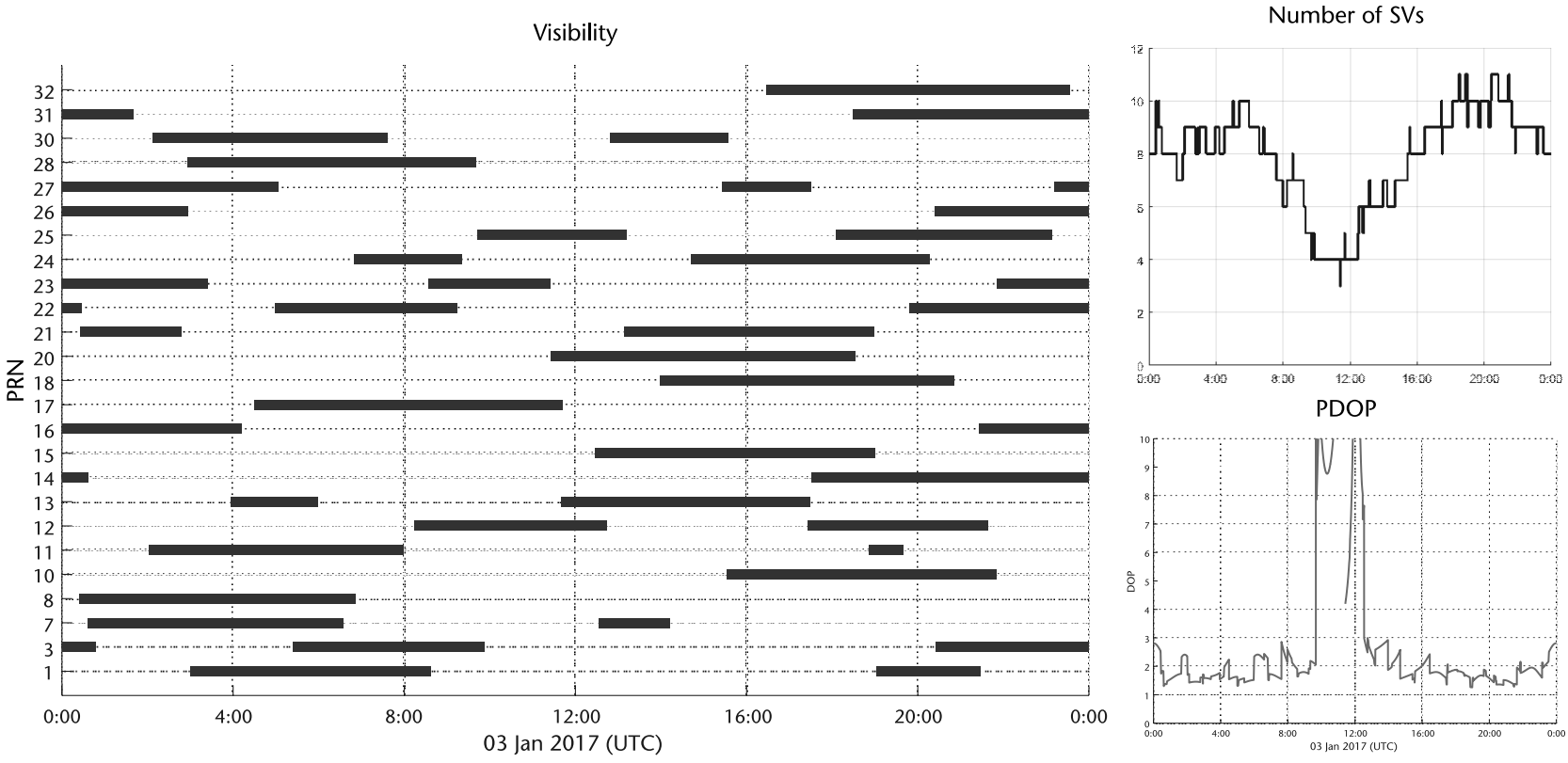


Figure 11.23 Satellite visibility and PDOP over a 24-h period with satellites 2, 5, 6, 9, and 29 removed. (Courtesy of Evan Lewis.)

anomalies were observed; 92% of the anomalies over this period were identified as being clock-related. The rate of occurrence of the anomalies dropped by a factor of 10 from 2009 to 2012.

Abnormally large broadcast ephemeris errors and errors in other broadcast navigation data elements have also been observed on rare occasions for both GPS and GLONASS. For both systems multiple instances of broadcast ephemeris errors in excess of 400m have been observed [16]. On April 1, 2014, all the GLONASS satellites broadcast erroneous ephemeris data, with errors of up to 200 km in magnitude [17]. It took approximately 10 hours to rectify the problem. In January 2016, multiple GPS satellites broadcast erroneous information regarding the offset between GPS time and UTC [18]. In 2002, incorrect single-frequency ionospheric correction data was broadcast by GPS. Single-frequency receivers may have experienced ranging errors of up to 16m before the problem was detected.

Other types of integrity anomalies can result in smaller ranging errors. An example of this occurred on GPS SVN 19. After approximately 8 months on orbit, an anomalous condition developed on the satellite that resulted in carrier leakage on the observed L1 signal spectrum which is normally carrier suppressed. In this case, no control segment problems were observed or user equipment problems were reported, so the SV was left to operate in the off-nominal mode. No incident reports or problems regarding the SVN 19 C/A code occurred until March 1993 during FAA field tests using differential navigation for aided landings. The differential navigation solution was corrupted with a 4-m bias [19]. The GPS SVN 19 event led to an immense amount of research on GNSS navigation signal quality and the establishment of *signal quality monitoring* (SQM) within some high-integrity differential systems such as those used for aviation.

The GPS ground monitoring network in the past had blind spots where it could not see some satellites some of the time [12]. Therefore, if an integrity problem were to occur, it may not have been detected immediately. An example of this occurred on July 28, 2001, when SVN22 experienced a clock failure over the southern Pacific Ocean region resulting in user range errors in excess of 200,000m. For about a half an hour, this was undetectable by the GPS operational control segment because the satellite was not in view of any OCS monitor stations [19]. With the addition of the NGA monitor stations to the GPS control segment (see Section 3.3), the blind spot has been eliminated. However, even when an error is detected manual intervention is required by the ground operators. The operators must decide a course of action, steer a dish antenna towards the satellite, and issue a command to change the operation of the satellite. This process requires several minutes to accomplish when the operators are already in communication with the satellite and tens of minutes if they are not. Certain payload failure types can be automatically detected onboard the satellites and the satellite can switch its signal to a nonstandard code automatically without operator intervention. Such fault types typically only last seconds or are removed before they become large enough to be harmful.

The GNSS service providers are continuously working to minimize integrity anomalies as much as possible by installing redundant hardware, robust software, and providing training to prevent human error. However, as previously stated, the best response time for many faults may still be several minutes, which is insufficient for aviation and certain other applications. There are methods by which the user is independently able to be notified of a satellite anomaly if it does occur.

11.4.3 Integrity Enhancement Techniques

The integrity problem is important for many applications, but crucial for aviation since the user is traveling at high speeds and can quickly deviate from the flight path. The integrity function becomes especially critical if GNSS is to be used as a primary navigation system.

Historically, integrity enhancements for GNSS were first developed for GPS. RTCA Special Committee 159 (SC-159), a federal advisory committee to the Federal Aviation Administration, has devoted much effort to developing techniques to provide integrity for airborne use of GPS [20]. Three methods used today for GPS integrity monitoring are receiver autonomous integrity monitoring (RAIM) (one element of a set of airborne GPS enhancements defined by the International Civil Aviation Organization [ICAO] as aircraft-based augmentation systems [ABAS]), satellite-based augmentation systems (SBAS), and ground-based augmentation systems (GBAS).

This section primarily concentrates on RAIM since SBAS and GBAS are differential techniques discussed in more detail in Chapter 12. Today, the preponderance of high-integrity GNSS equipment only utilizes GPS of the core constellations, so the discussion within the remainder of this section is of necessity GPS-centric. Efforts are underway on the development of standards for multiconstellation equipment [20] using extensions of the techniques discussed in this section.

11.4.3.1 RAIM and FDE

The use of stand-alone GPS or GPS in conjunction with use of ranging sources from other satellites such as geostationary satellites, GLONASS, Galileo, and/or BeiDou where integrity is provided by RAIM and Fault Detection and Exclusion (FDE) is referred to as an aircraft-based augmentation system (ABAS). As noted earlier, current airborne equipment standards only fully address GPS of the core GNSS constellations. The RAIM algorithm is contained within the receiver, hence the term "autonomous" monitoring. RAIM is a technique that uses an overdetermined solution to perform a consistency check on the satellite measurements [21].

RAIM algorithms require a minimum of five visible satellites in order to detect the presence of an unacceptably large position error for a given mode of flight. If a failure is detected, the pilot receives a warning flag in the cockpit which indicates that GPS should not be used for navigation. Certified GPS receivers that contain FDE, an extension of RAIM that uses a minimum of six visible satellites, can not only detect the faulty satellite, but can exclude it from the navigation solution so the operation can continue without interruption.

The inputs to the RAIM algorithm are the standard deviation of the measurement noise, the measurement geometry, as well as the maximum allowable probabilities for a false alert and a missed detection. The output of the algorithm is the horizontal protection level (HPL), which is the radius of a circle, centered at the true aircraft position that is assured to contain the indicated horizontal position with the given probability of false alert and missed detection that are discussed next. This section concentrates on the generation of HPL using a snapshot RAIM algorithm that has been developed in support of RTCA SC-159 [21].

The linearized GPS measurement equation is given as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon} \tag{11.42}$$

where $\mathbf{x}$ is the 4×1 vector whose elements are incremental deviations from the nominal state about which the linearization takes place. The first three elements are the east, north, and up position components, and the fourth element is the receiver clock bias. $\mathbf{y}$ is the $n \times 1$ vector whose elements are the differences between the noisy measured pseudoranges and the predicted ones based on the nominal position and clock bias (i.e., the linearization point). The value n is the number of visible satellites (number of measurements). $\mathbf{H}$ is the $n \times 4$ linear connection matrix between $\mathbf{x}$ and $\mathbf{y}$. It consists of three columns of direction cosines and a fourth column containing the value 1, which corresponds to the receiver clock state. $\boldsymbol{\epsilon}$ is the $n \times 1$ measurement error vector. It may contain both random and deterministic (bias) terms.

GPS RAIM is based on the self-consistency of measurements, where the number of measurements, n, is greater than or equal to 5. One measure of consistency is to work out the least squares estimate for $\mathbf{x}$, substitute it into the right side of (11.42), and then compare the result with the empirical measurements in $\mathbf{y}$. The difference between them is called the range residual vector, $\mathbf{w}$. In mathematical terms,

$$\begin{aligned}
&\hat{\mathbf{x}}_{\mathrm{LS}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{y} \qquad \text{(least squares estimate)} \\
&\hat{\mathbf{y}}_{\mathrm{LS}} = \mathbf{H}\hat{\mathbf{x}}_{\mathrm{LS}} \\
&\mathbf{w} = \mathbf{y} - \hat{\mathbf{y}}_{\mathrm{LS}} = \mathbf{y} - \mathbf{H}\left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{y} = \left[\mathbf{I}_n - \mathbf{H}\left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\right]\mathbf{y} \\
&= \left[\mathbf{I}_n - \mathbf{H}\left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\right](\mathbf{H}\mathbf{x} + \boldsymbol{\epsilon}) = \left[\mathbf{I}_n - \mathbf{H}\left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\right]\boldsymbol{\epsilon}
\end{aligned} \tag{11.43}$$

Since $\boldsymbol{\epsilon}$ is not known to the user aircraft, the last line of (11.43) is only used in simulations.

Let

$$\mathbf{S} \equiv \mathbf{I}_n - \mathbf{H}\left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}} \tag{11.44}$$

where $\mathbf{I}_n$ is the $n \times n$ unit matrix. Then, the $n \times 1$ range residual vector, $\mathbf{w}$, is given as $\mathbf{w} = \mathbf{S}\mathbf{y}$ (used in practice) or $\mathbf{w} = \mathbf{S}\boldsymbol{\epsilon}$ (used in the simulations). The range residual vector, $\mathbf{w}$, could be used as a measure of consistency. However, this is not ideal because there are four constraints (associated with the four unknown components of the vector $\mathbf{x}$) among the n elements of $\mathbf{w}$, which obscure some of the aspects of the inconsistency that are of interest. Therefore, it is useful to perform a transformation that eliminates the constraints and transforms the information contained in $\mathbf{w}$ into another vector known as the parity vector, $\mathbf{p}$.

Performing a transformation on $\mathbf{y}$,

$$\mathbf{p} = \mathbf{P}\mathbf{y}$$

where the parity transformation matrix $\mathbf{P}$ is defined as an $(n - 4) \times n$ matrix, which can be obtained by QR factorization of the $\mathbf{H}$ matrix [22]. The rows of $\mathbf{P}$ are mutually orthogonal, unity in magnitude, and mutually orthogonal to the columns of $\mathbf{H}$.

Due to these defining properties, the resultant **p** has special properties, especially with respect to the noise [21]. If $\boldsymbol{\epsilon}$ has independent random elements that are all $N(0, \sigma^2)$, then

$$\mathbf{p} = \mathbf{P}\mathbf{w} \tag{11.45a}$$

$$\mathbf{p} = \mathbf{P}\boldsymbol{\epsilon} \tag{11.45b}$$

$$\mathbf{p}^T\mathbf{p} = \mathbf{w}^T\mathbf{w} \tag{11.45c}$$

These equations state that the same transformation matrix **P** that takes **y** into the parity vector, **p**, also takes either **w** or $\boldsymbol{\epsilon}$ into **p**. The sum of the squared residuals is the same in both range space and parity space. In performing failure detection, it is much easier to work with **p** than with **w**.

Using a case of six visible satellites as an example, the following analysis demonstrates how the parity transformation affects a deterministic error in one of the range measurements. Suppose there is a range bias error, b, in satellite 3. From (11.45b),

$$\mathbf{p} = \begin{bmatrix} P_{11} & P_{12} & P_{13} \cdots P_{16} \\ P_{21} & P_{22} & P_{23} \cdots P_{26} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ b \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{ or}$$

$$\mathbf{p} = b \times (\text{3rd column of } \mathbf{P})$$

The third column of **P** defines a line in parity space called the characteristic bias line associated with satellite 3. Each satellite has its own characteristic bias line. The magnitude of the parity bias vector induced by the range bias b, is given by
$|\text{parity bias vector}| = b \bullet \text{norm} \mid [P_{13}\ P_{23}]^T \mid$, (bias on satellite 3, assuming $b > 0$)
where $|[P_{13}\ P_{23}]^T| = \sqrt{P_{13}^2 + P_{23}^2}$.

In general,

$$(\text{range bias, b, on the ith satellite}) = \frac{(\text{norm of parity bias vector})}{(\text{norm of ith column of } \mathbf{P})}$$

The position error vector **e** is defined as: $\mathbf{e} = \hat{\mathbf{x}}_{LS} - \mathbf{x}$

$$\begin{aligned} \mathbf{e} &= \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y} - \mathbf{x} \\ &= \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\left(\mathbf{H}\mathbf{x} + \boldsymbol{\epsilon}\right) - \mathbf{x} \\ &= \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\boldsymbol{\epsilon} \quad (\text{Vector position error}) \end{aligned}$$

which, for a bias b in the ith satellite, can be written as

$$(\text{position error vector}) = \left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}} \bullet \begin{bmatrix} 0 \\ \cdot \\ b \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

These equations provide a means of getting back and forth from a bias in parity space to the corresponding bias in range space, and finally to the corresponding position error. The norm of the first two components of the position error vector provides the horizontal radial position error.

The objective is to protect against excessive horizontal position error. The RAIM algorithm must detect if the horizontal error goes beyond a certain threshold within a specified level of confidence. Since the position error cannot be observed directly, something must be inferred from the quantity that can be observed, which in this case is the parity vector.

The magnitude of the parity vector is used as the test statistic (mathematical indicator) for detection of a satellite failure. The inputs to the parity space algorithm are the standard deviation of the measurement noise, the measurement geometry, as well as the maximum allowable probabilities for a false alert and a missed detection. The output of the algorithm is the horizontal protection level (HPL), which defines the smallest horizontal radial position error that can be detected for the specified false alert and missed detection probabilities.

A false alert is an indication of a positioning failure to the pilot when a positioning failure has not occurred, as the result of a false detection. The detection threshold for the RAIM and FDE algorithms is determined by integrating the probability density function from the detection threshold to infinity so that the area under the curve is equal to the probability of a false alert, P_{FA}.

The parity space method is based on modeling the test statistic using a chi-square distribution with $n - 4$ degrees of freedom for six or more visible satellites. The sum of the squared measurement residuals has a chi-square distribution. A Gaussian distribution is used for the case where five satellites are in view. The general formulas for the chi-square density functions are provided next.

For a central chi-square,

$$f\text{cent}(\mathrm{x}) = \left[x^{((k/2)-1)}e^{-x/2}\right] \Big/ \left[2^{k/2}\Gamma(k/2)\right],\ x > 0$$
$$= 0,\ x \le 0$$

where Γ is the gamma function.

For the probability of missed detection, the noncentral chi-square density function is integrated from 0 to the chi-square detection threshold to determine λ, the noncentrality parameter that provides the desired P_{md}. The minimum detectable bias based on the selected probabilities of false alert and missed detection is denoted as pbias, where pbias= $\sigma_{UERE}\sqrt{\lambda}$.

For a noncentral chi-square,

$$f_{\text{N.C.}}(x) = \left[e^{-(x+\lambda)/2}/2^{k/2}\right]\sum_{j=0}^{\infty}\left\{\lambda^j \cdot x^{(k/2)+j-1}/\left[\Gamma\left((k/2)+j\right)\cdot 2^{2j}\cdot j!\right]\right\},\ x>0$$

$$= 0,\ x \le 0$$

where λ is the noncentrality parameter. It is defined in terms of the normalized mean m and the number of degrees of freedom k, as $\lambda = km^2$.

The chi-square density functions for a case of six visible satellites (2 degrees of freedom) are shown in Figure 11.24. These density functions are used to define the detection threshold to satisfy the false alarm and missed detection probabilities. For supplemental navigation, the maximum allowable false alarm rate is one alarm per 15,000 samples or 0.002/h. One sample was considered to be a 2-minute interval based on the correlation time of SA. The maximum false alarm rate for GPS primary means navigation is 0.333×10^{-6} per sample. The minimum detection probability for both supplemental and primary means of navigation is 0.999, or a missed detection rate of 10^{-3} [23].

Figure 11.25 displays a linear no-noise model of the estimated horizontal position error versus the test statistic, forming a characteristic slope line for each visible satellite. These slopes are a function of the linear connection, or geometry matrix, **H**, and vary slowly with time as the satellites move about their orbits. The slope associated with each satellite is given by

$$\text{SLOPE}\ (i) = \sqrt{A_{1i}{}^2 + A_{2i}{}^2}\ /\ \sqrt{S_{ii}},\quad i = 1,2,\ldots,n$$

where

$$\mathbf{A} \equiv \left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}$$

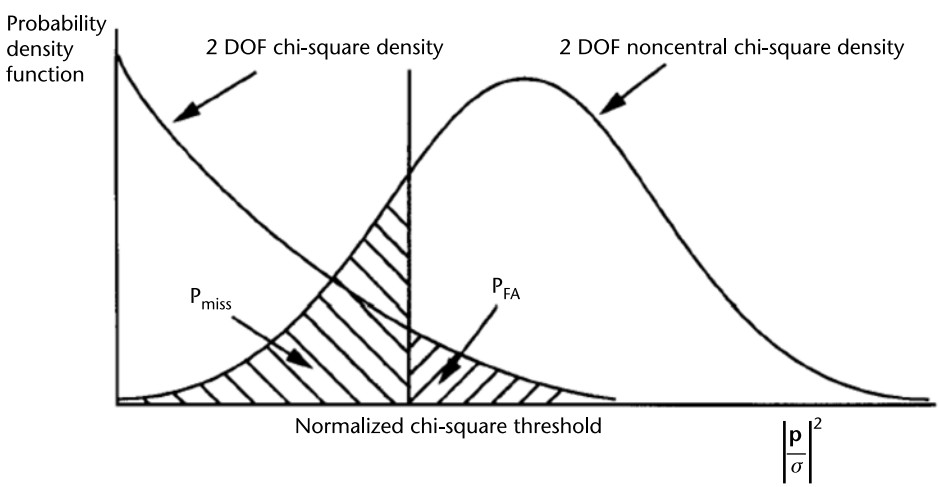


Figure 11.24 Chi-square density functions for 2 degrees of freedom.

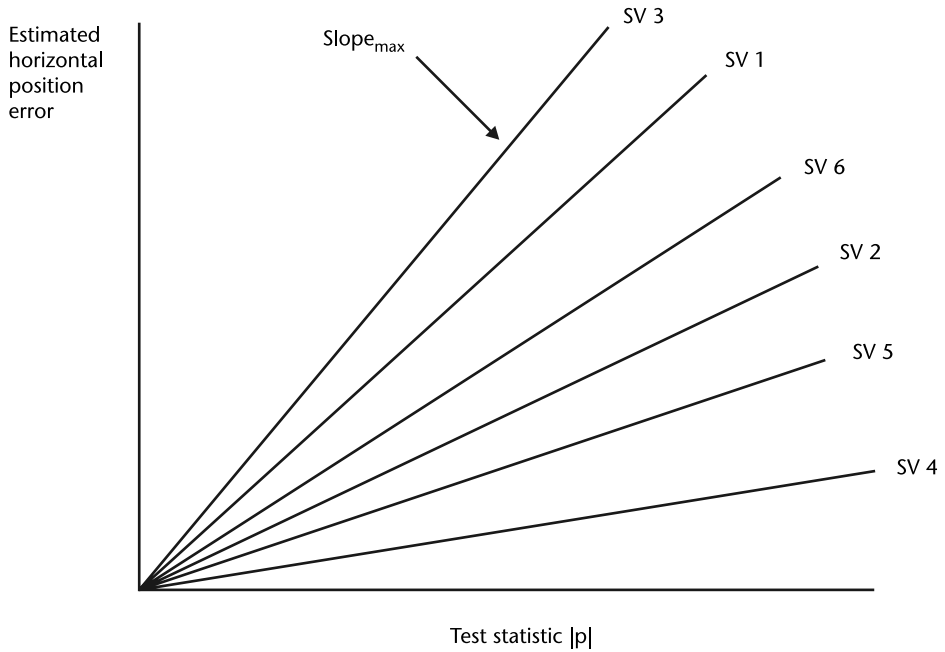


Figure 11.25 Characteristic slopes for six visible satellites.

and **S** was defined previously in (11.44) , but also can be computed directly from **P** as

$$\mathbf{S} = \mathbf{P}^{\mathrm{T}}\mathbf{P}$$

For a given position error, the satellite with the largest slope has the smallest test statistic and will be the most difficult to detect. Therefore, there is a poor coupling between the position error to be protected and the magnitude of the parity vector that can be observed when a bias actually occurs in the satellite with the maximum slope.

The oval-shaped cloud of data shown in Figure 11.26 is a depiction of the scatter that would occur if there were a bias on the satellite with the maximum slope. This bias is such that the fraction of data to the left of the detection threshold is equal to the missed detection rate. Any bias smaller than this value will move the data cloud to the left, increasing the missed detection rate beyond the allowable limit. This critical bias value in parity space is denoted as pbias. The pbias term is completely deterministic, but it is dependent on the number of visible satellites [21]:

$$\text{pbias} = \sigma_{UERE}\sqrt{\lambda}$$

where λ is the noncentrality parameter of the noncentral chi-square density function and σ_{UERE} is the standard deviation of the satellite pseudorange measurement error.

The HPL is determined by

$$\text{HPL} = \text{Slopemax} \times \text{pbias}$$

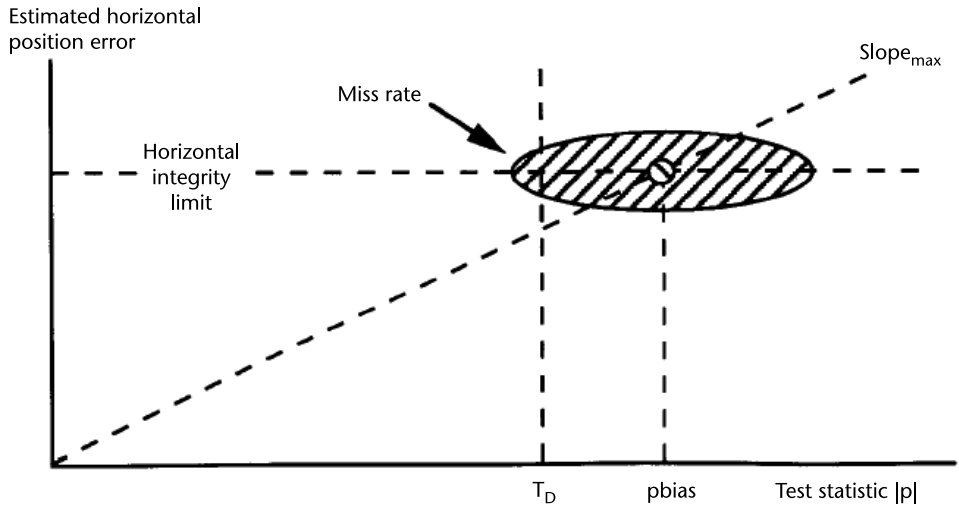


Figure 11.26 Scatter plot with critical bias on Slopemax satellite.

When selective availability (SA) was the dominant error source, other error terms that heavily depend on the elevation angle were negligible. For this reason, pre-2000 RAIM and FDE availability analyses typically assumed a fixed σ_{UERE} value of 33.3m for all satellites, regardless of the satellite elevation angles. After SA was discontinued, errors that depend on the elevation angles make σ_{UERE} values for each satellite significantly different.

Accounting for elevation-dependent errors is accomplished through weighting (or deweighting) of individual satellite range measurements [24]. The only difference between the weighted solution RAIM and the nonweighted solution RAIM is the formula for the maximum horizontal slope, which is shown next.

The threshold and pbias values are the same as with SA on. This is because the maximum false alarm rate is set at 0.333×10^{-6}/sample, which is consistent with the guidance in [10] for SA off.

$$\text{SLOPE}\,(i) = \sqrt{A_{1i}{}^2 + A_{2i}{}^2}\,\sigma_i \,/\, \sqrt{S_{ii}}$$

where

$$\mathbf{A} \equiv \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{W}$$

$$\mathbf{S} \equiv \mathbf{I}_n - \mathbf{H}\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{W}$$

$$\mathbf{W}^{-1} = \begin{bmatrix} \sigma_1^2 & 0 & .. & 0 \\ 0 & \sigma_2^2 & .. & 0 \\ .. & .. & .. & .. \\ 0 & 0 & .. & \sigma_n^2 \end{bmatrix}$$

$$\sigma_{\mathrm{i}}^2 = \sigma_{\mathrm{i,URA}}^2 + \sigma_{\mathrm{i,uire}}^2 + \sigma_{\mathrm{i,tropo}}^2 + \sigma_{\mathrm{i,mp}}^2 + \sigma_{\mathrm{i,rcvr}}^2 \tag{11.46}$$

where the error components are user range accuracy (clock and ephemeris error), user ionospheric range error, tropospheric error, multipath, and receiver noise.

The HPL is formed by the same method as nonweighted RAIM.

$$\text{HPL} = \text{Slopemax} \times \text{normalized pbias} = \text{Slopemax} \times \sqrt{\lambda}$$

Availability of RAIM

Availability of RAIM is determined by comparing the HPL to the maximum alert limit for the intended operation. RAIM was developed and primarily has been used to support aviation applications. Therefore, the focus of the availability analysis in this section will be on aviation applications. The horizontal alert limits for various phases of flight are shown in Table 11.2.

If the HPL is below the alert limit, RAIM is said to be available for that phase of flight. Since the HPL is dependent on the satellite geometry, it must be computed for each location and point in time. Since RAIM requires a minimum of five visible satellites in order to perform fault detection and a minimum of six for fault detection and exclusion, RAIM and FDE will have a lower availability than the navigation function. An analysis of the nominal 24-satellite constellation has been performed to evaluate the availability of RAIM [25–29].

Although a 7.5° mask angle is specified in Federal Aviation Administration (FAA) Technical Standard Order (TSO) C129, a 5° mask angle is specified for FAA TSO C146 receivers and most receivers use a 5° mask angle or lower. A 5° mask angle is applied to this analysis and availability is evaluated over a worldwide grid of points at 5-minute samples over a 24-hour period.

The analysis considers two cases referred to as "SA on" and "SA off." As discussed in Chapter 3, SA was discontinued in May 2000 and new GPS satellites do not even have the capability to generate SA errors. However, at the time of this writing, there were still many certified GPS avionics in operation (e.g., those compliant with TSO C129) that have the SA-on pseudorange error hardcoded into the software. The availability of this type of equipment is poorer than the availability of later equipment. The SA-on results to follow assume that the equipment believes that SA is on (even though it never will be again), and the SA-off results apply to equipment that understands that SA errors are no longer present.

The availability of RAIM fault detection is well above 99% for the en-route and terminal phases of flight and 97.3% for nonprecision approaches. In order to improve availability, the barometric altimeter can be included as an additional

Table 11.2 GNSS Integrity Performance Requirements

Phase of Flight	*Horizontal Alert Limit*
En route	2 nmi
Terminal	1 nmi
NPA	0.3 nmi

Source: [23].

measurement in the RAIM solution. With baro aiding, availability improves to 100% for en-route navigation with 99.99% availability for the terminal phase of flight and 99.9% for nonprecision approach. The maximum outage duration over the course of the day decreases from over half an hour to 15 minutes for nonprecision approach.

The availability of fault detection and exclusion with baro aiding ranges from 81.4% during nonprecision approaches to 98.16% for en-route navigation (FDE without baro aiding is not considered in this analysis due to its low availability). For a nonprecision approach, FDE outages can last for more than 1.5 hours at a location. These results are summarized in Tables 11.3 and 11.4.

The availability of RAIM and FDE with SA off applying a 5° mask angle are shown in Tables 11.5 and 11.6. As shown in Table 11.5, availability of the RAIM fault detection function has nearly 100% availability for equipment that recognizes

Table 11.3 RAIM/FDE Availability with a 5° Mask Angle, SA-On Equipment

RAIM/FDE Function	*En Route*	*Terminal*	*Nonprecision Approach*
Fault detection	99.98%	99.94%	97.26%
Fault detection with baro aiding	100%	99.99%	99.92%
Fault detection and exclusion with baro aiding	99.73%	97.11%	81.40%

Table 11.4 Maximum Duration of RAIM/FDE Outages with 5° Mask Angle, SA On Equipment

RAIM/FDE Function	*En Route*	*Terminal*	*Nonprecision Approach*
Fault detection	5 minutes	10 minutes	35 minutes
Fault detection with baro aiding	0 minutes	5 minutes	15 minutes
Fault detection and exclusion with baro aiding	25 minutes	55 minutes	100 minutes

Table 11.5 RAIM/FDE Availability with a 5° Mask Angle, SA Off Equipment

RAIM/FDE Function	*En Route*	*Terminal*	*Nonprecision Approach*
Fault detection	99.998%	99.990%	99.903%
Fault detection with baro aiding	100%	100%	99.998%
Fault detection and exclusion with baro aiding	99.923%	99.643%	99.100%

Table 11.6 Maximum Duration of RAIM/FDE Outages with 5° Mask Angle, SA Off Equipment

RAIM/FDE Function	*En Route*	*Terminal*	*Nonprecision Approach*
Fault detection	5 minutes	10 minutes	30 minutes
Fault detection with baro aiding	0 minutes	0 minutes	5 minutes
Fault detection and exclusion with baro aiding	10 minutes	35 minutes	60 minutes

that SA is off and that incorporates baro aiding. Recognition that SA is off allows better detection of a bias present on a satellite.

The availability of FDE also improves substantially for equipment that properly recognizes that SA is off such that greater than 99% availability can be achieved for en-route navigation through nonprecision approach. However, the outage duration for nonprecision approach can still be substantial, with outages lasting on the order of an hour.

As shown in Figure 11.27, outages can last up to 60 minutes in several locations, but there is virtually 100% coverage near the equator. This high availability of FDE near the equator is due to the increased number of visible satellites.

Another method for improving availability of RAIM and FDE is to lower the mask angle so that more satellites are visible to the user equipment. However, as mentioned previously, low-elevation satellites will have higher atmospheric errors. These satellites are deweighted in the solution according to (11.46). As demonstrated in Tables 11.7 and 11.8, availability of the fault detection function is very high even without baro aiding. For FDE with baro aiding outages remain, but the number of occurrences and duration are shortened.

Advanced Receiver Autonomous Integrity Monitoring

The advent of two civil frequencies from GPS (L1 and L5) suitable for aviation use combined with the emergence of three other core constellations (GLONASS, Galileo, BeiDou) beyond GPS has created interest in using RAIM for vertical guidance rather than horizontal only. Advanced Receiver Autonomous Integrity Monitoring

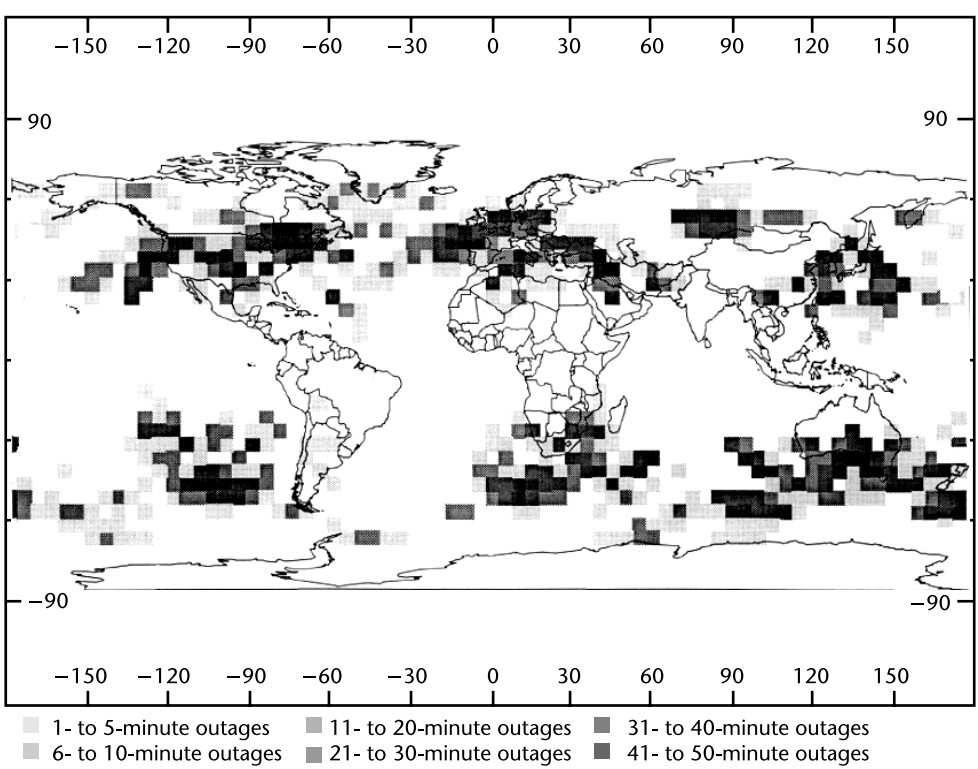


Figure 11.27 FDE availability for NPA with baro aiding with a 5° mask angle.

Table 11.7 RAIM/FDE Availability with 2° Mask Angle, SA Off Equipment

RAIM/FDE Function	*En Route*	*Terminal*	*Nonprecision Approach*
Fault detection	100%	100%	99.988%
Fault detection with baro aiding	100%	100%	100%
Fault detection and exclusion with baro aiding	99.981%	99.904%	99.854%

Table 11.8 Maximum Duration of RAIM/FDE Outages with 2° Mask Angle, SA Off Equipment

RAIM/FDE Function	*En Route*	*Terminal*	*Nonprecision Approach*
Fault detection	0 minutes	0 minutes	5 minutes
Fault detection with baro aiding	0 minutes	0 minutes	0 minutes
Fault detection and exclusion with baro aiding	10 minutes	15 minutes	30 minutes

(ARAIM) is an advanced version of RAIM, which has been known to the aviation community since the late 1980s. While RAIM supports lateral navigation only, ARAIM is intended to support horizontal and vertical guidance, which changes the severity level of misleading information from major (RAIM) to hazardous (ARAIM). The original version of RAIM was based on a set of fixed assertions regarding the nominal performance and fault rates of GPS. In contrast, ARAIM allows a ground system to provide updates regarding the nominal performance and fault rates of the multiplicity of contributing constellations. This integrity data is contained in the Integrity Support Message (ISM) that is developed on the ground and provided to the airborne fleet. The ISM enables this update advantage for evolving constellations without requiring equipment changes [30]. ARAIM is not yet included within any standards, but is anticipated to be included within standards for next-generation civilian GNSS airborne equipment by approximately 2020 [20].

Satellite-Based Augmentation Systems

As discussed in the previous section, one of the limitations of the RAIM and FDE algorithms is that they do not always have enough ranging sources with sufficient geometry to meet availability requirements. Even with the availability improvement obtained with the discontinuance of SA in May 2000 and employing a 2° mask angle outages of up to 30 minutes can occur for the nonprecision approach phase of flight with all 24 satellites operational. Satellites occasionally are taken out of service for maintenance, further reducing the availability of RAIM and FDE.

Therefore, aviation authorities have developed augmentation systems to GPS. One such augmentation is SBAS. The U.S. version of SBAS is known as the Wide Area Augmentation System (WAAS). Other operational SBAS systems are the European Geostationary Overlay Service (EGNOS), the Japanese Multifunction Satellite Augmentation System (MSAS), and the Indian GPS Aided GEO Augmented Navigation (GAGAN) System.

SBAS systems consist of widely dispersed reference stations that monitor and gather data on the GPS satellites. These data are forwarded to the SBAS master

stations for processing to determine the integrity and differential corrections for each monitored satellite. The integrity information and differential corrections are then sent to a ground uplink station and transmitted to a geostationary satellite, along with the geostationary satellite navigation message.

The geostationary satellites downlink the integrity and differential corrections for each monitored satellite using the GPS L1 frequency with a modulation similar to that used by GPS. Therefore, the geostationary satellite also can serve as an additional GPS ranging signal. Based on this information, the user receiver forms horizontal and vertical protection levels based on a weighted solution. The U.S. WAAS system currently utilizes three geostationary satellites at 133°W, 107.3°W, and 98°W. SBAS systems are discussed in much greater detail in Chapter 12. As discussed in Chapter 12, SBAS service providers worldwide are planning to evolve their systems to support additional core constellations beyond GPS.

Ground-Based Augmentation Systems

Ground-based augmentation systems (GBAS) are designed to be specific to an airfield to support precision approach and perhaps terminal area and surface navigation. GBAS systems, such as the version originally referred to as the Local Area Augmentation System (LAAS) developed by the FAA, utilize multiple GPS reference receivers. Data from the reference receivers are processed using an averaging technique to determine integrity and develop differential corrections.

The GBAS integrity algorithm involves placing an upper confidence bound on the lateral and vertical position error by computing lateral and vertical protection levels (LPL and VPL) using an assumed fault hypothesis. There are two fault hypotheses for GBAS: H_0 and H_1. The H_0 hypothesis refers to normal measurement conditions (i.e., no faults) in all reference receivers and for all satellites. The H_1 hypothesis represents a latent fault associated with one reference receiver. A latent fault includes any erroneous measurement(s) that are not immediately detected by the ground subsystem, such that the broadcast data are affected and there is an induced position error in the airborne subsystem. The differential corrections and integrity parameters for each monitored satellite are broadcast to the aircraft via a VHF datalink. GBAS systems are discussed in detail in Chapter 12.

11.5 Continuity

Continuity, as defined in [9], is "…the probability that the specified system performance will be maintained for the duration of a phase of operation, presuming that the system was available at the beginning of that phase of operation." The level of continuity provided by GNSS thus varies with the specific performance requirements for any given application. For example, the level of continuity of GNSS for a low-accuracy time-transfer application will be much higher than the level of GNSS continuity for an aircraft non precision approach. The former application only requires a single visible GNSS satellite, whereas the latter requires at least 5 visible satellites with good geometry to support RAIM.

11.5.1 GPS

Some useful information regarding the continuity of the GPS C/A-code signal-in-space is provided in the GPS SPS Performance Standard [10]. This reference assures a greater than or equal to 0.9998 probability over any hour of not losing the GPS C/A-code from a slot in the baseline 24-satellite constellation due to an unscheduled interruption. This continuity standard is based upon an average over all slots in the 24-slot constellation, normalized annually, and assumes that the signal is available from the slot at the beginning of the hour. Scheduled interruptions (e.g., due to maintenance) that are announced at least 48 hours in advance publicly are not considered within [10] to constitute a loss of continuity.

An earlier edition of [10] provided data regarding continuity of GPS based upon observed performance from January 1994 to July 2000. During this timespan on average, each in-orbit GPS satellite ceased functioning 2.7 times per year and was out-of-service for a total downtime of 58 hours. The majority of these instances (referred to as *downing events*) were related to scheduled maintenance, accounting for 1.9 downing events per year and an average total downtime of 18.7 hours. The remaining 0.9 downing event per year per satellite were unscheduled and accounted for a total average downtime of 39.3 hours. (Note that the component values provided in the third edition of [10] of 0.9 and 1.9 do not add to the total of 2.7, also in the third edition. This is presumably due to rounding errors.) Causes of unscheduled outages include failures of one or more satellite subsystems that resulted in a loss of service. For many applications, only unscheduled downing events are of concern. Scheduled maintenance activities are generally announced well in advance and can often be planned around. For such applications, the probability that any given GPS satellite will fail over a 1-hour time interval is approximately 0.0001. This value is computed by dividing the average of 0.9 unscheduled downing event per year by the number of hours in a year, 8,760.

The GPS PPS Performance Standard [31] essentially provides the same assurance for L1/L2 P(Y)-code continuity as in [10] for the C/A-code described above. Performance Standards for the modernized GPS civilian and military signals are not yet available.

11.5.2 GLONASS

Performance standards for GLONASS are now being developed. A GLONASS signal-in-space continuity level of 0.9995 per hour is proposed in [32], for similar conditions as the GPS C/A-code 0.9998 continuity standard in [10] (discussed in Section 11.5.1).

11.5.3 Galileo

Established continuity requirements for Galileo are formulated differently than continuity requirements for other three core constellations. Galileo continuity requirements from [33] establish a maximum probability per time interval that a defined service (with associated levels of accuracy and integrity) is lost. Continuity risk requirements for the Galileo Safety-of-Life Service (SoL) and Public Regulated Service (PRS) are included in [33]. For both SoL (critical level) and the PRS, the continuity

risk requirement is 10^{-5}/15 seconds. These services are described in Chapter 5. As discussed within that chapter, the SoL service is currently being reprofiled.

11.5.4 BeiDou

The BeiDou Navigation Satellite System (BDS) Open Service (OS) Performance Standard [34] defines signal-in-space (SIS) continuity as "the probability that a healthy BDS OS SIS will continue working without unscheduled interruptions over a specified time interval." It provides standards for SIS continuity levels of ≥ 0.995/hour for the BeiDou GEO and IGSO satellites, and ≥ 0.994/hour for the BeiDou MEO satellites. (See Chapter 6 for an overview of BeiDou.)

References

[1] Rodríguez, J. A. A., et al., "Combined Galileo/GPS Frequency and Signal Performance Analysis," *Proc. of ION GNSS 2004*, Long Beach, CA, September 2004, pp. 632–649.

[2] Nelson, W., *Use of Circular Error Probability in Target Detection*, United States Air Force, ESD-TR-88-109, Hanscom Air Force Base, Bedford, Massachusetts, May 1988.

[3] Special Committee 159, *Minimum Operational Performance Standards for Global Positioning System/Wide Area Augmentation System Airborne Equipment*, DO-229D with Change 1, RTCA, Inc., Washington, D.C., December 2013.

[4] van Graas, F., and A. Soloviev, "Precise Velocity Estimation Using a Stand-Alone GPS Receiver," *NAVIGATION: Journal of The Institute of Navigation*, Winter 2004-2005.

[5] Moudrak, A., et al., 2004, "GPS Galileo Time Offset: How It Affects Positioning Accuracy and How to Cope with It," *Proc. of The Institute of Navigation ION GNSS 2004*, Long Beach, CA, September 2004.

[6] Hegarty, C., "Panel Discussion on GNSS Interoperability," *Proc. of the 36th Precise Time and Time Interval Meeting*, Washington, D.C., December 2004. http://tycho.usno.navy.mil/ptti/2004papers/panel.pdf.

[7] Working Group C, *Combined Performances for Open GPS/Galileo Receivers*, United States – European Union Working Group C established for Cooperation on Satellite Navigation, July 2010. http://www.gps.gov/policy/cooperation/europe/2010/working-group-c/combined-open-GPS-Galileo.pdf.

[8] Kouba, J., and P. Héroux, "GPS Precise Point Positioning Using IGS Orbit Products," *GPS Solutions*, Vol. 5, No. 2, Fall 2000, pp. 12–28.

[9] DOD/DOT/DHS, 2014 *Federal Radionavigation Plan*, DOT-VNTSC-OST-R-15-01, U.S. Departments of Transportation, Defense, and Homeland Security, Washington, D.C., May 2015.

[10] U.S. Department of Defense, *Global Positioning System Standard Positioning Service Performance Standard*, 4th ed., Washington, D.C., September 2008.

[11] Sotolongo, G. L., "Proposed Analysis Requirements for the Statistical Characterization of the Performance of the GPSSU RAIM Algorithm for Appendix A of the MOPS," RTCA 308-94/SC159-544, July 20, 1994.

[12] Shank, C., and J. Lavrakas, "GPS Integrity: An MCS Perspective," *Proc. of ION GPS-93, Sixth International Technical Meeting of the Satellite Division of the Institute of Navigation*, Salt Lake City, UT, September 22–24, 1993, pp. 465–474.

[13] Walter, T., and Blanch, J., "Characterization of GNSS Clock and Ephemeris Errors to Support ARAIM," *Proc. of the ION 2015 Pacific PNT Meeting*, Honolulu, HI, April 2015.

[14] RTCA, "Aberration Characterization Sheet (ACS)," RTCA Paper No. 034-01/SC-159-867, July 1998.

[15] Heng, L., et al., "GLONASS Signal-in-Space Anomalies Since 2009," *Proc. of the 25th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2012)*, Nashville, TN, September 2012, pp. 833–842.

[16] Heng, L., "Safe Satellite Navigation with Multiple Constellations: Global Monitoring of GPS and GLONASS Signal-in-Space Anomalies," Ph.D. Dissertation, Stanford University, Stanford, CA, December 2012.

[17] Beutler, G., et al., "The System: GLONASS in April, What Went Wrong," *GPS World*, June 2014.

[18] Kovach, K., et al., "GPS Receiver Impact from the UTC Offset (UTCO) Anomaly of 25-26 January 2016," *Proc. of ION-GNSS 2016*, Portland, OR, September 2016.

[19] Van Dyke, K., et al., "GPS Integrity Failure Modes and Effects Analysis (IFMEA)," *Proc. of The Institute of Navigation National Technical Meeting*, Anaheim, CA, January 2003.

[20] Hegarty, C., et al., "RTCA SC-159: 30 Years of Aviation GPS Standards," *Proceedings of ION GNSS+ 2015*, Tampa, FL, September 2015, pp. 877–896.

[21] Brown, R. G., "GPS RAIM: Calculation of Thresholds and Protection Radius Using Chi-Square Methods-A Geometric Approach," *ION Red Book Series, Volume 5, Global Positioning System*, Papers Published in *NAVIGATION*, 1998.

[22] van Graas, F., and P. A. Kline, *Hybrid GPS/LORAN-C*, Ohio University technical memorandum OU/AEC923TM00006/46+46A-1, Athens, OH, July 1992.

[23] RTCA, *Minimum Operational Performance Standards for Airborne Supplemental Navigation Equipment Using Global Positioning System (GPS)*, Document No. RTCA/DO-208, prepared by SC-159, July 1991.

[24] RTCA SC-159 Response to the JHU/APL Recommendation Regarding Receiver Autonomous Integrity Monitoring, July 19, 2002.

[25] Van Dyke, K. L., "Analysis of Worldwide RAIM Availability for Supplemental GPS Navigation," DOT-VNTSC-FA360-PM-93-4, May 1993.

[26] Brown, R. G., et al., "ARP Fault Detection and Isolation: Method and Results," DOT-VNTSC-FA460-PM-93-21, December 1993.

[27] Van Dyke, K. L., "RAIM Availability for Supplemental GPS Navigation," *NAVIGATION, Journal of the Institute of Navigation*, Vol. 39 No. 4, Winter 1992–93, pp. 429–443.

[28] Van Dyke, K. L., "Fault Detection and Exclusion Performance Using GPS and GLONASS," *Proc. of the ION National Technical Meeting*, Anaheim, CA, January 18–20, 1995, pp. 241–250.

[29] Van Dyke, K. L., "World After SA: Improvements in RAIM/FDE Availability," *IEEE PLANS Symposium*, April 2000.

[30] GPS-Galileo Working Group C, *ARAIM Technical Subgroup Milestone 2 Report,* February 11, 2015. http://www.gps.gov/policy/cooperation/europe/2015/working-group-c/.

[31] U.S. Department of Defense, *Global Positioning System Precise Positioning Service Performance Standard*, Washington, D.C.: Department of Defense, February 2007.

[32] Bolkunov, A., "GLONASS Open Service Performance Parameters Standard and GNSS Open Service Performance Parameters Template Status," *International Committee on GNSS*, Prague, Czech Republic, November 2014. http://www.unoosa.org/pdf/icg/2014/wg/wga4.2.pdf.

[33] European Commission and European Space Agency, *Galileo Mission High Level Definition*, version 3, September 2002.

[34] China Satellite Navigation Office, *BeiDou Navigation Satellite System Open Service Performance Standard*, version 1.0, Beijing, December 2013.

CHAPTER 12

Differential GNSS and Precise Point Positioning

S. Bisnath, M. Uijt de Haag, D. W. Diggle, C. Hegarty, D. Milbert, and T. Walter

12.1 Introduction

As discussed in Chapter 11, a dual-frequency or multifrequency, multiconstellation GNSS user can often attain better than 1 to 2m, 95% positioning and 1 to 2 ns, 95% timing accuracy worldwide with high levels of integrity, availability, and continuity. However, there are many applications that demand yet higher levels of accuracy, integrity, availability, and continuity. For such applications, augmentation is required. There are several classes of augmentation, which can be used singly or in combination: DGNSS[1], precise point positioning (PPP), and the use of external sensors. This chapter 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. The supplied information may include:

- Corrections to the raw end-user's pseudorange or carrier phase measurements, corrections to GNSS satellite-provided clock and ephemeris data, or data to replace the broadcast clock and ephemeris information, atmospheric corrections, and so forth.

1. Regarding terminology, since GPS was the first operational GNSS constellation, followed shortly thereafter by GLONASS, there is a wealth of literature on differential GPS (DGPS) techniques and to a lesser extent differential GLONASS (DGLONASS) techniques. Although numerous operational DGNSS systems today only provide data applicable to GPS, the techniques employed may be applied to all GNSS constellations and we will emphasize this point through the use of the term DGNSS even for these systems.

- Raw reference station measurements (e.g., pseudorange and carrier phase).
- Integrity data, for example, "use" or "don't use" indications for each visible satellite, or statistical indicators of the accuracy of provided corrections.
- Auxiliary data including the location, health, and meteorological data of the reference station(s).

The reference station(s) data may be supplied in real-time to the end user using any one of a variety of data links, for example, radio links at frequencies ranging from low frequencies (LF) below 300 kHz to L-band (1,000–2,000 MHz) and beyond, the Internet, and importantly, the link may not be real time. For instance, it is possible to implement DGNSS methods using two GNSS receivers that each simply log data to a hard drive or other storage device.

All DGNSS and, to some extent, PPP systems work by exploiting the spatial and time correlation characteristics of GNSS errors (see Chapter 10). Because many GNSS error sources are highly correlated spatially and temporally, if these errors can be measured using one or more reference stations at known locations, this information provided in a sufficiently timely manner can greatly benefit the end user. Since correlation of GNSS errors is generally higher for shorter distances, accuracy is generally improved in DGNSS when the reference stations are closer to the end user.

DGNSS systems provide corrections to either the raw measurements the end user makes or to the broadcast navigation data for each visible GNSS satellite. PPP systems can be similar in architecture but supply data to replace, rather than correct, that provided in the GNSS signals' broadcast navigation data.

DGNSS techniques may be categorized in different ways: as *absolute* or *relative* differential positioning; as *local-area*, *regional-area*, or *wide-area*; and as *code-based* or *carrier-based*.

Absolute differential positioning is the determination of the user's position with respect to an Earth-centered, Earth-fixed (ECEF) coordinate system (see Section 2.2.2). This is the most common goal of DGNSS. For absolute differential positioning, the reference station(s) must be accurately known with respect to the same ECEF coordinate system that the user position is desired. Aircraft use this type of positioning as an aid for remaining within certain bounds of the desired flight path; ships use it as an aid for remaining within a harbor channel.

Relative differential positioning is the determination of the user's position with respect to a coordinate system attached to the reference station(s), whose absolute ECEF position(s) may not be perfectly known. For instance, if DGNSS is implemented to land aircraft on an aircraft carrier, the ECEF positions of the reference stations may be imperfectly known and time-varying. In this case, only the position of the aircraft with respect to the carrier is required.

DGNSS systems may also be categorized in terms of the geographic area that is to be served. The simplest DGNSS systems are designed to function only over a very small geographic area (e.g., with the user typically separated by less than 10–200 km from a single reference station). The separation between the user and the reference station is referred to as the *baseline*, which may be interpreted as a vector. The terms *short baseline*, *medium baseline*, and *long baseline* are frequently encountered, but unfortunately there are no universally agreed-upon definitions.

The most common usage is for short baselines to span approximately 0–20 km, medium baselines to span from 20–100 km, and long baselines to span greater than 100 km. To effectively cover larger geographic regions, typically multiple reference stations and different algorithms are employed. The terms regional-area and wide-area are frequently used in the literature to describe DGNSS systems covering larger geographic regions with regional-area systems generally covering areas up to around 1,000 km and wide-area systems covering yet larger regions such as a continent. However, there are not universally agreed-upon demarcations in terms of distance for the applicability of each term.

One final categorization of DGNSS systems is between code-based or carrier-based techniques. Code-based DGNSS systems rely primarily on GNSS code (i.e., pseudorange) measurements, whereas carrier-based DGNSS systems ultimately rely primarily on carrier-phase measurements[2]. As discussed in Chapter 8, carrier-phase measurements are much more precise than pseudorange measurements, but contain unknown integer wavelength components that must be resolved. Code-based differential systems can provide decimeter-level position accuracies, whereas state-of-the-art carrier-based systems can provide millimeter-level performance.

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. Section 12.6 details a number of operational and planned DGNSS and PPP systems.

12.2 Code-Based DGNSS

Many code-based DGNSS techniques have been proposed to provide improvements in performance over stand-alone GNSS. These techniques vary in sophistication and complexity from a single reference station that calculates the errors at its position for use with nearby GNSS receivers to worldwide networks that provide data for estimating errors from detailed error models at any position on or near the Earth's surface. As discussed in Section 12.1, they may be sorted into three categories, local area, regional area, and wide area, depending on the geographic area that they are intended to serve. This section discusses code-based techniques for each of these categories.

12.2.1 Local-Area DGNSS

A local-area DGNSS (LADGNSS) system improves on the accuracy of stand-alone GNSS by estimating errors corrupting the stand-alone GNSS position solution and transmitting these estimates to nearby users.

2. It should be noted that virtually all DGNSS systems employ both pseudorange and carrier-phase measurements, so the distinction between code-based and carrier-based techniques is a matter of degree of reliance on the respective measurement type. Most DGNSS systems that are referred to as carrier-based resolve integer ambiguities in either the end user's raw carrier-phase measurements or more commonly within the differences of these measurements and the reference station(s) measurements.

12.2.1.1 Position Domain Corrections

Conceptually, the simplest way to implement LADGNSS is to place a single GNSS reference receiver at a surveyed location, compute the coordinate differences (in latitude, longitude, and geodetic height) between that surveyed position and the position estimate derived from GNSS measurements, and transmit these latitude, longitude, and height differences to nearby users. For the most part, the coordinate differences represent the common errors in the reference and user receiver GNSS position solutions at the measurement time. The user receivers can use these coordinate differences to correct their own GNSS position solutions.

Although extremely simple, this technique has a number of significant deficiencies. First, it requires that all receivers make pseudorange measurements to the same set of satellites to ensure that common errors are experienced. Therefore, the user receivers must coordinate their choice of satellites with the reference station or the reference station may determine and transmit position corrections for all combinations of visible satellites. When eight or more satellites are visible, the number of combinations becomes impractically large (70 or more combinations of four satellites). A second problem may also arise if the user and reference station receivers employ different position solution techniques. Unless both receivers employ the same technique (e.g., least-squares, weighted least-squares, or Kalman filters, with equivalent smoothing time-constants, filter tunings, and so forth), position domain corrections may yield erratic results. For these reasons, position domain corrections are seldom if ever employed in operational DGNSS systems.

12.2.1.2 Pseudorange Domain Corrections

In most operational code-based, local-area DGNSS systems (see Figure 12.1), instead of determining position coordinate errors, the reference station determines and disseminates pseudorange corrections for each visible satellite. If the reference station is sufficiently close to the user, the errors in the reference station's pseudorange measurements for visible satellites are expected to be very similar to those experienced by the user. If the reference station estimates the errors by leveraging its known surveyed position and provides this information in the form of corrections to the user, it is expected that the user's position accuracy will be improved as a result. Accuracy is dependent upon both distance to the reference station and latency of the supplied corrections. As discussed in Chapter 10, for many GNSS error sources spatial correlation monotonically decreases with distance. Thus, DGNSS accuracy is generally better for short baselines than for medium or long. Time correlations (i.e., how rapidly the errors change with time) are also of interest, because in general, DGNSS systems cannot instantaneously provide data to the end user; even with a high-speed radio link, there is some finite delay associated with the generation, transmission, reception, and application of the data.

The local-area DGNSS concept is explained in detail in the following mathematical treatment. In order for the user receiver to determine its position accurately with respect to the Earth (i.e., for absolute DGNSS applications), the reference station must have accurate knowledge of its own position in ECEF coordinates. Given that the reported position of the ith satellite is (x_i, y_i, z_i) and the position of

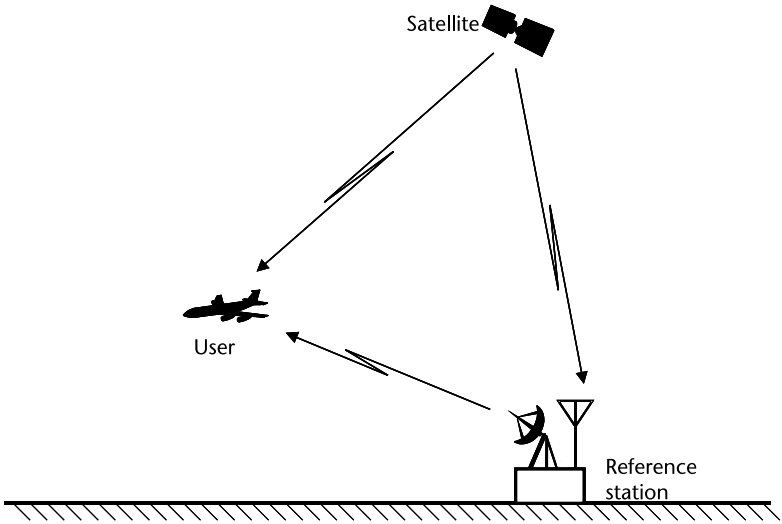


Figure 12.1 Local-area DGNSS concept.

the reference station is known through a survey to be at position (x_m, y_m, z_m), the computed geometric distance, R_m^i, from the reference station to the satellite is

$$R_m^i = \sqrt{\left(x_i - x_m\right)^2 + \left(y_i - y_m\right)^2 + \left(z_i - z_m\right)^2} \tag{12.1}$$

The reference station then makes a pseudorange measurement, ρ_m^i, to the ith satellite. This measurement contains the range to the satellite along with the errors discussed in Chapter 10.

$$\rho_m^i = R_m^i + c\,\delta t_m + \varepsilon_m \tag{12.2}$$

where ε_m are the pseudorange errors and $c\delta t_m$ represents the reference station clock offset from a convenient common timescale (e.g., GPS or other GNSS System Time).

The reference station differences the computed geometric range, R_m^i, with the pseudorange measurement to form the differential correction

$$\Delta\rho_m^i = R_m^i - \rho_m^i = -c\,\delta t_m - \varepsilon_m \tag{12.3}$$

This correction, which may be a positive or negative quantity, is broadcast to the user receiver where it is added to the user receiver's pseudorange measurement to the same satellite

$$\begin{aligned}\rho_u^i + \Delta\rho_m^i &= R_u^i + c\,\delta t_u + \varepsilon_u \\ &\quad + \left(-c\,\delta t_m - \varepsilon_m\right)\end{aligned} \tag{12.4}$$

To a significant extent, the user receiver's pseudorange error components will be common to those experienced by the reference station with the exception of multipath and receiver noise. The corrected pseudorange can be expressed as

$$\rho^i_{u,cor} = R^i_u + \varepsilon_{um} + c\,\delta t_{um} \tag{12.5}$$

where $\varepsilon_{um} = \varepsilon_u - \varepsilon_m$ represents residual pseudorange errors and δt_{um} is the difference in user and reference station clock offsets, $\delta t_u - \delta t_m$.

In Cartesian coordinates, (12.5) becomes

$$\rho^i_{u,cor} = \sqrt{(x_i - x_u)^2 + (y_i - y_u)^2 + (z_i - z_u)^2} + \varepsilon_{um} + c\,\delta t_{um} \tag{12.6}$$

By making pseudorange measurements to four or more satellites, the user receiver can compute its position by using one of the position determination techniques discussed in Chapters 2 and 11. Since the residual pseudorange error, ε_{um}, is generally smaller statistically than the error of the uncorrected pseudorange, a more accurate position solution is generally attained.

Importantly, when pseudorange corrections are applied, the clock offset produced by the position solution is the difference between the user's clock error and the reference station clock error. For applications where the user requires accurate time, the reference station clock offset may be estimated using the standard position solution technique and removed from the pseudorange corrections. Removal of the reference station clock offset is generally desirable, even when the user does not require accurate time, since a large reference station clock bias could result in excessively large pseudorange corrections (e.g., to fit within a fixed-size data field in a digital message).

Because pseudorange errors vary with time, as discussed in Chapter 10, the transmitted pseudorange correction,

$$\Delta\rho^i_m(t_m) = \left[R^i_m(t_m) - \rho^i_m(t_m)\right] \tag{12.7}$$

which is an estimate of the pseudorange error with the sign inverted and is most accurate at the instant of time t_m, for which the correction was calculated. To enable the user receiver to compensate for pseudorange error rate, the station may also transmit a pseudorange rate correction, $\Delta\dot{\rho}^i_m(t_m)$. The user receiver then adjusts the pseudorange correction to correspond to the time of its own pseudorange measurement, t, as follows:

$$\Delta\rho^i_m(t) = \Delta\rho^i_m(t_m) + \Delta\dot{\rho}^i_m(t_m)(t - t_m) \tag{12.8}$$

The corrected user receiver pseudorange, $\rho^i_{cor}(t)$, for time t is then calculated from

$$\rho^i_{u,corr}(t) = \rho^i(t) + \Delta\rho^i_m(t) \tag{12.9}$$

12.2.1.3 Performance of Code-Based LADGNSS

Using the information presented in Chapter 10 on the spatial and time correlation characteristics of GNSS errors, Table 12.1 presents an error budget for a LADGNSS system in which the reference station and the user rely only on single-frequency pseudorange measurements that are assumed to be made to one or more generic GNSS constellations that provide 1-sigma signal-in-space errors at the 1-m level (using the representative clock and ephemeris error values from Table 10.3). The values in the table assume that latency errors are negligible (e.g., that the pseudorange corrections are transmitted over a high-speed data link). It is also assumed that the reference station and user are either at the same altitude or that a tropospheric height difference correction is employed. Note that multipath is the dominant error component over short baselines. For longer baselines, the residual ionospheric or tropospheric errors may dominate. Over very long baselines, performance may be improved by applying a local tropospheric error model at both the reference station and user locations, rather than the conventional short-baseline design in which neither side applies a model.

12.2.2 Regional-Area DGNSS

To extend the region over which LADGNSS corrections can be used without the decorrelation of errors that accompanies the separation of the user from the station, three or more reference stations may be distributed along the perimeter of the region of coverage in a concept referred to as regional-area DGNSS. The user receiver can then obtain a more accurate position solution by employing a weighted average of pseudorange corrections from the stations. Because the error in the broadcast corrections grows with distance from each station, the weights may be determined by geometric considerations alone to give the largest weight to the closest station, such as by choosing those weights that describe the user position as the weighted sum of the station positions [1]. For example, with three stations at locations denoted by latitude ϕ and longitude λ, the three weights, w_1, w_2, and w_3, of stations

Table 12.1 Pseudorange Error Budget with and without LADGNSS Corrections

Segment Source	*Error Source*	*1σ Error (m)*	
		GNSS-only	*with LADGNSS*
Space/control	Broadcast clock	0.4	0.0
	Broadcast ephemeris	0.3	0.1–0.6 mm/km × baseline in kilometers
User	Ionospheric delay	7.0	0.2–4 cm/km × baseline in kilometers
	Tropospheric delay	0.2	1–4 cm/km × baseline in kilometers
	Receiver noise and resolution	0.1	0.1
	Multipath	0.2	0.3
System UERE	Total (rss)	7.0	0.3m + 1–6 cm/km × baseline in kilometers

$M_1(\phi_1, \lambda_1)$, $M_2(\phi_2, \lambda_2)$, and $M_3(\phi_3, \lambda_3)$, for user $U(\phi, \lambda)$ may be determined by the following set of three equations (Figure 12.2):

$$\phi = w_1 \phi_1 + w_2 \phi_2 + w_3 \phi_3$$
$$\lambda = w_1 \lambda_1 + w_2 \lambda_2 + w_3 \lambda_3$$
$$(w_1 + w_2 + w_3) = 1$$

A two-step approach to using multiple reference stations to improve the accuracy of the user's position estimate is described in [1]. In the first step, the pseudorange corrections from each reference station are used to determine the position of the user individually. The second step entails computing a weighted average of the individual position estimates to provide a more accurate estimate. Each weight is formed from the inverse of the product of the distance of the reference station from the user and the standard deviation from the average of the estimates from that station, normalized by the sum of the weights. The error introduced by each reference station receiver is thus diluted by its weight, so that if, for example, the weights were all equal, then each reference station receiver error would be diluted by a factor of $1/n$. However, since the errors are uncorrelated, the standard deviation of their sum is $1/\sqrt{n}$; thus, the standard deviation of the total error due to the reference stations is decreased by a factor of $\sqrt{n}$ from that of one reference station.

12.2.3 Wide-Area DGNSS

Wide-area DGNSS (WADGNSS) attempts to attain submeter-level accuracy over a large region while using a fraction of the number of reference stations that

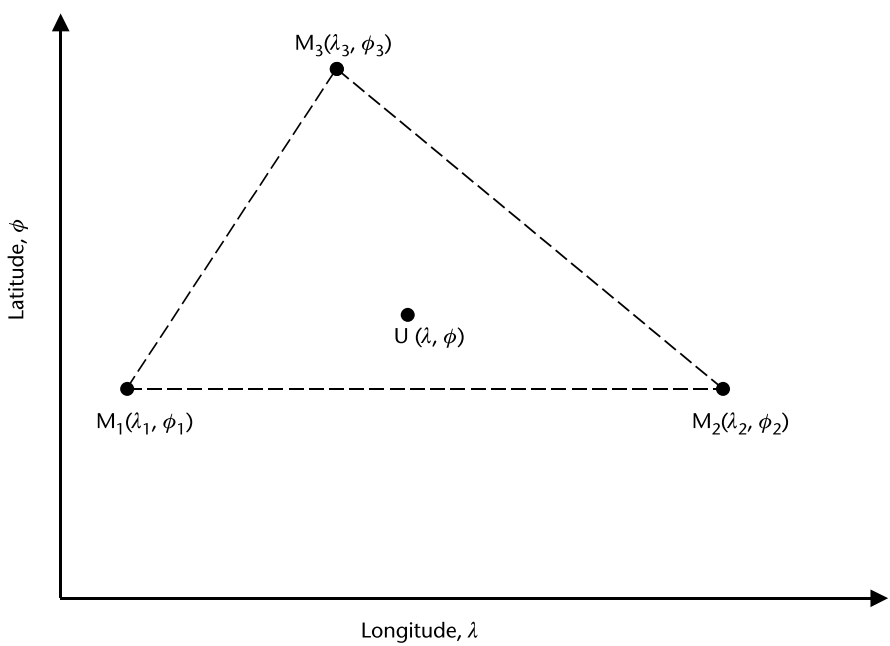


Figure 12.2 Calculating the correction weights.

LADGNSS would require to attain the same accuracy within the same coverage region. The general approach (see, e.g., [2–4])—in contrast to that of LADGNSS—is to break out the total pseudorange error into its components and to estimate the variation of each component over the entire region, rather than just at the station positions. The accuracy then does not depend on the proximity of the user to a single reference station.

The WADGNSS concept, illustrated in Figure 12.3, includes a network of reference stations, one or more central processing sites, and a data link to provide corrections to users. Each reference station includes one or more GNSS receivers that measure pseudorange and carrier-phase for the broadcast signals from all visible satellites. This data is provided to the central processing site(s), which process the raw data to develop estimates of the broadcast ephemeris and broadcast clock errors for each satellite. WADGNSS systems intended to provide service for single-frequency users also estimate ionospheric errors throughout the service volume. Tropospheric delays are typically addressed through the use of models employed by the reference stations and by the user.

12.2.3.1 Satellite Ephemeris and Clock Errors

Using pseudorange and carrier-phase data from the entire network of reference stations, each central processing site can develop precise estimates of the true locations and clock errors of the GNSS satellites that are visible to the network. For each

Figure 12.3 WADGNSS concept.

satellite, the three-dimensional position error (e.g., in an ECEF coordinate system) between the WADGNSS position estimate and the broadcast position is provided to the user. The user then maps this satellite position correction into a pseudorange correction by projecting the position error onto the line-of-sight direction to the satellite. A separate clock correction is also broadcast to the user that can be directly applied as an additional pseudorange correction.

The central processing site can estimate the true GNSS satellite positions and clocks by reversing the basic GNSS algorithm. Here four or more widely separated ground stations whose positions are accurately known each calculate the pseudorange to a given satellite, after estimating and removing the atmospheric delays [2, 3]. Synchronization of the reference station clocks are required, which may be accomplished using GNSS. In practice, extremely accurate position and clock estimates can be achieved by combining the concept of a reverse- GNSS solution with sophisticated models to describe the motion of the GNSS satellites over time. Such modeling is a standard method used for orbit determination for many satellite systems, including the ground networks for each GNSS constellation (see, e.g., Section 3.3.1.4). An excellent introduction to the methods of satellite orbit determination may be found in [5].

12.2.3.2 Determining Ionospheric Propagation Delays

Ionospheric delays can be addressed in various ways within a WADGNSS system. The simplest approach is for the user to directly measure ionospheric delays using a dual-frequency or multifrequency receiver. This option requires the use of a second frequency, which is becoming increasingly available. Historically, GPS was the first operational GNSS constellation but lacked a second civilian frequency until the first L2C-capable satellite was launched in 2005. Decades earlier, codeless and semicodeless methods were developed within civilian equipment to track the encrypted GPS L2 P(Y)-code signal (see Section 8.7.4), but such methods were fragile. Because of this history, some important operational WADGPS systems discussed in Section 12.6 are designed to support users with single-frequency L1 C/A code receivers. These systems estimate ionospheric delays throughout their service volumes using dual-frequency receivers in their reference stations. The slant ionospheric delays measured by the reference stations are used by the central processing site, along with models of the ionosphere, to develop estimates of vertical ionospheric delays for discrete latitude/longitude points across the coverage volume. These vertical delay estimates are broadcast to the user. The user equipment then interpolates among these points to develop a vertical ionospheric delay correction each visible GNSS signal. The vertical delay correction is mapped into an appropriate slant delay correction based upon the elevation angle for each visible satellite. The vertical delay corrections for the visible satellites are generally not the same, since the points of intersection between the signal paths and the ionosphere are not collocated.

12.3 Carrier-Based DGNSS

The constant motion of many GNSS satellites and additional possible motion of the user requires that a GNSS receiver, in general, be capable of accounting for the

changing Doppler frequency shift on each tracked frequency. The shift in frequency arises due to the relative motion between the satellites and the receiver(s). For example, typical GNSS satellite motion in medium Earth orbit with respect to an Earth-fixed observer can result in a maximum range of Doppler frequencies of approximately ±4,000 Hz with respect to the L-band carrier frequencies. Integration of the Doppler frequency offset within phase tracking loops results in an extremely accurate measurement of the advance in signal carrier phase between time epochs (see Section 8.6). Interferometric techniques can take advantage of these precise phase measurements and, assuming sources of error can be mitigated, real-time positional accuracies in the centimeter range are achievable. While changes in signal phase from epoch to epoch can be measured with extreme accuracy, the number of whole carrier cycles along the propagation path from satellite to receiver remains ambiguous. Determining the number of whole carrier cycles in the propagation path is known as carrier-cycle integer ambiguity resolution and remains an active area of investigation in the field of kinematic DGNSS research. Integer ambiguity resolution was first studied for GPS. Remondi [6] made extensive use of the ambiguity function for resolving these unknown integer wavelength multiples, but the pioneering work in this area arose from the efforts of Counselman and Gourevitch [7] and Greenspan et al. [8]. A number of ambiguity resolution (AR) techniques have since been developed such as the least-squares ambiguity adjustment (LAMBDA) [9], allowing for on-the-fly (OTF) resolution of phase measurements in static and kinematics situations. Coupled with real-time communications links between reference and remote receivers, *real-time kinematic* (RTK) has been come the industry standard for few centimeter-level positioning. And, in a regional context, making use of regional corrections, network RTK was developed.

Advantage can be taken by combining the multiple frequencies to speed the ambiguity resolution process, and this approach has been the subject of a number of articles in the literature (e.g., Hatch [10] for GPS). These dual-frequency receiver measurements can be combined to produce the sum and difference of frequencies. Using the difference wavelength (known as the wide-lane) makes the integer ambiguity search more efficient. A change of one wide-lane wavelength results in virtually a fourfold increase in distance over that of one wavelength at either the GPS L1 and L2 frequencies alone. Obviously, the search for the proper combination of integer ambiguities progresses more quickly using wide-lane observables, but the requirements on the receiver for simultaneous dual-frequency tracking—here, the P(Y) code is generally used—are more stringent. In particular, the noise factor for the wide-lane processing goes up by a factor of nearly 6 [11]. These matters aside, wide-lane techniques offer great advantage for obtaining rapid RTK integer ambiguity resolution, and the methodology will be presented later in this chapter.

12.3.1 Precise Baseline Determination in Real Time

Determination of the carrier-cycle ambiguities on-the-fly is key to any application where precise positioning at the centimeter level, in real time, is required. Such techniques have been successfully applied to aircraft precision approach and automatic landing for approach baselines extending to 50 km in some instances [12–15].

However, they are equally applicable to land-based or land-sea applications (e.g., precise desert navigation, off-shore oil exploration). In contrast, land-surveying applications and the like, often involving long baselines, have had the luxury of the postprocessing environment and as a result, accuracies at the millimeter level are commonplace today. Techniques applied in such instances involve resolution of carrier cycle ambiguities on the data sets collected over long periods of time (generally an hour or more). In addition, postprocessing of the data lends itself to recognition and repair of receiver cycle slips. Precision can be further enhanced by use of precise satellite ephemerides. These topics, while of interest, are beyond the scope of this book. Texts such as [16] ably cover these applications.

The following discussion focuses on an integer ambiguity resolution technique first proposed in [17], which capitalizes on some concepts from [18] to resolve the inconsistencies between redundant measurements. The latter work maintains that "all information about the 'inconsistencies' resides in a set of linear relationships known as *parity equations*." While these techniques were originally applied to inertial systems and their associated instruments (e.g., accelerometers and gyros), there is similar applicability to GPS measurement inconsistencies that, in this instance, manifest themselves in the integer wavelength ambiguities inherent in the carrier-phase observables. In [19], it has been shown that a similar approach using a technique that minimizes least-square residuals has application to the rapid resolution of the ambiguities albeit in a static, nonkinematic, environment. This reference also suggests the use of the wide-lane measurements to reduce computational overhead, thus speeding up the ambiguity-resolution process.

12.3.1.1 Combining Receiver Measurements

As mentioned in Chapter 8, two distinct measurements are provided by a GNSS receiver: the pseudorange measurement, also referred to as the code measurement, and the carrier-phase measurement. Code and carrier-phase measurements are available from each signal broadcast by each SV tracked by the receiver. Dual- or triple-frequency GNSS receivers provide such measurements for multiple signals on multiple frequencies. Unfortunately, these measurements are subject to some detrimental effects. Inherent in GNSS signals is a variety of errors (see Chapter 10), errors due to signal propagation through the ionosphere and troposphere, satellite ephemeris errors and clock errors, and noise. Receivers have their own set of problems: clock instability, signal multipath, and also noise. Fortunately, the term DGNSS implies that we have similar sets of measurements from at least two GNSS receivers separated by some fixed distance called a baseline. By forming linear combinations (differences) of like measurements from two receivers, it becomes possible to eliminate errors that are common to both receivers. Such a combination is referred to as a single difference (SD). By differencing two SD measurements from the same SV, we form what is called the double difference (DD). The result is that by using DD processing techniques on the measurements, most of the error sources are removed [6]. One major exception is multipath—it can be mitigated, but not eliminated. Note that receiver noise is still present, but its contribution is generally much less than that of multipath.

12.3.1.2 Carrier Phase Measurement

Once the receiver locks on to a particular satellite, it not only makes pseudorange measurements on each signal, but it also keeps a running cycle count based upon the Doppler frequency shift present on each carrier frequency (one cycle represents an advance of 2π radians of carrier phase or one wavelength). Each epoch, this running cycle count (the value from the previous epoch plus the advance in phase during the present epoch) is available from the receiver. More specifically, the advance in carrier phase during an epoch is determined by integrating the carrier Doppler frequency offset (f_D) over the interval of the epoch. Frequency f_D is the time rate of change of the carrier phase; hence, integration over an epoch yields the carrier phase advance (or recession) during the epoch. Then, at the conclusion of each epoch, a fractional phase measurement is made by the receiver. This measurement is derived from the carrier phase tracking loop of the receiver. Mathematically, for two frequencies (e.g., GPS L1 and L2) the relationship is as follows:

$$\begin{aligned}\phi_{l1_n} &= \phi_{l1_{n-1}} + \int_{t_{n-1}}^{t_n} f_{D_{l1}}(\tau)d\tau + \phi_{r_{1_n}} \quad \text{where } \phi_{l1_0} = A_{l1} \\ \phi_{l2_n} &= \phi_{l2_{n-1}} + \int_{t_{n-1}}^{t_n} f_{D_{l2}}(\tau)d\tau + \phi_{r_{2_n}} \quad \text{where } \phi_{l2_0} = A_{l2}\end{aligned} \tag{12.10}$$

where:

ϕ is the accumulated phase at the epoch shown;
$l1$ and $l2$ are the two L-band frequencies (e.g., L1 and L2 for GPS);
n and $n - 1$ are the current and immediately past epochs;
f_D is the Doppler frequency as a function of time;
ϕ_r is the fractional phase measured at the epoch shown;
A_{l1},A_{l2} are whole plus fractional cycle count (arbitrary) at receiver acquisition.

Even though the receiver carrier-phase measurement can be made with some precision (0.001 cycle for receivers in the marketplace) and any advance in carrier cycles since satellite acquisition by the receiver can be accurately counted, the overall phase measurement contains an unknown number of carrier cycles. This is called the carrier-cycle integer ambiguity (N). This ambiguity exists because the receiver merely begins counting carrier cycles from the time a satellite signal is placed in active track. Were it possible to relate N to the problem geometry, the length of the path between the satellite and the user receiver, in terms of carrier cycles or wavelengths, could be determined with the excellent precision mentioned above.

Figure 12.4 depicts such a situation and also illustrates the effect of the calculated carrier-phase advance as a function of time (e.g., ϕ_1, ϕ_2, and so forth). Clearly, determining N for each satellite used to generate the user position is of paramount concern when interferometric techniques are used. As the term interferometry implies, phase measurements taken at two or more locations are combined. Normally, the baseline(s) between the antennas are known and the problem becomes one of reducing the combined phase differences to determine the precise location of the

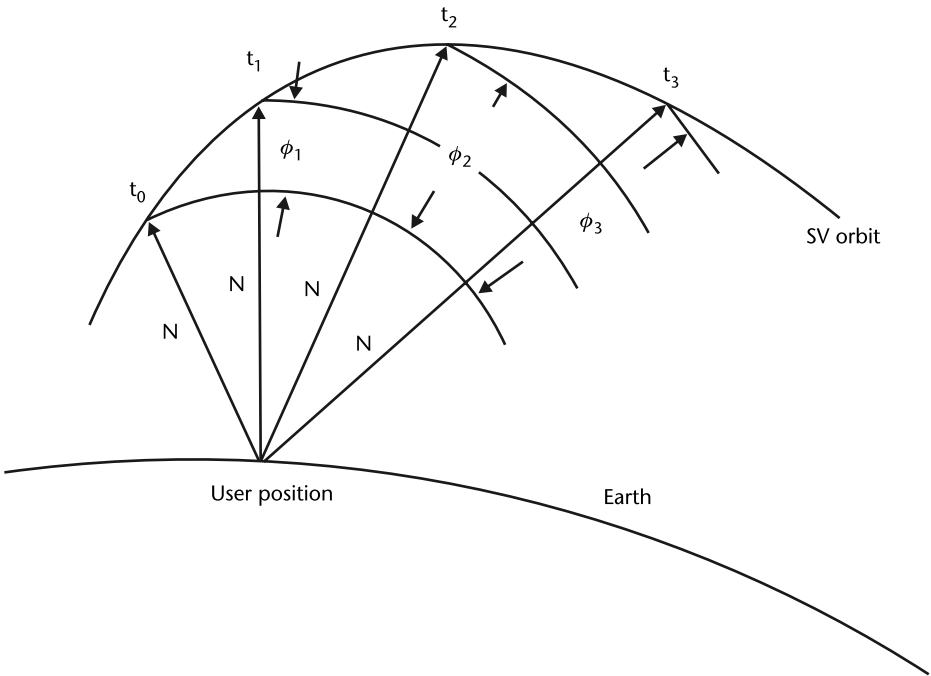


Figure 12.4 Carrier-phase geometric relationships.

source of the signal. In the case of DGNSS, the baseline is unknown, but the location of the signal sources (the GNSS SVs) can be precisely determined using ephemerides available from the navigation data in the satellite transmission.

12.3.1.3 Double-Difference Formation

Generation of both carrier-phase and pseudorange (code) DDs is key to determining the baseline vector between the ground and airborne platform antennas. In so doing, satellite ephemerides must be properly manipulated to ensure that the carrier-phase and code measurements made at the two receiver locations are adjusted to a common measurement time base with respect to GNSS time scale. Formation of the DD offers tremendous advantage because of the ultimate cancellation of receiver and satellite clock biases as well as most of the ionospheric propagation delay. If the two antennas are located at the same elevation, the tropospheric propagation delay will largely cancel as well. This is not the case if one of the antennas is on an airborne platform, and thus the path delay due to the troposphere experienced at the two antenna locations differs based upon their altitude differential.

Carrier-Phase Double Difference

Figure 12.5 schematically depicts a simple GNSS interferometer interacting with a single satellite. The phase centers of two antennas are located at k and m, and **b** represents the unknown baseline between them. SV p is in orbit at a distance of several Earth radii and we assume the paths of propagation between the satellite and the two antennas are parallel. The lengths of the propagation path between SV

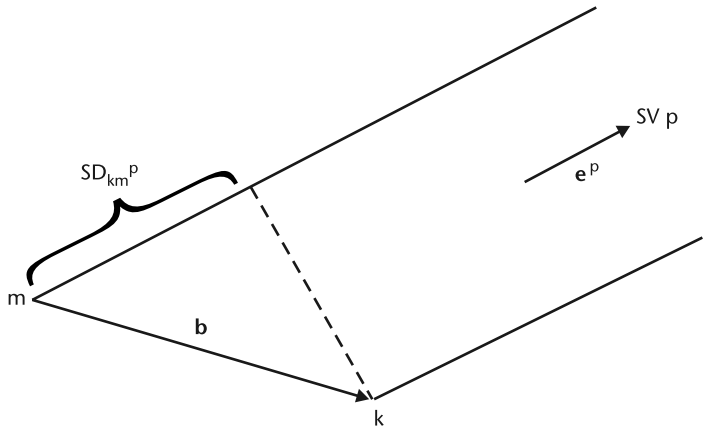


Figure 12.5 GNSS interferometer—one satellite.

p and k (Φ_k^p) or SV p and m (Φ_m^p), in terms of fractional and integer carrier cycles, are as follows:

$$\begin{aligned}\Phi_k^p(t) &= \phi_k^p(t) - \phi^p(t) + N_k^p + S_k + f\tau_p + f\tau_k - \beta_{iono} + \delta_{tropo} \\ \Phi_m^p(t) &= \phi_m^p(t) - \phi^p(t) + N_m^p + S_m + f\tau_p + f\tau_m - \beta_{iono} + \delta_{tropo}\end{aligned} \tag{12.11}$$

where

k and m refer to the receiver/receiver antennas phase centers;

p is the satellite signal source;

ϕ^p is the transmitted satellite signal phase as a function of time;

$\phi_k^p(t)$ and $\phi_m^p(t)$ are the receiver-measured satellite signal phase as a function of time;

N is the unknown integer number of carrier cycles from SV p to k or SV p to m;

S is phase noise due to all sources (e.g., receiver, multipath);

f is the carrier frequency;

τ is the associated satellite or receiver clock bias;

β_{iono} is the advance of the carrier (cycles) due to the ionosphere;

δ_{tropo} is the delay of the carrier (cycles) due to the troposphere.

The minus sign associated with the ionospheric effects will be discussed later in this section.

The interferometric variable, the single difference (SD), is now created by differencing the carrier-cycle propagation path lengths (SV p to k and SV p to m):

$$SD_{km}^p = \phi_{km}^p + N_{km}^p + S_{km}^p + f\tau_{km} \tag{12.12}$$

The nomenclature remains the same as in (12.11), but certain advantages accrue in forming the SD metric. Prime among these are the cancellation of the transmitted

satellite signal phase and clock biases and the formation of a combined integer ambiguity term that represents the integer number of carrier cycles along the path from m to the projection of k onto the mp line of sight. A combined phase-noise value has been created as well as a combined receiver clock-bias term. With regard to the ionosphere and troposphere, these effects largely cancel too if the receivers are co-altitude and closely spaced (baselines less than 50 km). This condition will be assumed to exist for purposes of the discussion. (See Chapter 10 for a discussion of differential ionospheric and tropospheric error characteristics.) Errors in satellite ephemerides (see Chapter 10) have not been considered, but usually have very small effect. Since they are a common term like the satellite clock bias, they cancel when the single difference is formed.

Figure 12.6 extends the GNSS interferometer to two satellites. For q, the additional SV, a second SD metric can be formed:

$$SD_{km}^{q} = \phi_{km}^{q} + N_{km}^{q} + S_{km}^{q} + f\tau_{km} \tag{12.13}$$

As with (12.12), the expected cancellation of SV transmitted signal phase and clock bias occurs, and a short baseline will be assumed such that ionospheric and tropospheric propagation delays cancel as well.

The interferometric DD is now formed using the two SDs. Involved in this metric are two separate satellites and the two receivers, one at either end of the baseline, **b**. Differencing equations (12.12) and (12.13) yield the following:

$$DD_{km}^{pq} = \phi_{km}^{pq} + N_{km}^{pq} + S_{km}^{pq} \tag{12.14}$$

where the superscripts p and q refer to the individual satellites, and k and m are the individual receivers. With the formation of the DD, the receiver clock-bias terms now cancel. Remaining is a phase term representing the combined carrier-phase measurements made at k and m by the receivers using SVs p and q, an integer term made up of the combined unknown integer ambiguities and a system phase-noise term consisting primarily of combined multipath and receiver effects [19]. It now remains to relate the DD to the unknown baseline **b**, which exists between the two receiver antennas.

Referring again to Figure 12.6, it is evident that the projection of **b** onto the line of sight between p and m can be written as the inner (dot) product of **b** with a unit vector $\mathbf{e}^{p}$ in the direction of SV p. This projection of **b** (if converted to wavelengths by dividing by λ) is SD_{km}^{p}. Similarly, the dot product of **b** with a unit vector $\mathbf{e}^{q}$ in the direction of SV q would result in SD_{km}^{q}. Rewriting SD equations (12.12) and (12.13) with this substitution yields:

$$\begin{aligned} SD_{km}^{p} &= \left(\mathbf{b}\cdot\mathbf{e}^{p}\right)\lambda^{-1} = \phi_{km}^{p} + N_{km}^{p} + S_{km}^{p} + f\tau_{km} \\ SD_{km}^{q} &= \left(\mathbf{b}\cdot\mathbf{e}^{q}\right)\lambda^{-1} = \phi_{km}^{q} + N_{km}^{q} + S_{km}^{q} + f\tau_{km} \end{aligned} \tag{12.15}$$

Clearly, we can incorporate this result into the double difference as well:

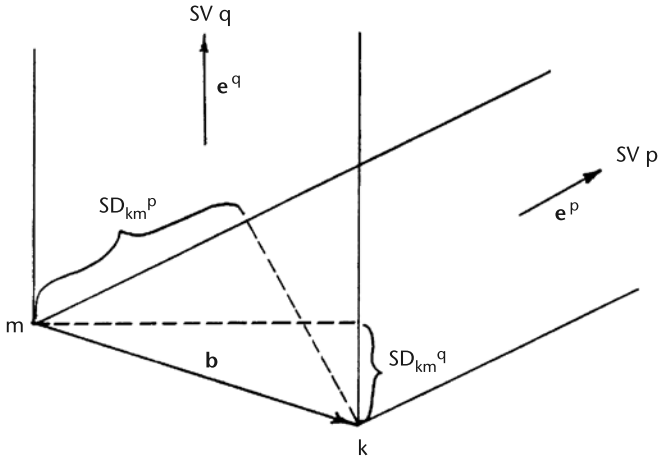


Figure 12.6 GNSS interferometer—two satellites.

$$DD_{km}^{pq} = \left(\mathbf{b} \cdot \mathbf{e}^{pq}\right)\lambda^{-1} = \phi_{km}^{pq} + N_{km}^{pq} + S_{km}^{pq} \tag{12.16}$$

where $\mathbf{b} \cdot \mathbf{e}^{pq}$ is the inner product between the unknown baseline vector and the difference of the unit vectors to SVs p and q. Since determining the unknown baseline between the antennas is at the heart of the matter, it is this second formulation for the DDs, (12.16), that will serve as the basis for further derivation.

Of the variables shown in (12.16), there is only one that can be precisely measured by the receiver and that is the carrier phase. In actuality, then, it is the carrier-phase measurements of the receivers that are combined to produce the DDs. The term DD_{cp} is adopted to represent this and implicit in its formulation is conversion to meters. The noise term will be dropped to simplify the expression. In the end, as the carrier-cycle ambiguity search progresses, the noise sources tend to cancel. There remains to be determined the baseline vector ($\mathbf{b}$), which has three components (b_x, b_y, b_z), plus an unknown integer carrier-cycle ambiguity (N) associated with each of the DD_{cp} terms. Toward this end, four double-differences will be used. While additional double-differences could be formed depending on the number of satellites in track by the receiver, this is a sufficient number and minimizes the computational requirements of the carrier-cycle ambiguity-search algorithm. In terms of satellites, two satellites are required to form each double difference. Thus, in order to form four DD equations, a minimum of five satellites is necessary. The transfiguration and extension of (12.16) to four double differences appears as follows:

$$\begin{bmatrix} DD_{cp1} \\ DD_{cp2} \\ DD_{cp3} \\ DD_{cp4} \end{bmatrix} = \begin{bmatrix} e_{12x} & e_{12y} & e_{12z} \\ e_{13x} & e_{13y} & e_{13z} \\ e_{14x} & e_{14y} & e_{14z} \\ e_{15x} & e_{15y} & e_{15z} \end{bmatrix} \begin{bmatrix} b_x \\ b_y \\ b_z \end{bmatrix} + \begin{bmatrix} N_1 \\ N_2 \\ N_3 \\ N_4 \end{bmatrix} \lambda \tag{12.17}$$

where DD_{cp1}, for example, is the first of four double-differences; $\mathbf{e}_{12}$ represents the differenced unit vector between the two satellites under consideration; **b** is the baseline vector; N_1 is the associated integer carrier-cycle ambiguity; and λ is the applicable wavelength. The wavelength is introduced at this point to provide consistency with DD_{cp} and **b**, which are now in meters. During this and subsequent discussion, all double-difference formulations will be in units of length. Using the matrix notation, (12.17) takes the following form:

$$\mathbf{DD}_{cp} = \mathbf{Hb} + \mathbf{N}\lambda \tag{12.18}$$

where $\mathbf{DD}_{cp}$ is a 4×1 column matrix of carrier-phase double differences; **H** is a 4×3 data matrix containing the differenced unit vectors between the two satellites represented in the corresponding DD, **b** is a 3×1 column matrix of the baseline coordinates, and **N** is a 4×1 column matrix of integer ambiguities. Once the carrier-phase DDs are formed, a similar set of DDs is determined using the pseudoranges between each antenna and the same set of satellites.

Pseudorange (Code) Double Difference

As in the case of the carrier phase measurement, the receiver makes a pseudorange measurement each epoch for all satellites and all signals being actively tracked. The pseudorange suffers from similar propagation and timing effects as is the case for the carrier phase. The only basic difference is that where the ionosphere advances the carrier phase, the pseudorange information experiences a group delay. In considering the propagation of electromagnetic waves through a plasma, of which the ionosphere is an example, the propagation velocity (v_g) of the modulation on a carrier is retarded, while the phase velocity (v_p) of the carrier itself is advanced [20] (see Chapter 10). The following relationship holds:

$$v_g v_p = c^2 \tag{12.19}$$

where c is the speed of light. Thus, when the code double difference is formed, the effects of the ionospheric delay are additive. Formulation of the code double difference begins with the pseudorange equation, as follows:

$$\begin{aligned} P_k^p(t) &= t_k^p(t) - t^p(t) + Q_k + \tau_p + \tau_k + \gamma_{iono} + \delta_{tropo} \\ P_m^p(t) &= t_m^p(t) - t^p(t) + Q_m + \tau_p + \tau_m + \gamma_{iono} + \delta_{tropo} \end{aligned} \tag{12.20}$$

where

P is the receiver-measured pseudorange as a function of time in seconds;

k, m refer to receiver/receiver antennas phase centers;

p is the satellite-signal source;

t_k^p or t_m^p is signal-reception time as measured by the receiver clocks;

t^p is signal-transmission time as determined from the SV clock;

Q is noise (timing jitter) due to all sources (e.g., receiver, multipath);

τ is the associated satellite or receiver clock bias;

γ_{iono} represents group delay (s) of the modulation due to the ionosphere;

δ_{tropo} represents the delay (s) of the modulation due to the troposphere.

Note the absence of the integer carrier-cycle ambiguity N—the pseudorange measurement is unambiguous. In other words, code DD observables formed from the pseudoranges measured by the receivers contain no carrier-cycle ambiguities. Unfortunately, pseudorange cannot be measured as precisely as the carrier phase, so it is noisier. Also of note is the change in the sign for the ionospheric effects from that in (12.11) due to the group delay. The unambiguous nature of the code DD will serve as the basis for code/carrier smoothing to be described in the next section.

Pseudorange SDs are now formed:

$$\begin{aligned} SD_{km}^{p} &= t_{km}^{p} + Q_{km}^{p} + \tau_{km} \\ SD_{km}^{q} &= t_{km}^{q} + Q_{km}^{q} + \tau_{km} \end{aligned} \tag{12.21}$$

Finally, the pseudorange DD, in meters, is formed:

$$DD_{km}^{pq} = t_{km}^{pq} + Q_{km}^{pq} \tag{12.22}$$

Paralleling the development of the carrier phase DDs, the same five satellites are used to form four code DDs. Figure 12.7 is similar to Figure 12.5 with the exception that it has been labeled in terms of pseudoranges. It is evident that the inner product of the baseline **b** and the unit vector to satellite p can be expressed as the difference of two pseudoranges to the SV, one measured at receiver antenna k, the other at m. Recasting the baseline vector **b** in terms of the code SDs and DDs is virtually identical to that previously done with the carrier phase SD and DD formulations. There is one very important difference, however, and it is that there are no ambiguities when code measurements are used. Further, the DDs are converted

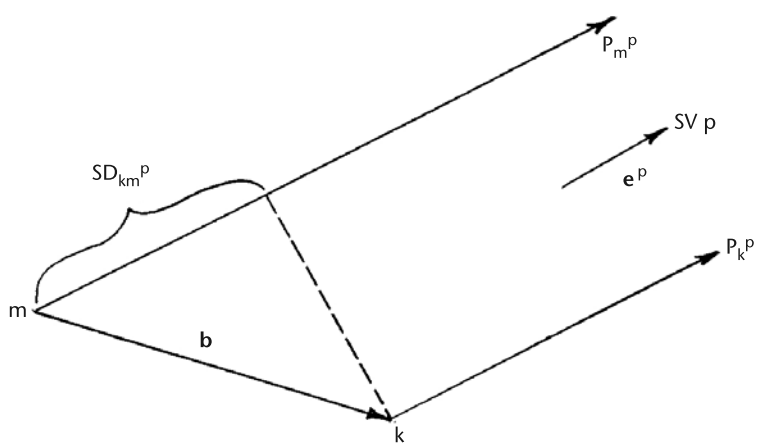


Figure 12.7 Code-equivalent GNSS interferometer.

to units of length by multiplying by the speed of light and, for simplicity, the noise term is dropped. The pseudorange-based equivalent of (12.17) is depicted here:

$$\begin{bmatrix} DD_{pr1} \\ DD_{pr2} \\ DD_{pr3} \\ DD_{pr4} \end{bmatrix} = \begin{bmatrix} e_{12x} & e_{12y} & e_{12z} \\ e_{13x} & e_{13y} & e_{13z} \\ e_{14x} & e_{14y} & e_{14z} \\ e_{15x} & e_{15y} & e_{15z} \end{bmatrix} \begin{bmatrix} b_x \\ b_y \\ b_z \end{bmatrix} \tag{12.23}$$

Once again, the integer ambiguities N, as appear in (12.17), are absent because the pseudorange is unambiguous. Using matrix notation to express (12.23) yields the following, which is the code double-difference counterpart of (12.18):

$$\mathbf{DD}_{pr} = \mathbf{Hb} \tag{12.24}$$

where $\mathbf{DD}_{pr}$ is the 4×1 column matrix of pseudorange (code) double differences; $\mathbf{H}$ is a 4×3 data matrix containing the differenced unit vectors between the two satellites represented in the corresponding DD; and $\mathbf{b}$ is a 3×1 column matrix of the baseline coordinates.

12.3.1.4 Pseudorange (Code) Smoothing

Thus far in this description of GNSS interferometry, two distinct sets of DDs have been created. The first set is based upon differencing the low noise (less than 1 cm) but ambiguous carrier phase measurements; the second set is formed from the unambiguous but noisier (1 to 2m) pseudorange (code) measurements. The two sets of measurements can be combined using a variety of techniques to produce a smoothed-code DD measurement. This is extremely important since the baseline vector $\mathbf{b}$ determined from the smoothed code DDs provides an initial solution estimate for resolving the carrier-cycle integer ambiguities. Based on [17], a complementary Kalman filter is used to combine the two measurement sets. The technique uses the average of the noisier code DDs to center the quieter carrier phase DDs, thereby placing a known limit on the size of the integer ambiguity.

The filter equations are as follows:

$$\begin{aligned} DD_{s_n}^{-} &= DD_{s_{n-1}}^{+} + \left(DD_{cp_n} - DD_{cp_{n-1}}\right) \\ p_n^{-} &= p_{n-1}^{+} + q \\ k_n &= p_n^{-}\left(p_n^{-} + r\right)^{-1} \\ DD_{s_n}^{+} &= DD_{s_n}^{-} + k_n\left(DD_{pr_n} - DD_{s_n}^{-}\right) \\ p_n^{+} &= (1 - k_n)p_n^{-} \end{aligned} \tag{12.25}$$

The first line of (12.25) propagates the smoothed-code double difference to the current time epoch (n) using the estimate of the smoothed-code double difference from

the previous epoch ($n-1$) and the difference of the carrier-phase double difference across the current and past epochs. The estimate (DD_s^+), which is based upon averaging the DD_{pr} (code) difference, centers the calculation; the DD_{cp} (carrier-phase) difference adds the latest low-noise information. Note that differencing two carrier-phase double differences across an epoch removes the integer ambiguity; hence, the propagated smoothed-code double difference (DD_s^-) remains unambiguous. The estimation-error variance (p_n^-) is brought forward (line two) using its previously estimated value plus the variance of the carrier-phase double-difference measurement q. The Kalman gain is next calculated in preparation for weighting the effect of the current code double-difference measurement. Line three shows that as the variance on the code double difference r approaches zero, the Kalman gain tends to unity. This is not surprising since the higher the accuracy of a measurement (smaller the variance), the greater is its effect on the outcome of the process. Lines four and five of (12.25) propagate the estimate of the smoothed-code double difference (DD^+) and estimation-error variance to the current epoch (n) in preparation for repeating the process in the next epoch ($n+1$). DD^+ (to be used in the next epoch) involves the sum of the current value of the smoothed-code double difference (just predicted) and its difference from the current code double difference (just measured) weighted by the Kalman gain. Intuitively, if the prediction is accurate, then there is little need to update it with the current measurement. Finally, the estimation-error variance p is updated. The update maintains a careful balance between the goodness of the code and that of the carrier-phase DDs based upon whether the Kalman gain approaches unity or zero or lies somewhere in between.

Equation (12.25) represents a set of scalar complementary Kalman filter equations that can operate on each of the requisite double-difference measurement pairs (code and carrier-phase) in turn. Alternatively, these equations can be set up in matrix form and accomplish the same end once all DD measurements for a given epoch are calculated and collected together in respective arrays. Either approach is satisfactory, but, for ease of programming, the scalar formulation is used here.

Figure 12.8 shows actual carrier phase (top) and code (bottom) DD measurements collected over a period of 20 minutes during a flight test [17]. The offset between the two plots is arbitrary, but can be thought of in terms of some unknown ambiguity included in the carrier-phase DD measurements. It is apparent that the two sets of data are quite similar with the exception of apparent noise on the code DDs.

Figure 12.9 shows the output of the complementary Kalman filter (i.e., the smoothed-code DDs (bottom) and the original carrier phase DDs (top) over the same 20-min interval). With the exception of the first few epochs (nominally about 10), the smoothed-code DD virtually mirrors the carrier phase DD. It has the added advantage that it is centered about the original code DD measurements and is thus unambiguous.

Depending upon the multipath in the local environment, the smoothed-code DD, once the complementary Kalman filter is initialized, is generally within ±1 to 2m. In terms of carrier wavelengths, this represents approximately ±5 to 10λ at L1.

Figure 12.10 shows the difference between the carrier phase and smooth-code DDs of Figure 12.9 with the nominal offset removed from the former. For this particular set of data, the difference is well within ±1m and is indicative of low multipath at both the ground and airborne antennas. Again, the behavior prior to

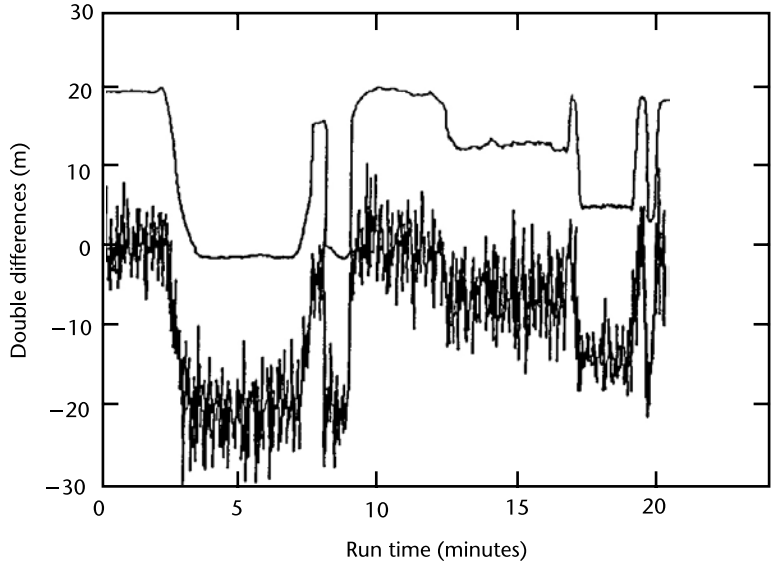


Figure 12.8 Carrier-phase (top) and raw-code DDs (bottom).

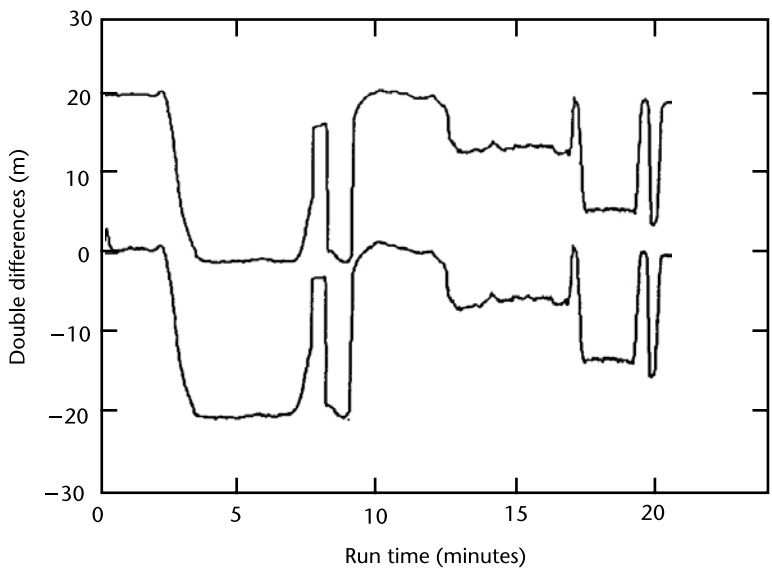


Figure 12.9 Carrier-phase (top) and smoothed-code DDs (bottom).

completing the initialization of the complementary Kalman filter is clearly evident during the first few epochs, but once initialized, the difference is very well behaved.

12.3.1.5 Initial Baseline Determination (Float Solution)

The smoothed-code DD from the complementary Kalman filter, once the filter is initialized, is key to determining the float solution. The float baseline solution is a least squares fit yielding an estimate of the baseline vector **b**, accurate to within a few integer wavelengths depending upon the effects of satellite geometry and the

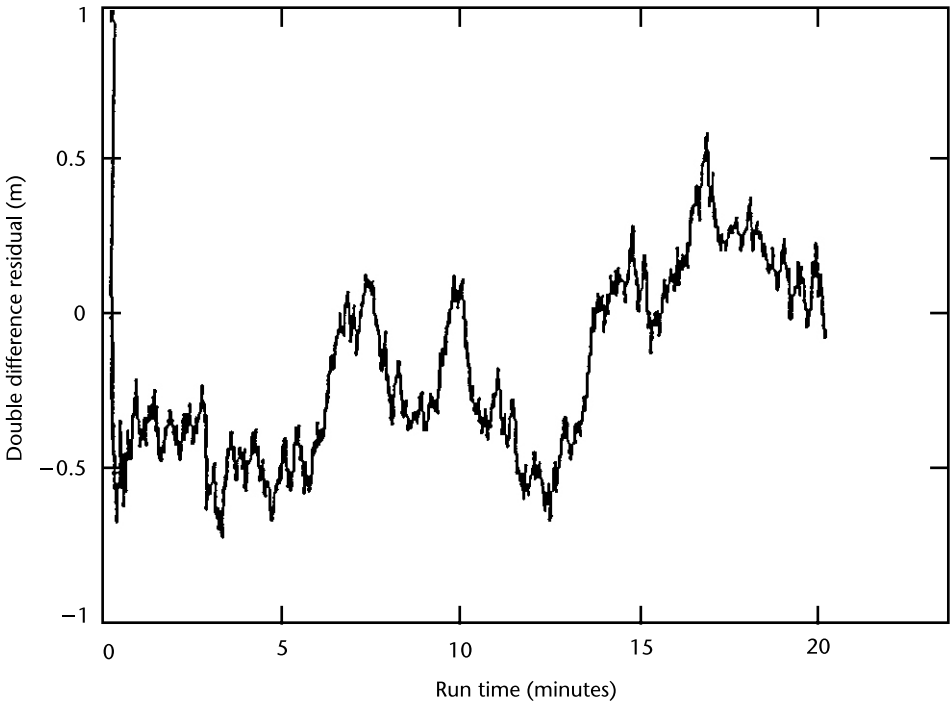


Figure 12.10 Carrier-phase minus smoothed-code DDs.

severity of the multipath environment surrounding the antenna at either end of the baseline.

Using the vector notation introduced with (12.24), the DD baseline equation for the smoothed-code DDs is as follows:

$$\mathbf{DD}_s = \mathbf{H}\mathbf{b}_{float} \tag{12.26}$$

In a general least-squares sense, $\mathbf{DD}_s$ is an $m \times 1$ column matrix of DDs for $m + 1$ SVs, $\mathbf{H}$ is an $m \times 3$ data matrix containing the differenced unit vectors between the two SVs represented in the corresponding DD, and $\mathbf{b}$ is a 3×1 column matrix of the estimated float baseline solution coordinates. Were the least squares solution for $\mathbf{b}$ the only desired result, the generalized inverse approach $\mathbf{H}^T\mathbf{H}$ could be applied immediately. In this situation, however, the float baseline solution represents an intermediate step along the way to the desired final result, which is an integer-ambiguity resolved or *fixed* baseline solution. With this end in mind, some matrix conditioning is performed on the elements of (12.26) prior to determining the floating baseline solution. The $\mathbf{H}$ matrix is decomposed using $\mathbf{QR}$ factorization where $\mathbf{Q}$ is a real, orthonormal matrix (thus, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$) and $\mathbf{R}$ is an upper triangular matrix [21]. $\mathbf{QR}$ factorization allows the least-squares residual vector to be obtained by projecting the DDs onto a measurement space that is orthogonal to the least-squares solution space spanned by the columns of $\mathbf{H}$. Hence, the least-squares residual vector is projected onto the left null space of $\mathbf{H}$, called parity space, while the least-squares solution is mapped onto the column space of $\mathbf{H}$, known as the estimation space [18]. Since the parity space and the estimation space are orthogonal, the residuals therein are independent of the estimate. This property will be used to

an advantage to isolate the carrier-cycle integer ambiguities and subsequently adjust the smoothed-code DDs. Incorporating the properties of the **QR** factorization into (12.26) yields:

$$\mathbf{DD}_s = \mathbf{QRb}_{float} \tag{12.27}$$

Capitalizing on the property of the orthonormal matrix where the inverse and transpose are equivalent, and then rearranging gives:

$$\mathbf{Rb}_{float} = \mathbf{Q}^T\mathbf{DD}_s \tag{12.28}$$

Expanding the matrices for clarity yields:

$$\begin{bmatrix} R_{11} & R_{12} & R_{13} \\ 0 & R_{22} & R_{23} \\ 0 & 0 & R_{33} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_x \\ b_y \\ b_z \end{bmatrix} = \begin{bmatrix} Q_{11}^T & Q_{12}^T & Q_{13}^T & Q_{14}^T \\ Q_{21}^T & Q_{22}^T & Q_{23}^T & Q_{24}^T \\ Q_{31}^T & Q_{32}^T & Q_{33}^T & Q_{34}^T \\ q_1 & q_2 & q_3 & q_4 \end{bmatrix} \begin{bmatrix} DD_{s_1} \\ DD_{s_2} \\ DD_{s_3} \\ DD_{s_4} \end{bmatrix} \tag{12.29}$$

Equation (12.29) lends itself readily to horizontal partitioning, and elements of the $\mathbf{Q}^T$ matrix have been labeled with capital Q to show the portion which corresponds to the least-squares solution (estimation space) and with small q to indicate the elements making up the least-squares residual vector (parity space). The partitioning of (12.29) is:

$$\mathbf{R}_u\mathbf{b}_{float} = \mathbf{Q}_u{}^T\mathbf{DD}_s \tag{12.30}$$

$$\mathbf{0} = \mathbf{q}\ \mathbf{DD}_\mathbf{s} \tag{12.31}$$

Solving (12.30) gives the float baseline solution shown here:

$$\mathbf{b}_{float} = \mathbf{R}_u{}^{-1}\mathbf{Q}_u{}^T\mathbf{DD}_s \tag{12.32}$$

Equation (12.31), while ideally equal to zero, is the least-squares residual vector and can be exploited to provide the means for resolving the carrier-cycle integer ambiguities, a discussion of which follows in the next section. The float baseline solution is freshly calculated for each epoch and serves as a temporal benchmark during the ambiguity resolution process while the fixed baseline solution is being pursued. Once the fixed baseline solution is in hand, the float solution subsequently serves as a cross-check to ensure the continued integrity of the former. Recall that this is a dynamic process—one end of the baseline is usually in motion (e.g., airborne); thus, both the fixed and float baseline solutions will vary from epoch to epoch and must be constantly monitored. However, the carrier-cycle integer ambiguities, once resolved, remain fixed in the solution since the receivers dynamically track the change (i.e., growth or contraction) in the number of carrier cycles

between the baseline antennas and the respective SVs used in the solution for the baseline. This holds true as long as all SVs remain in constant track by the receivers with no cycle slips occurring.

12.3.1.6 Carrier-Cycle Ambiguity Resolution

Using the complementary Kalman filter to produce the smoothed-code DDs ensures that each of the DD measurements contributes to a solution whose accuracy is within 1 to 2m, as previously stated. In terms of integer wavelengths at L1, for example, the DDs values are each within about ± 5 to 10λ. Intuitively, it would seem possible to iterate each DD through this range of carrier wavelengths, recalculate the least-squares solution for each iteration, and then examine the residuals. Residuals near zero, since there is noise in the process, would be identified, and the number of integer wavelengths added to each of the DDs would be kept as a candidate integer ambiguity set for the particular trial. This could be done on an epoch-by-epoch basis and those sets of integer ambiguities that continued to remain valid would be marked and tallied. The list would diminish over time and eventually one set of integer ambiguities would emerge victorious. The approach just described would take place in parity space since we would be adjusting the DDs measurements (iteratively) and subsequently examining the new set of least-squares residuals that resulted. However, to search the uncertainty volume about the float baseline solution would be computational inefficiency at its extreme. For example, an uncertainty of $\pm 11\lambda$ would require initially that 23^4 least-squares solutions be generated each epoch and the residuals for each examined. Even though the number would diminish over time, the technique would in general remain computationally inefficient.

A far better approach is to screen candidate integer ambiguity sets/test points using predetermined criteria, and then test only those sets which meet that criteria. Examples of prominent techniques include the fast ambiguity resolution approach (FARA) [22], the least-squares ambiguity search technique (LSAST) [23], the ambiguity function method (AFM) [7] (later refined in [6]), the fast ambiguity search filter (FASF) [24], and the LAMBDA [25]. As these algorithms developed, they lent themselves to real-time and kinematic applications, and once ambiguity fixing was made robust, such baseline solutions have since been referred to as RTK.

To generalize the foregoing, an initial solution is obtained, a search domain about the solution is established, and some methodology is used to preselect candidate test points/ambiguity sets within the domain, which are subsequently used to generate candidate fixed baseline solutions. Using a given selection criteria, the candidate fixed baseline solutions are accepted or rejected until ultimately only one remains. Finally, some validation metric is used to test the selected candidate ambiguities. AFM and LSAST can potentially accomplish this in a single period of several minutes; FARA generally requires two to three such periods. It was subsequently pointed out in [25] that double differences are usually highly correlated and subject to poor precision. This leads to the possibility that the search space, while centered on the float baseline solution, may not in fact even contain the carrier-cycle ambiguities. Toward this end, LAMBDA uses an ambiguity transformation that decorrelates the ambiguities and reshapes the search space. For example, a highly elongated ellipse in a two-dimensional ambiguity example becomes near

circular in the transformed domain. The new ambiguity-search space is constructed to be integer in nature and volume preserving. Further, the search space can be appropriately scaled and all but guarantees that the proper set of integer ambiguities resides within.

Thus far in the development of the parity-space methodology used for resolving the carrier-cycle integer ambiguities, the steps necessary to develop the initial solution have been covered in detail. These include the formation of both carrier-phase and pseudorange (code) DDs, the creation of the smoothed-code DDs, the formulation of the DD baseline equations, and the separation of these equations into a least-squares (float) baseline solution and a least-squares residual vector. A search volume has been established based upon the accuracy inherent in the smoothed-code DDs—nominally ±1 to 2m. It remains to formulate the DDs in a manner such that they can be selectively examined over the search volume as a function of integer carrier-cycle ambiguities. Once this is accomplished, candidate ambiguity sets can be isolated, thresholded, and eventually retained or eliminated. From those few remaining sets, fixed baseline solutions are determined. These solutions are then subjected to additional checks (e.g., comparison with the float solution, among others), until the ultimate fixed baseline solution emerges.

The **QR** factorization is a powerful technique that allows the least-squares residuals to be isolated from the least-squares solution space without the necessity of performing the least-squares solution itself. Application of the residuals to the process of sorting out the integer ambiguities is the next area of interest. To do so requires that the carrier-phase DD measurement be examined in light of its constituent parts. The following equation so illustrates:

$$DD_{cp} = \left(\phi_{DD} + \hat{n} + R_b + S\right)\lambda \tag{12.33}$$

where ϕ_{DD} is the double-difference fractional phase from the receiver measurements, $\hat{n}$ is the unknown double-difference ambiguity, R_b is the inherent receiver channel bias plus residual propagation delays, S is the noise due to all sources (e.g., receiver, multipath), and the use of λ converts the DD to units of length. Strictly speaking, multipath is not noise. However, it does add a noise-like uncertainty to the DD measurement, which unfortunately cannot be uniquely separated at a given instant in time from other noise sources. To solve this dilemma, multipath is simply included with the noise.

Equation (12.33) can be reexpressed using the smoothed-code DDs and with the knowledge that the uncertainty in the sources on the right-hand side of the equation is bounded. The terms ϕ_{DD} and $\hat{n}$ are replaced with ρ_{DD} and $\tilde{n}$. This follows from the knowledge that the smoothed-code DDs are accurate to within 1 to 2m, their inherent noise level. This noise level is equivalent to, for example, ±11 wavelengths at GPS L1 and allows the integer ambiguity to be bounded; hence, $-11 \le \tilde{n} \le +11$. The term ρ_{DD}, then, represents the geometric distance (in carrier-cycles) of the smoothed-code double difference within the noise bound. The equation now appears as follows:

$$DD_s = \left(\rho_{DD} + \tilde{n} + R_b + S\right)\lambda \tag{12.34}$$

The resolution of $\tilde{n}$ can now be attacked using the residuals from the least-squares solution developed as (12.30). This equation is expanded and shown here:

$$\begin{aligned} qDD_s = &\Big[q_1 \left(\rho_{DD_1} + \tilde{n}_1 + R_{b_1} + S_1 \right) + q_2 \left(\rho_{DD_2} + \tilde{n}_2 + R_{b_1} + S_2 \right) \\ &+ q_3 \left(\rho_{DD_3} + \tilde{n}_3 + R_{b_3} + S_3 \right) + q_4 \left(\rho_{DD_4} + \tilde{n}_4 + R_{b_4} + S_4 \right) \Big] \lambda \\ &= \eta \end{aligned} \tag{12.35}$$

where the q_r are the elements of the least-squares residual vector and $\tilde{n}_r$ represents a wavelength ambiguity number associated with the applicable DD. Ideally, the value of η, the measurement inconsistency, would be zero, but this could only be true in the presence of noiseless measurements and resolved carrier-cycle integer ambiguities.

At any particular epoch, the values for q remain constant—the residual of the least squares solution does not change until another set of measurements is taken, the DDs are computed and smoothed, and the **QR** factorization completed. In modern receivers, great effort is expended to minimize interchannel biases; the same holds true for receiver noise. This leaves multipath as the major component of noise. Fortunately, code multipath, over time, behaves in a noise-like fashion, although not necessarily tending to a zero mean [26]. It is worthwhile then to consider (12.35) with emphasis on the component that is constant from epoch to epoch, knowing that the other sources of error will be mostly random or small over an extended period of time. This component is the unknown carrier-cycle integer ambiguity in each of the smoothed-code DDs. If the ambiguity can be removed from the DD, then the only remaining error sources are noise-like and will approach zero or, in the case of multipath, some mean value. Equation (12.35) is rewritten here in light of these ideas:

$$\begin{aligned} &q_1 \left[DD_{s1} - \tilde{n}_1 \lambda \right] + q_2 \left[DD_{s2} - \tilde{n}_2 \lambda \right] \\ &+ q_3 \left[DD_{s3} - \tilde{n}_3 \lambda \right] + q_4 \left[DD_{s4} - \tilde{n}_4 \lambda \right] = \gamma \end{aligned} \tag{12.36}$$

Once again it is noted that the smoothed-code double differences are bounded within ±1 to 2m depending upon the multipath environment. With this in mind, the values for $\tilde{n}$ in (12.36) can be adjusted such that the result is near to zero—at least within some predetermined threshold (γ). Assuming that the receiver noise and interchannel biases can be kept to below $\lambda/2$ (which is generally the case), it becomes possible to use (12.36) to resolve the carrier-cycle ambiguities. Putting (12.36) into matrix form:

$$\mathbf{q} \left[\mathbf{DD}_s - \mathbf{N} \lambda \right] = \gamma \tag{12.37}$$

where $\mathbf{N} = \left[\tilde{n}_1 \quad \tilde{n}_2 \quad \tilde{n}_3 \quad \tilde{n}_4 \right]$ and represents a set of integer values that when substituted into the equation, satisfy the threshold constraint (i.e., γ). The question now becomes one of how to find the **N** vectors that produce such a result.

Since there are only four multiplication operations and three additions required to examine each case, one answer to such a question is to use an exhaustive search. With a ±1 to 2m bound on the accuracy using the smoothed-code DDs, such a search requires that components of **N** contain iterations covering $\pm 11\lambda$ at L1 where the wavelength is 19.03 cm. There are 23^4, slightly less than 300,000, possible candidates for the first epoch, which is not an unreasonable number. Were it necessary, more efficient search strategies could be implemented; however, when the wide-lane wavelength is examined at the end of this chapter, the number of candidates will drop to less than 3,000, which then makes the exhaustive search almost trivial. In any event, as the integer values are cycled from [–11 –11 –11 –11] to [+11 +11 +11 +11], those integer sets that are within the threshold are retained and become candidates for the fixed baseline solution.

12.3.1.7 Final Baseline Determination (Fixed Solution)

Each epoch, the various **N** sets that meet the γ threshold constraint of (12.37) are stored or, if stored previously, a counter (j) is incremented to indicate persistence of the particular ambiguity set. For those sets that persist, a sample mean (η_{avg}) is calculated based upon the first 10 values of the residual. The variance (η_{σ^2}) about the sample mean is determined as well. These calculations are as follows:

$$\eta_{avg_j} = \frac{\left[\eta_{avg_{j-1}}(j-1) + \eta_j\right]}{j} \qquad j \le 10 \text{ and } \eta_{avg_0} = 0 \tag{12.38}$$

$$\begin{aligned} \eta_{\sigma_j^2} &= \eta_{\sigma_{j-1}^2} + \left(\eta_{avg_j} - \eta_j\right)^2 \qquad j = 1,2 \\ \eta_{\sigma_j^2} &= \frac{\left[(j-2)\eta_{\sigma_{j-1}^2} + (\eta_{avg_j} - \eta_j)\right]^2}{(j-1)} \qquad j > 2 \end{aligned} \tag{12.39}$$

Those ambiguity sets with the smallest variance (usually about 10 in number) are then ranked in ascending order. Persistence is defined as a minimum of 10 epochs (seconds for the research upon which this discussion is based) and has been determined experimentally. For a particular ambiguity set to be selected for the fixed solution, one additional requirement must now be met. The ratio of the residual calculated for the ambiguity set with the smallest and next smallest variances must exceed a minimum value. This value has also been determined experimentally and set to 0.5.

Upon selection of an ambiguity set, the $\tilde{n}$ values of the **N** vector multiplied by λ become literally the amount of path length used to adjust the current smoothed-code DD to create the exact (resolved) DD path length. To complete the process, the smoothed-code DDs are recomputed using the ambiguity set(s) that were generated during the search/selection process. The following relationship is used:

$$\mathbf{DD}_r = \mathbf{DD}_s - \mathbf{N}\lambda \tag{12.40}$$

The resolved smoothed-code double differences ($\mathbf{DD}_r$) are then used to calculate the fixed baseline solution using (12.32) as modified here:

$$\mathbf{b}_{fixed} = \mathbf{R}_u^{-1}\mathbf{Q}_u^T\mathbf{DD}_r \tag{12.41}$$

The RMS of the difference between the float and fixed baseline solution is calculated for the current and subsequent epochs and monitored for consistency. Should the difference begin to diverge, the fixed baseline solution is discarded and a new search for integer ambiguities begun. Recall that the receivers, once acquisition of a given SV is established, keep track of advances or retreats in the receiver-to-satellite path length. Hence, a valid integer-ambiguity set in one epoch remains equally valid in the next and subsequent epochs. This being the case, the fixed baseline solution can be recalculated each epoch by adjusting the current set of smoothed-code DDs with the resolved ambiguity set (**N**), followed by an updated least squares solution [i.e., successive application of (12.40) and (12.41)]. Particularly noteworthy is that during the entire carrier-cycle ambiguity resolution process, it is unnecessary to generate the least-squares solution. All calculations remain in the measurement (parity) space using the least-squares residual vector obtained during the **QR** factorization. It is only after the proper consistency among the measurements (DDs) emerges (i.e., the emergence of a final resolved integer-ambiguity set) that the fixed baseline solution is calculated. True, the float baseline solution is calculated each epoch, but this is more for monitoring than mathematical necessity. Remaining in the measurement space minimizes computational overhead and speeds the process as a result.

Two separate phenomena work to accelerate the process. First, the GNSS constellation is dynamic. Its movement in relationship to the ground and user receiver antennas provides an overall change in geometry that has a very positive influence when interferometric techniques are used. Second, under most conditions, the user platform is also in motion. This movement provides additional, although less significant, changes in geometry. Further, if the user is airborne, there is a substantial averaging effect on the multipath seen by the airborne antenna. In point of fact, with kinematic GNSS implementations, multipath from the ground site is the single biggest contributor to error in the overall airborne system.

As a further aid to resolving the ambiguities, SVs in track by the receiver beyond the minimum five required can be used for cross-checking, thereby accelerating the ambiguity-resolution process. With six SVs, for example, two sets of four DDs can be generated. This provides a second floating baseline solution and a corresponding least-squares residual vector that can be searched. Double-difference measurements that are common between the two floating baseline solutions will produce associated integer ambiguities, which can be compared for consistency. Such redundancy usually leads to faster isolation of the proper ambiguity set.

12.3.1.8 Wide-Lane Considerations

With some receivers, it is possible to track dual- or triple-frequency GNSS signals. Use of dual-frequency techniques permits the ionospheric path delay to be precisely determined and, in some cases, eliminated. Great utility in isolating the

carrier-cycle integer ambiguities can be obtained by combining multiple frequencies to produce wide-lane observables. The creation of the wide-lane carrier phase (ϕ_{wl}) is straightforward:

$$\phi_{wl} = \phi_{Lm} - \phi_{Ln}$$

where *Lm* and *Ln* are two carrier frequencies (e.g., L1, L2, or L5) for GPS. Note that such wide-lane combinations can be formed with dual-frequency signals from any GNSS satellite [16]. As an example, the following beat frequency and wide-lane wavelength result for GPS L1 and L2:

$$f_{wl} = 1575.42 - 1227.6 = 347.82 \text{ MHz}$$
$$\lambda_{\text{wl}} = 86.19 \text{ cm}$$

When applied to searching the uncertainties of smoothed-code DD measurements, the bound of ±1 to 2m on the search volume can be spanned in theory with $\pm 3\lambda_{\text{wl}}$ instead of $\pm 11\lambda$ at, for example, GPS L1. This results in a hundredfold decrease in the number of integer-ambiguity set residuals that must be computed and examined during a given epoch. GPS IIF and newer satellites transmit the L5 signal at 1,176.45 MHz. This signal enables a wide-lane observable from the difference between L2 and L5. The resulting wavelength is 5.86m, which permits extremely rapid ambiguity searches.

The penalty for using the wide-lane wavelength is an increased noise level noise level (S_{wl}) as shown here:

$$S_{wl} = \lambda_{wl} \sqrt{\left(\frac{S_{Lm}}{\lambda_m}\right)^2 + \left(\frac{S_{Ln}}{\lambda_n}\right)^2} \tag{12.42}$$

However, current receiver technology can readily cope with this increase in noise and, assuming the magnitude of the noise level on each carrier is approximately equal, the equation reduces to, for example, 5.7 times either S_{L1} or S_{L2}, the L1 or L2 noise levels, respectively, for GPS. Considering the increase in noise that will tend to expand the search volume, in practice it may be become necessary to search beyond $\pm 3\lambda_{\text{wl}}$.

Just as there exists a combination (the difference) of carrier-phase observables that yields a wide-lane observable, there exists an alternative combination (the sum) that yields a narrow-lane observable. It can be shown that frequency-independent errors (e.g., clock, troposphere, and ephemeris errors) are unchanged in either the wide-lane or narrow-lane observations from their original values [11]. Such is not the case with frequency-dependent effects (e.g., ionospheric, multipath, and noise effects), so wide-lane carrier phase observables must be paired with narrow-lane pseudorange observables to realize the same frequency-dependent effects. A detailed explanation can be found in [27]. The narrow-lane pseudorange relationship (P_{nl}) is presented without further elaboration:

$$P_{nl} = \frac{f_{Lm} \cdot P_{Lm} + f_{Ln} \cdot P_{Ln}}{f_{Lm} + f_{Ln}} \tag{12.43}$$

There is no change in the formation of either the carrier-phase or the pseudorange (code) DDs once the wide-lane carrier phase and narrow-lane pseudorange observables are formed, and the methodology previously described in terms of the L1 carrier and code measurements is directly applicable. The prime advantage accrues from the fact that the search volume can be canvassed far more efficiently since fewer wide-lane wavelengths need to be searched. As mentioned earlier, to search the same $\pm 11\lambda$ at L1 could be done, in theory, with $\pm 3\lambda_{wl}$. In terms of **N**, the iterations would range from [–3 –3 –3 –3] to [+3 +3 +3 +3]. The integers in the ambiguity sets that result from the search represent a greater physical span, but other than that, the procedure for isolating the proper set of carrier-cycle integer ambiguity values is unchanged.

Once the proper wide-lane integer ambiguity set is determined, it is most advantageous to revert to single-frequency tracking: the signal strength of L1 C/A code is 3 to 6 dB greater than that of the P(Y)-code on L2 and there is an almost sixfold reduction in noise when using single-frequency observables over their dual-frequency counterparts. In essence, such a move significantly improves system robustness. While the transformation is quite straightforward, it is not without pitfalls. A close look at the formation of the wide-lane carrier phase DD shows the following:

$$\mathbf{DD}_{cp_{wl}} = \mathbf{DD}_{cp_{l1}} - \mathbf{DD}_{cp_{l2}} \tag{12.44}$$

This being the case, the integer ambiguity set for L1 can be determined by expanding and rearranging (12.44) as shown here:

$$\mathbf{DD}_{cp_{l1}} - \mathbf{N}_{l1}\lambda_{l1} = \mathbf{Hb} = \mathbf{DD}_{cp_{wl}} - \mathbf{N}_{wl}\lambda_{wl} \tag{12.45}$$

Combining (12.44) and (12.45) allows the recovery of the L1 integer ambiguity set:

$$\mathbf{N}_{l1} = \frac{\mathbf{N}_{wl}\lambda_{wl} - \mathbf{DD}_{cp_{l2}}}{\lambda_{l1}} \tag{12.46}$$

Care must be taken at this point since the calculation of the L1 ambiguity set will occasionally be incorrect. Referring to (12.14), it is shown that the carrier phase DD also contains an amount of noise; ultimately, this noise is swept into the resolved ambiguities. An intuitive glance at (12.46) leads to the conclusion that conversion to the L1 ambiguity set seldom if ever produces integer values. Generally speaking, the results are very close to integers, and the proper set can usually be realized by picking the nearest integer values. Occasionally, however, there is enough noise on one or more of the wide-lane measurements to cause the next

higher or lower integer ambiguity value to emerge from the conversion process. Reference [12] uses wide-lane techniques with subsequent conversion to the GPS L1 wavelength for ambiguity resolution and points out that a phase error as small as 1.2 cm can produce a conversion error of 9.51 cm ($\lambda/2$ at L1), which results in the selection of the wrong ambiguity if the nearest integer is chosen. The conclusion is that while reversion to single-frequency tracking adds robustness, the conversion process must be done with care. The L1 integer-ambiguity values that are generated by rounding the results from (12.46) must be near integer values to begin with or the operation potentially becomes suspect. One approach to solving this problem would be to follow (12.46) with a limited search around the L1 ambiguities.

12.3.2 Static Application

While land surveying is probably the most common of static applications, there are many other surveying and nonsurveying applications that take advantage of relative techniques. Among these could be counted precise dredging requirements for harbors and inland waterways, accurate leveling of land for highway construction and agricultural needs (especially land under irrigation), trackage surveys done to exacting standards for high speed rail service, and a whole host of others. Generally, the driving factor in low-dynamic applications is the necessity for centimeter-level accuracy. For land surveying, requirements for accuracies in the millimeter regime in three dimensions are not uncommon. The classical approach, used initially in [7] when only GPS was operational, demanded occupation times of up to several hours with simultaneous collection of GPS pseudorange and phase data at both ends of a prescribed baseline. This classic paper reported "analyses of data from different observation periods yielded baseline determinations consistent within less than 1 cm in all vector components." That was in December 1980, the baseline was 92.07m, and the occupation time was a minimum of 1 hour. The survey data was processed after the fact, as remains typical today. The requirement for the occupation time of at least 1 hour was driven by the need to have sufficient movement in the GPS satellite constellation to allow the carrier-cycle integer ambiguities to be resolved. Another key consideration was the overall lack of GPS satellites, which eliminated the use of redundant measurements for resolving the carrier cycle ambiguities. In this pioneering work, the ambiguity function method was used for determining the integer-cycle ambiguities.

Once the level of accuracy using first GPS then GNSS relative techniques was established, it became a natural desire to improve the efficiency of their application. The technique of kinematic surveying came into being as a result. Here, through use of a known survey point and an existing baseline, the carrier-cycle ambiguities are first determined. One technique that can be used to do this quickly is an antenna swap wherein GNSS data is collected for several minutes at each end of the baseline, the receivers/antennas are then exchanged without losing the SV lock, and another period of GNSS data is collected. Several minutes of GNSS data, with epoch times on the order of 10 seconds, are required during each occupation period to collect sufficient data to resolve the ambiguities. Four (and preferably more) SVs yielding improved satellite geometry are required to accomplish this. Subsequent

to the antenna swap, one receiver/antenna is moved to each of the points making up the survey. Generally, the receiver/antenna at the known survey point becomes the control point (base station) for the survey and the other becomes the rover. Following a 1- to 2-minute occupation of each survey point, the rover is returned to its initial starting location to provide data for closure of the overall survey. In all instances, it remains necessary to have continuous track on a minimum of the same four (but preferably more) GNSS satellites. The GNSS data are postprocessed and the survey results are calculated. For baselines of up to 10 km, the effects of the ionosphere are minimal and centimeter-level accuracies can be expected. There are variations on the static and kinematic surveying methods, but generally the resulting accuracies remain at or near the centimeter level. Further, it is the kinematic method that allows for the extension of GNSS relative techniques to the near-static or low-dynamic environment mentioned previously.

Nowadays, with complete GPS and GLONASS constellations and the availability of low-noise receivers that can track multiple frequencies from multiple constellations, it has become possible to resolve the carrier-cycle ambiguities without the need of either the presurveyed baseline or an initial period of GNSS data collection (e.g., the antenna swap procedure). The term applied to this technique is RTK. Implicit in this approach is differential, carrier phase integer-cycle ambiguity resolution. As a rule, the base station broadcasts either differential corrections or raw measurement data over a datalink and the rover computes its position relative to the base station by combining its own measurements with the information received over the datalink. Such an implementation reduces the dependence on postprocessing and permits the user to know immediately whether the survey is progressing in a successful manner. In most instances, the base station is located at a precisely known surveyed point; thus, the rover can determine its absolute position (i.e., latitude, longitude, and elevation), since it has calculated the baseline vector between it and the base station. Accuracies on the order of a few centimeters are achievable in real time with the rover in motion. Network RTK is the extension of RTK over a region, where a set of reference stations is used together to compute additional corrections for satellite orbits and atmospheric refraction.

12.3.3 Airborne Application

Flight reference systems (FRSs) using carrier phase or interferometric GNSS (IGNSS) techniques have been implemented and flight tested on transport-category aircraft. The underlying principle of operation is similar to that used for kinematic surveying and is also referred to as differential carrier-phase tracking. Figure 12.11 depicts such a system where differential techniques are employed. In this case, raw observables from all SVs in view are transmitted from a ground subsystem via datalink. The carrier- cycle ambiguity resolution is done OTF aboard the aircraft. Onboard the aircraft, position relative to the runway touchdown point is calculated in near real time and provided to the aircraft autoland system [13].

The objectives of an IGNSS FRS include such things as 0.1-m accuracy RMS (each axis), one or more updates per second, UTC time synchronization better than 0.1-ms real time, all-weather operation, and repeatable flight paths. The latter

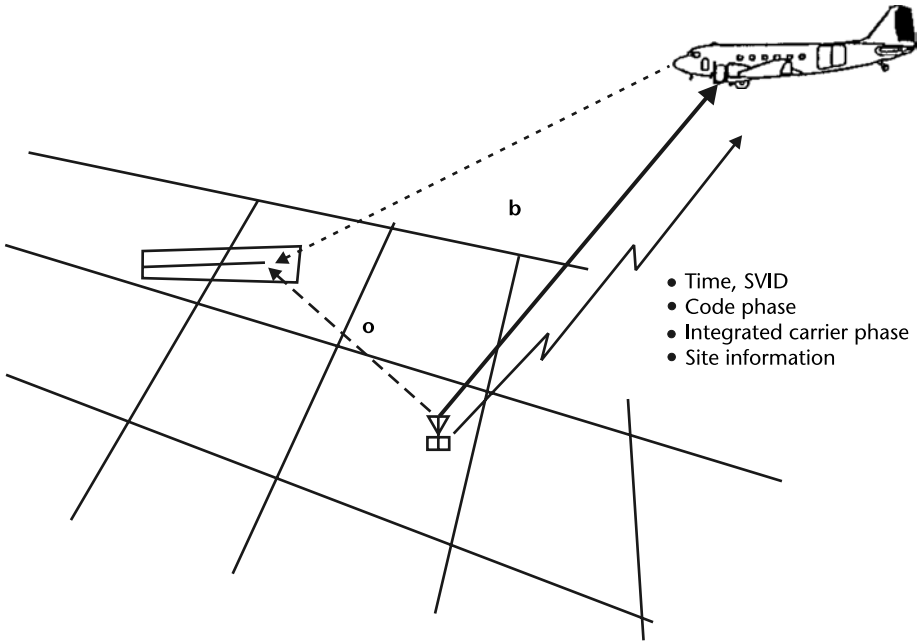


Figure 12.11 Interferometric GNSS flight reference system.

requirement calls for full aircraft integration and coupled flight. Specific applications for such systems include evaluation of approach/landing systems [e.g., instrument landing system (ILS)], and test range instrumentation calibration (e.g., all manner of tracking systems: laser, infrared, optical and radar).

Precision instrumentation of test ranges themselves can be accomplished with special GNSS receiver/datalink equipment aboard aircraft using the test range. Such a system would perform the ambiguity resolution at an appropriate ground site for each vehicle using the test range and provide position data to designated test range tracking facilities. In addition, feasibility studies in the areas of precision landing/autoland, low-visibility surface operations (taxiing, docking), high-speed turn-off, parallel runway operations, input to electronic charts, and four-dimensional navigation can also be supported.

12.3.3.1 Stand-Alone Ambiguity Resolution

Using the approach outlined in Section 12.3.1, approximately 2 to 5 minutes are required to resolve the carrier-phase integer-cycle ambiguities. The time required depends upon a sufficiency of satellites, six or more are generally needed. Good satellite geometry is also beneficial, as is motion of the airborne platform (though not necessarily required). The latter supplements the normal motion of the GNSS constellations and reduces carrier-phase and code multipath. Motion of the constellation is vital to the resolution of the carrier-cycle ambiguities. As the various candidate ambiguity sets are identified and evaluated over time, there exists only one set that can persist given the dynamics of the constellation and, to a lesser degree, the added motion of the platform. Simply stated, without motion, the technique presented would not work.

12.3.3.2 Pseudolite Ambiguity Resolution

During the initial feasibility studies for the FAA Local Area Augmentation System (LAAS), integrity beacons (a form of pseudolite), were used for rapid GPS carrier-cycle integer-ambiguity resolution. These devices were low-power transmitters, two of which were placed within several miles of a runway threshold along the nominal approach path. These transmitters operated at L1 and were modulated with unused PRN codes such that they would not be mistaken for an SV. The several minutes of time required to resolve the carrier-cycle ambiguities as described above were reduced to seconds with this method owing to the rapid change in geometry as the aircraft passes through the signal bubble created above the pseudolites. A second GPS antenna mounted on the belly of the aircraft was used to acquire the pseudolite signals. The presence of the two pseudolites also reduced the requirement of visible SVs to four and ensured that as the aircraft exits the bubble, the carrier-phase integer-cycle ambiguities are resolved. Centimeter-level positioning accuracy was thus ensured from this point to touchdown and rollout. Both the real-time cycle ambiguity resolution and the centimeter-level positioning accuracy were demonstrated in flight testing with transport category aircraft [14, 15]. Pseudolites have also been investigated as a means of improving local GPS satellite availability [28]. However, at the present time, pseudolites are no longer planned for use in civil aviation DGNSS systems for reasons including cost, spectrum regulatory concerns, and the emergence of additional GNSS core constellations.

12.3.3.3 Accuracy

Once the carrier-phase ambiguities are resolved, the accuracy of the DD measurement is determined by the carrier-phase measurement. In this case, multipath is the dominant error source. If the reflected signal is weaker than the direct signal, the phase measurement can be in error by up to 0.25 wavelength. If the reflected signal is stronger than the direct signal, cycle slips are likely to occur. Typical wide-lane DD measurement errors are on the order of 2 to 10 cm (2σ). Due to geometry, vertical positioning errors are between 1.5 to 2 times the DD measurement error, resulting in up to 20 cm (2σ) vertical positioning errors. Horizontal positioning errors are generally less than 20 cm (2σ). If both the ground and the airborne antennas are placed in a rich multipath environment, vertical positioning error further degrades to approximately 40 cm. However, as soon as the aircraft is in motion, airborne multipath is mitigated due to the rapid changing path length difference between the direct and reflected signals, which tends to average the multipath error.

The use of dual-frequency measurements can be very important for an IGNSS FRS in some applications since it allows for the mitigation of ionospheric errors, particularly for longer baselines. Once the ambiguities are resolved, the system could revert back to a single-frequency system. Because of the shorter wavelength of the L1 signal, multipath error would be reduced by approximately a factor of 4.5.

12.3.3.4 Carrier Cycle Slips

The carrier-phase observable must be tracked continuously by the receiver or the agreement between the fixed and floating baseline solutions will diverge rapidly.

Loss of signal can occur due to the setting of a satellite, excessive maneuvering of the user (a large bank angle in the case of an airborne user during approach or takeoff), or an obstructed view of the sky in the direction of the satellite. In any event, a loss of signal continuity, no matter how brief, results in an unknown signal loss or gain of carrier cycles when the signal is reacquired by the receiver. In a kinematic environment, detection of cycle slips is vital, since allowing corrupted carrier phase measurements to propagate forward usually causes immediate loss of the fixed solution. As such, identification of the cycle slip becomes paramount and, rather than attempt repair, the offending SV is ignored for a predetermined number of epochs with the assumption that the signal will quickly return to normal. At the conclusion of this time-out period, the data from the offending SV is once again accepted, and the carrier-cycle integer ambiguity resolution process restarted for the SV. In the interim, if a minimum of four SVs (not including the offending SV) had their ambiguities resolved at the time of cycle-slip detection, the fixed baseline solution is maintained. Otherwise, at best only a floating baseline solution can be provided.

12.3.4 Attitude Determination

An additional application of the interferometric techniques described earlier in this section is *attitude determination*. If antennas are placed on a rigid body, such as an aircraft, then the baseline vectors between each pair of antennas are known quantities within the body-frame coordinate frame defined as shown in Figure 12.12. The x-axis extends through the nose of the vehicle, the z-axis points downward, and the y-axis is mutually perpendicular to the x- and z-axes to form a right-handed coordinate system (e.g., through the right wing as viewed by the pilot for an aircraft). Typically, the nominal center of mass of the platform is chosen as the origin.

If carrier-phase measurements are taken from each of the antennas, the integer ambiguities may be resolved, as discussed above, to determine the baseline vectors within the local north, east, down (NED) coordinate frame. This may be mechanized by first solving for these quantities in an ECEF coordinate frame (e.g., WGS-84) and then applying the appropriate transformation (see Chapter 2). At any given time, the relationship between the coordinates of three antennas expressed in the body frame (known from the installation) and expressed in the NED frame (computed from GNSS carrier-phase measurements) may be written as [17]:

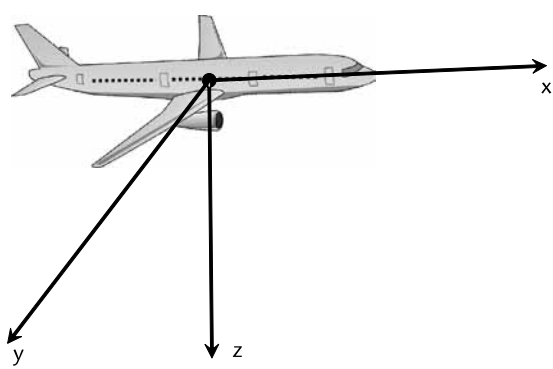


Figure 12.12 Body-frame coordinate system

$$\mathbf{R}_{ned} = \mathbf{T}\mathbf{R}_{body} \tag{12.47}$$

where $\mathbf{R}_{ned}$ is the matrix of antenna coordinates in the NED frame, $\mathbf{R}_{body}$ is the matrix of antenna coordinates in the body frame, and $\mathbf{T}$ is the 3×3 transformation matrix:

$$\mathbf{T} = \begin{bmatrix} \cos\psi\cos\theta & -\sin\psi\cos\phi + \cos\psi\sin\theta\sin\phi & \sin\psi\sin\phi + \cos\psi\sin\theta\cos\phi \\ \sin\psi\cos\theta & \cos\psi\cos\phi + \sin\psi\sin\theta\sin\phi & -\cos\psi\sin\phi + \sin\psi\sin\theta\cos\phi \\ -\sin\theta & \cos\theta\sin\phi & \cos\theta\cos\phi \end{bmatrix} \tag{12.48}$$

The desired end-quantities are the *Euler angles* (ψ, θ, ϕ) that represent *heading*, *pitch*, and *roll* (more formally, heading, elevation, and bank angle [29]), respectively. Following [17], the Euler angles may be found by first determining a least-squares estimate of $\mathbf{T}$:

$$\mathbf{T} = \mathbf{R}_{ned}\mathbf{R}_{body}{}^{T}\left(\mathbf{R}_{body}\mathbf{R}_{ned}{}^{T}\right)^{-1} \tag{12.49}$$

followed by a solution of (12.48):

$$\theta = \sin^{-1}\left(-T_{31}\right); \phi = \sin^{-1}\left(\frac{T_{32}}{\cos\theta}\right); \psi = \sin^{-1}\left(\frac{T_{21}}{\cos\theta}\right) \tag{12.50}$$

where T_{ij} refers to the (i,j)th element of the matrix $\mathbf{T}$.

From an inspection of (12.50), it is apparent that this approach is not stable for pitch angles approaching ±90°. Nonetheless, the method outlined above is practical for a number of applications where such an attitude is not encountered. The preferred mechanization for platforms that may experience any attitude is the use of *quaternions*. A quaternion is a mathematical construct that essentially extends the notion of a complex number to four dimensions. Whereas a complex number can be viewed as a mapping from a 2-vector (a,b) to a complex number $a + ib$, a 4-vector (a,b,c,d) can be mapped into a quaternion $a + \mathrm{i}b + \mathrm{j}c + \mathrm{k}d$ (referred to as a pure quaternion if $a = 0$), which has its own associated set of mathematical rules. An excellent introduction to quaternions is provided in [29].

GNSS attitude determination systems are often implemented with four antennas or more, even though all three Euler angles can be determined with only three. Additional antennas provide redundancy, and are especially important for all-attitude platforms that can rotate such that they block visibility of one or more of the antennas to the visible GNSS satellites. Whereas four satellites are normally required for carrier-phase positioning, only two satellites are needed for attitude determination provided that a common receiver is utilized for phase measurements for each antenna and that the baseline lengths between the antennas are precisely known [30]. The common receiver results in cancellation of the receiver clock bias when single differences of carrier phase measurements between antennas are

formed. Although each antenna has a different analog path to the receiver, which results from differing electrical path lengths, these *line biases* can be mostly removed through calibration procedures.

Multipath is the error source that limits performance for most GNSS attitude determination systems. Typical one-sigma accuracy for each Euler angle in radians is the one-sigma single-difference carrier phase multipath error divided by the antenna baseline length (with both the one-sigma multipath error and the baseline length in the same units of length) [30]. Additional error sources that may be significant for GNSS attitude determination applications that have not been previously discussed include structural flexing and tropospheric refraction. Structural flexing is the bending of the platform on which the multiple GNSS antennas are installed due to applied forces or temperature changes. If flexing is non-negligible, its effects can be mitigated through estimation or modeling. Tropospheric refraction is the bending of GNSS signals as they pass through the troposphere. The very slight bending of each GNSS signal path does not significantly alter pseudorange and carrier-phase measurements but may introduce unacceptable Euler angle errors for some applications. Tropospheric refraction effects can be mitigated through modeling (e.g., the use of Snell's law in conjunction with a slab model for the troposphere [30]). More comprehensive treatments of GNSS attitude determination concepts may be found in [30, 31].

12.4 Precise Point Positioning

The GNSS processing technique that has come to be known as *precise point positioning* (PPP) has quickly grown in use over the past decade and a half, and its various forms can be categorized as point positioning and differential positioning. PPP's popularity comes from the fact that nearby reference stations are not required for processing, and the technique gives the appearance that only a single (geodetic) receiver is needed. The reality is that PPP service providers use widespread (often global) networks of reference receivers to produce accurate estimates of GNSS satellite orbits and clock errors. A PPP ground network is thus similar in form and function to a WADGNSS ground network (see Section 12.2.3). Strictly speaking, however, PPP is not differential since in most implementations the ground network's clock and ephemeris estimates are supplied directly to the end user to be substituted for the broadcast data from each GNSS satellite (as opposed to being in the form of differential corrections to the broadcast data as in WADGNSS).

In its conventional (and original form), PPP consists of point positioning (see Chapter 11), but using precise satellite orbits and clocks rather than satellite broadcast corrections, additional error modeling (which will be discussed), and sequential filtering (e.g., via least-squares or Kalman filter approaches) of available dual-frequency pseudorange and carrier-phase measurements. The idea of using improved ephemerides in satellite navigation can be traced back to TRANSIT in 1976 [32], initially for pseudorange processing in GPS in 1995 [33], and for what is now conventional GPS PPP in 1997 [34]. In this section, the fundamental concepts of conventional PPP are described, its performance and utility, followed by recent advances in the technology and future prospects.

12.4.1 Conventional PPP

The idea behind the original PPP technique is to make use of dual-frequency GNSS pseudorange and carrier-phase measurements in a functional model where precise (centimeter-level) orbits and clocks, for example, from the International GNSS Service (see Section 12.6.2.2) replace broadcast (meter-level) products; additional error modeling is incorporated to allow for decimeter-to-subcentimeter positioning accuracy; and sequential measurement filtering is used to increase positional accuracy over time.

Mathematical Model

The fundamental PPP functional model parameterization is given in (12.51), in which four dual-frequency GNSS pseudorange and carrier-phase observables are combined into ionospheric-free linear combinations (see Section 10.2.4.1.)

$$\begin{aligned} P^s_{r_{IF}}(t) &= \rho(t) + c\left[d\tau^s(t) - d\tau_r(t)\right] + \delta_{tropo}(t) + mp_{P_{IF}}(t) + e_{P_{IF}}(t) \\ \Phi^s_{r_{IF}}(t) &= \rho(t) + c\left[d\tau^s(t) - d\tau_r(t)\right] + \delta_{tropo}(t) - \lambda_{IF}N_{IF} + mp_{\Phi_{IF}}(t) + e_{\Phi_{IF}}(t) \end{aligned} \tag{12.51}$$

Note that in this formulation, the ionosphere-free carrier-phase ambiguity N_{IF} is not an integer, as satellite and receiver equipment delays (also referred to fractional phase biases) are not removed as is the case with double-differencing. Also note that the magnitude of this pseudorange noise is approximately 100 times more than that on the carrier phase, which results in the characteristic positioning performance of PPP that will be shown later in this section. This model effectively removes the effects of ionospheric refraction on the measurements, while retaining the geometric range, clock terms, and zenith tropospheric refraction in the parameterization with only a small (centimeter-level) increase in propagated measurement error due to the linear combination.

The effect of applying this parameterization is for the meter-level precision pseudorange measurements to provide initial position accuracy, while the centimeter-level precision carrier-phase measurements refine this position estimation over time as receiver position, receiver clock error, tropospheric refraction and float phase ambiguities are more accurately estimated. The necessary estimation process can be found in, for example, [35], in this case using a sequential least-squares filter. The result is an initial code-like submeter-level positioning performance, followed by decimeter- to subcentimeter-level positioning over hours of continuous processing, that is, RTK-like performance, but without the need for a nearby reference station.

Error Modeling

In order to obtain these levels of positioning results, the user equipment must account for a number of error sources that are negligible for most other stand-alone or differential GPS applications. These error sources include:

- *Satellite antenna lever-arm:* Most often, in orbit determination, it is the location of the satellite center of mass that is estimated not the satellite's antenna

phase center. The satellite antenna lever arm is the vector difference between these two locations.

- *Phase wind-up:* Relative rotation between a GNSS satellite and the user antenna can cause carrier-phase measurements to change by up to one cycle. This effect is referred to as phase wind-up. A correction for this effect is provided in [36].
- *Solid earth tides and ocean loading:* The Earth's surface is not rigid, but rather somewhat pliable. Its shape varies with time, dominated by diurnal and semidiurnal components, in response to gravitational forces. These Earth surface movements are referred to as solid Earth tides. Additional motion of the Earth's surface, especially in coastal locations due to ocean tides, is referred to as ocean loading. Solid Earth tides and ocean-loading site displacements can be as large as 30 cm and a few centimeters, respectively. By convention, ECEF coordinate systems such as ITRF are explicitly defined to not include solid Earth tides and ocean-loading effects. Thus, these effects should be removed for applications where the user position in ECEF coordinates are desired. Accurate models for the Earth's deformation due to solid earth tides and ocean loading can be found in [37].

Performance Characteristics and Use

The result of this form of measurement processing is a characteristic initial convergence period, followed by steady state position accuracy as can be seen in Figure 12.13 for 24 hours of static, dual-frequency GPS data. Note in the inset that it takes

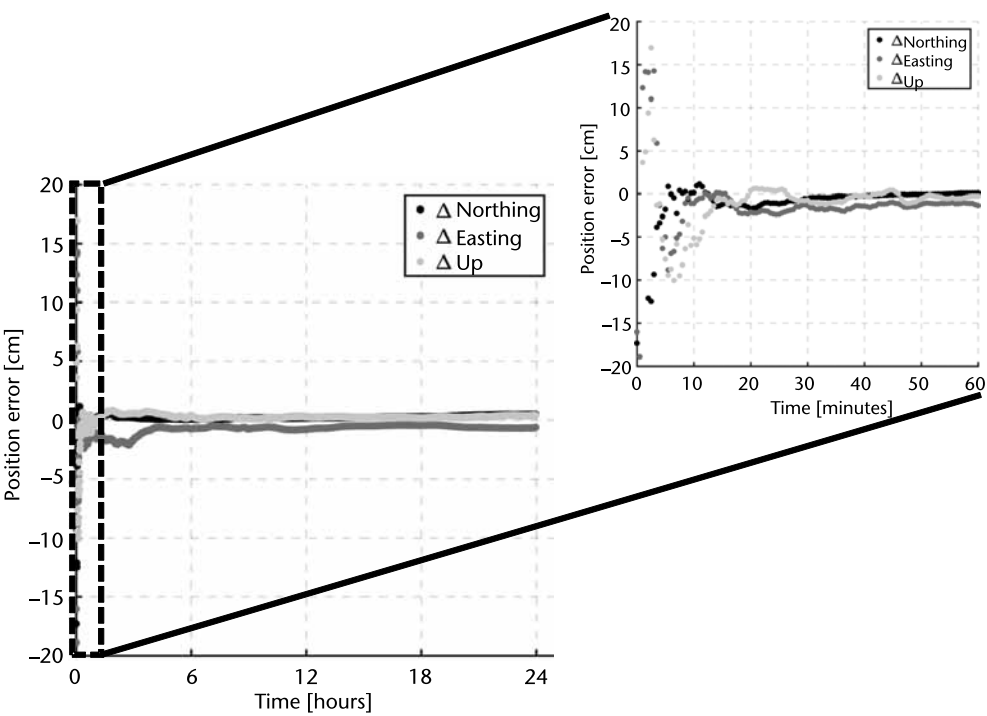


Figure 12.13 GPS PPP performance characteristics of initial convergence period and steady state.

approximately 15 to 20 minutes from a cold start for the position solution to converge to better than a few centimeters error in the north, east, and up components. Initial convergence tends to be defined for PPP applications as the time necessary to obtain a specified level of horizontal, vertical, or three-dimensional positioning accuracy.

As the conventional PPP technique only utilizes user measurements, PPP is more susceptible to the quality of available measurements that is baseline processing. With additional measurements (e.g., from GLONASS satellites) and given the appropriate spatial and temporal datum modeling between GPS and GLONASS and GLONASS equipment bias modeling, the initial convergence period can be reduced by minutes from a GPS-only solution, as is illustrated in Figure 12.14. Note that while the addition of GLONASS reduced initial convergence period, there is little or no improvement in steady-state performance with good sky view as adequate geometric strength is provided by GPS.

Given PPP's initial convergence period and its consistent performance worldwide without baseline limitations, the technique has become a dominant GNSS technology for precision positioning and navigation in remote areas or regions of low economic density, where LADGNSS use is limited or prevented [38]. The caveat is that continuous satellite tracking in needed. Commercial uses include offshore positioning, precision agriculture, geodetic surveying, airborne mapping; and scientific applications include plate tectonics, seismic monitoring, tsunami warning, and precise orbit determination.

12.4.2 PPP with Ambiguity Resolution

As was the case in the development of DGNSS processing, the question was raised as to if carrier-phase integer ambiguities in conventional or float PPP mode could be fixed, and if a similar level of performance (in terms of initial convergence,

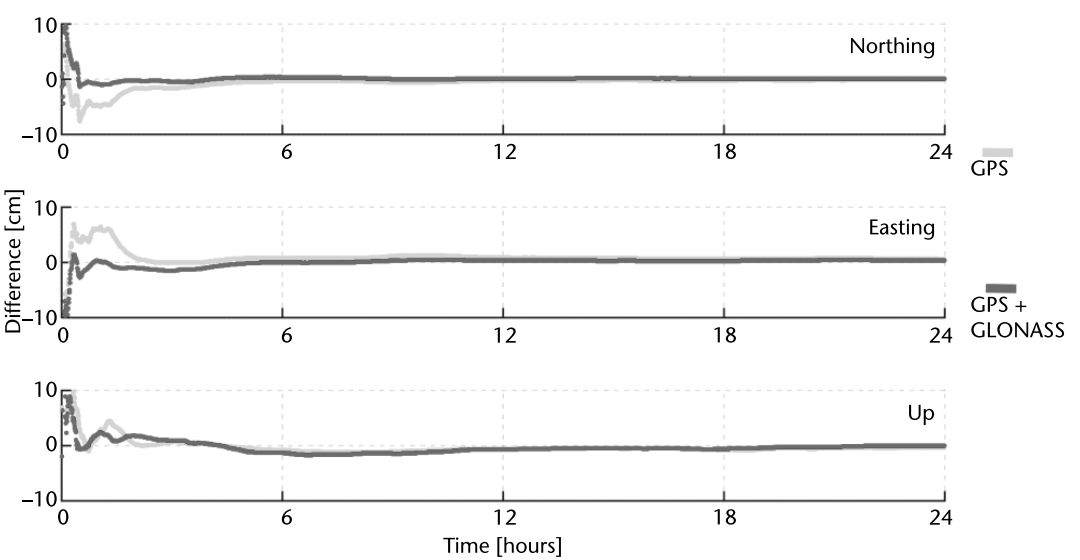


Figure 12.14 Reduced initial convergence period with GPS + GLONASS PPP.

precision, and accuracy) could be attained. The answer is yes, but the process by which to obtain fixed PPP solutions requires a hybrid solution combining PPP with some elements of carrier-based DGNSS processing.

Equipment Delays and Corrections

The challenge with using undifferenced PPP measurements for resolving carrier-phase ambiguities is that, referring to (12.51), the satellite and receiver equipment delays are not eliminated as is the case in DGNSS double-differencing (see Section 12.3.1.3). Equation (12.51) can be expanded to explicitly parameterize these equipment delays as:

$$\begin{aligned} P^s_{r_{IF}}(t) &= \rho(t) + c\left[d\tau^s(t) - d\tau_r(t)\right] + \delta_{tropo}(t) + b^s_{IF}(t) + b_{r_{IF}}(t) + e_{IF}(t) \\ \Phi^s_{r_{IF}}(t) &= \rho(t) + c\left[d\tau^s(t) - d\tau_r(t)\right] + \delta_{tropo}(t) + \lambda_{IF} b^s_{IF} + \lambda_{IF} b_{r_{IF}}(t) - \lambda_{IF} N_{IF} \\ &+ \varepsilon_{IF}(t) \end{aligned} \tag{12.52}$$

where b^s_{IF} and b_{rIF} are the ionospheric-free combination of the satellite and receiver equipment delays, respectively. In the literature, these terms are also referred to as hardware delays, or hardware biases, and for the phase terms, fractional phase biases. Note that for the latter, the phase ambiguity is now an integer prior to the linear combination. There is not enough resolving power to estimate all of the terms in this functional model as was the case in the float processing, where the equipment biases are absorbed into the measurement error, receiver clock and phase ambiguity terms. So an alternate approach must be taken.

References [39–42] describe some of the earliest methods to isolate satellite equipment delays and eliminate receiver equipment delays. Typically, the functional model is expanded to include the Melbourne-Wübbena linear combination [43, 44] to estimate wide-lane ambiguities and satellite and receiver equipment delays by differencing the narrow-lane pseudorange from the wide-lane carrier-phase, as this combination is uncorrelated from the other parameters in (12.52). A relatively sparse global network of GNSS reference stations can be used to estimate the satellite equipment delays, usually as part of the same network that is used to estimate satellite orbits and clocks. These satellite equipment delays can be broadcast to the user along with the traditional PPP precise orbit and clock products to allow user receivers to remove the satellite equipment delays, eliminate the receiver equipment delays through between satellite single differences, and therefore estimate and fix just the integer carrier-phase ambiguity terms. The result still gives the appearance of stand-alone positioning, as there is no baseline limitation/restriction. Figure 12.15 provides an example of the quality of float and fixed GPS PPP positioning, noting that fixing begins at hour 1 to conservatively avoid incorrect fixes. While ambiguity-resolved PPP (referred to as PPP-AR or occasionally as PPP-RTK or RTK-PPP) improves position solution precision and accuracy and reduces initial solution convergence period, quick convergence baseline RTK-like performance is still not attainable.

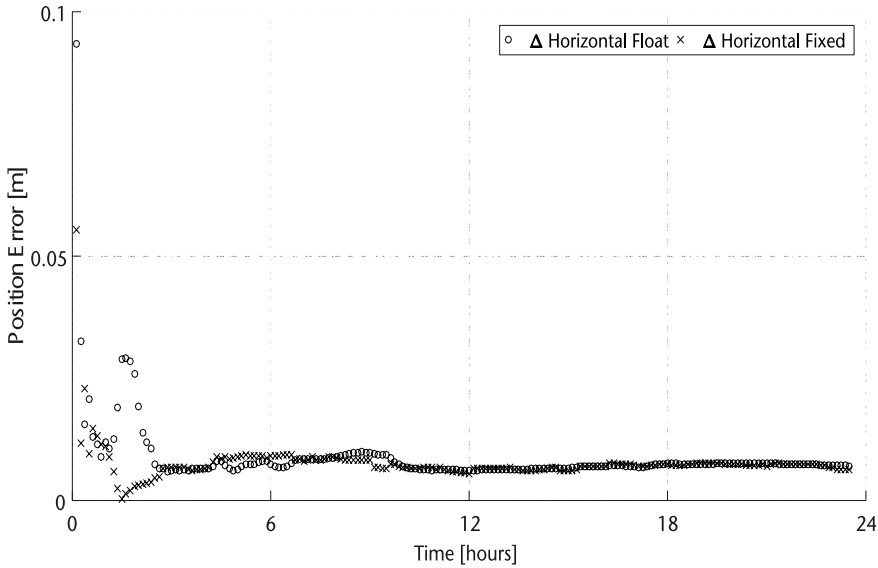


Figure 12.15 GPS PPP and PPP-AR positioning performance.

Ionospheric Modeling for Rapid Reconvergence

Another limitation posed by PPP is that in sky obstructed areas, significant measurement gaps result in estimation filter reinitialization, making PPP (unlike RTK with its rapid ambiguity fixing capability) unsuitable for many urban environments. Reference [45] showed that by estimating slant ionospheric delay, rather than eliminating this refraction effect in the functional model [(12.52)], these slant ionospheric delay estimates can be used as bridging parameters through small spatial or temporal measurement gaps. As direct GNSS estimated slant ionospheric delays contain significant geometric information and vary slowly (e.g., over tens of meters and tens of seconds), slant ionospheric estimates and appropriately de-weighted associated covariances can be used as a priori estimates after a data gap. For example, the user receiver travels under a highway overpass or cluster of tall buildings, resulting in very quick filter estimation of other state parameters (especially position, user clock error, and float ambiguities). Also, the estimation, rather than elimination, of ionospheric parameters has allowed for a priori information from regional and global ionospheric models to be included in PPP estimation to reduce initial position solution error and therefore reduce initial convergence period, as well as the seamless integration of CORS and network RTK services into PPP user receiver processing (see, e.g., [46]). The overall result is that this innovation has made PPP much more robust in obstructed environments.

Performance and Future Prospects

Given that PPP relies so heavily on the number, quality and geometry of the user receiver measurements (as opposed to DGNSS approaches making full use of observable double-differencing), rapid development and upgrade of global and regional navigation systems (GPS, GLONASS, BeiDou, Galileo, QZSS, and NAVIC) have led to noticeable improvements in initial positioning filter convergence period, and solution precision and accuracy under open-sky and slightly obstructed

conditions. However, significant challenges exist in processing signals from a single constellation and from multiple constellations. These well-known issues include between-code biases, between-phase biases, between-code and between-phase biases, between-frequency biases, and between-constellation spatial and temporal datum biases. All such biases, as well as appropriate stochastic modeling for all measurements, must be determined for quality multiconstellation PPP. It must be noted that at some point, the law of diminishing returns takes effect and additional signals only marginally improve the solution. For example, Figure 12.16 illustrates the effect of the accumulation of GNSS measurements by constellation in dual-frequency, float PPP positioning. Note that for this example from Japan, signals from four Galileo satellites were available at the start of processing. After significant positioning improvement to the GPS solution is made by processing GPS+GLONASS measurements, little advantage is gained by adding Galileo or BeiDou in this example.

Simulations of future PPP performance have tended to be optimistic, and no single innovation (e.g., PPP-AR) has produced quick, minute-level initial convergence. However, academic, institutional and commercial PPP research and development activities are growing at a significant pace. More and improved correction products are available; improved protocols and delivery mechanisms have been or are being developed; and, more online and commercial services are available (see Section 12.6.3). The introduction of ambiguity resolution, ionospheric estimation and modeling, multiconstellation processing, and triple-frequency processing have all contributed to reducing initial filter convergence period from many tens of minutes to approximately 10 to 15 minutes. As the technology improves in the coming decade, this initial convergence issue may be reduced to the point where PPP would be practical in all service areas.

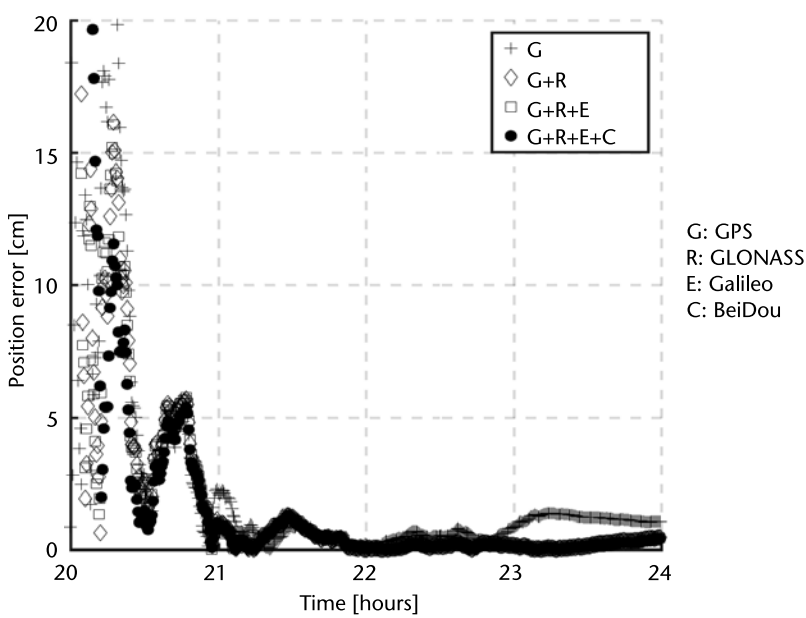


Figure 12.16 Dual-frequency float PPP positioning error from accumulated GNSS measurements by constellation.

12.5 RTCM SC-104 Message Formats

Many messaging protocols have been developed throughout the industry for the dissemination of code- and carrier-based DGNSS data between reference stations and users and for PPP data. This section will present some widely-used DGNSS and PPP message standards developed by the Radio Technical Commission for Maritime Services (RTCM) Study Committee 104 (SC-104). Although originally developed for maritime applications, RTCM SC-104 messages are now supported by the vast majority of commercial GNSS receivers, including low-cost recreational devices.

From the 1980s to the mid-2000s, there was only one set of SC-104 messages, referred to since 1990 as Version 2, to support both code- and carrier-based local-area DGNSS services. This message set evolved over time with Version 2.3 with Amendment 1 published in May 2010 [47] being the most recent version. In February 2004, RTCM published guidelines for a new set of messages that use a more efficient protocol referred to as Version 3. This later protocol, now up to Version 3.3 [48], provides message formats suitable for code and carrier-phase DGNSS as well as for PPP. Both protocols (Versions 2.3 and 3.3) describe digital message formats that can be broadcast from a reference station or network of reference stations to a user using any arbitrary data link. The Version 2.3 and 3.3 standards are described in Sections 12.5.1 and 12.5.2, respectively.

As mobile packet-switched cellular networks proliferated around the world, delivery of DGNSS and PPP data through the Internet Protocol (IP) has become increasingly popular. Section 12.5.3 describes one last widely used RTCM standard, which is for the Networked Transport of RTCM via IP (NTRIP).

12.5.1 Version 2.3

Figure 12.17 shows the basic frame format of Version 2.3, which consists of a variable number of 30-bit words. The last 6 bits in every word are parity, and the 30-bit word format is derived from the GPS navigation message. The first two words of each frame are referred to as the *header*. The content of the header is shown in Figure 12.18. The first word of the header contains an 8-bit preamble, consisting of the fixed sequence 0110110, followed by the 6-bit *Frame ID*, which identifies one of 64 possible message types (see Table 12.2). Next, a 10-bit *Station ID* identifies the reference station. The first 13 bits of the second word in the header, the *Modified Z-count*, comprise the time reference for the message. The following three bits form the *Sequence Number*, which increments on each frame and is used to verify frame synchronization. The frame length is needed to identify the beginning of the next frame, since the length of the frame is variable, depending on the message type

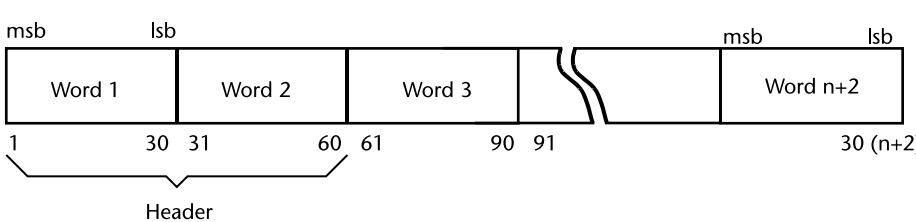


Figure 12.17 RTCM SC-104 Version 2.3 message frame.

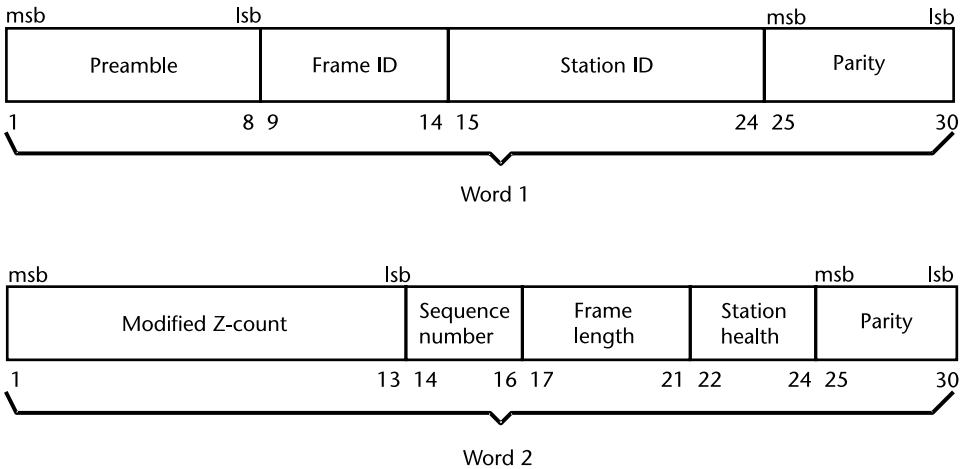


Figure 12.18 RTCM SC-104 Version 2.3 Message Header.

and the number of visible satellites. The 3-bit *Station Health* indicates if the reference station is not functioning properly, or if the reference station transmissions are unmonitored. Six of the possible eight patterns of the 3-bit Station Health are used to provide a scale factor for a field that appears in various message types referred to as *User Differential Range Error* (UDRE) that will be described later.

For code-based DGNSS systems that only supply corrections for the GPS C/A-code signal, Message Types 1 and 9 are among the most important messages. The content of Message Type 1 is shown in Figure 12.19 (note that the two-word header that is appended at the beginning of every message type is not explicitly shown). For every visible satellite, the Type 1 message includes the following parameters:

- *Scale factor:* 1 bit to indicate the resolution of the pseudorange and range-rate corrections to follow. If unset (set), resolutions of 0.02m (0.32m) and 0.002 m/s (0.032 m/s) apply for the pseudorange and range-rate corrections, respectively.
- *UDRE:* 2 bits that indicate ranges of expected one-sigma errors of the pseudorange corrections. As mentioned above, 6-bit patterns in the Station Health field of the header are used to provide a scale factor for UDRE. UDRE values ranging from ≤0.1m to > 8m are possible with the scale factor applied.
- *Satellite ID:* 5 bits to indicate the satellite number for which DGPS corrections are being provided.
- *Pseudorange correction:* 16-bit correction $\Delta\rho_m^i(t_0)$ for the indicated satellite, applicable at the time t_0 provided by the Z-count in the header.
- *Range-rate correction:* 8-bit rate correction $\Delta\dot{\rho}_m^i(t_0)$ (see discussion in Section 12.2.1.2).
- *Issue of data (IOD):* The IOD indicates the specific set of GPS navigation data that was used in generating the corrections. As noted in Chapter 3, approximately every 2 hours, the broadcast clock and ephemeris data from each GPS satellite is changed. The GPS navigation message tags each set of clock and ephemeris data with IOD values referred to as IODC (IOD, clock)

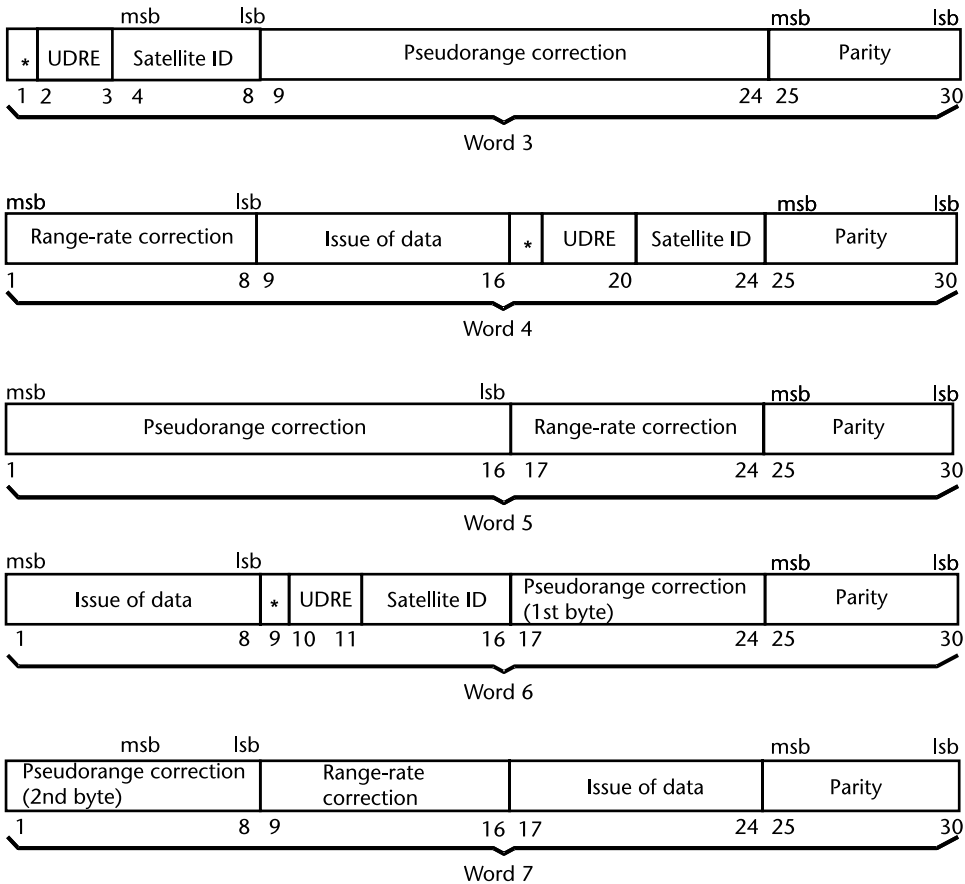
Table 12.2 RTCM SC-104 Version 2.3 Message Types

Message Type	*Status*	*Use*
1	Fixed	Differential GPS Corrections
2	Fixed	Delta Differential GPS corrections
3	Fixed	GPS Reference Station Parameters
4	Tentative	Reference Station Datum
5	Fixed	GPS Constellation Health
6	Fixed	GPS Null Frame
7	Fixed	DGPS Radiobeacon Almanac
8	Tentative	Pseudolite Almanac
9	Fixed	GPS Partial Correction Set
10	Reserved	P-code differential Corrections
11	Reserved	C/A-code L1, L2 Delta Corrections
12	Reserved	Pseudolite Station Parameters
13	Tentative	Ground Transmitter Parameters
14	Fixed	GPS Time of Week
15	Fixed	Ionosphere Delay Message
16	Fixed	GPS Special Message
17	Fixed	GPS Ephemerides
18	Fixed	RTK Uncorrected Carrier Phases
19	Fixed	RTK Uncorrected Pseudoranges
20	Fixed	RTK Carrier Phase Corrections
21	Fixed	RTK/Hi-Accuracy Pseudorange Corrections
22	Tentative	Extended Reference Station Parameters
23	Tentative	Antenna Type Definition Record
24	Tentative	Antenna Reference Point (ARP)
25–26	—	Undefined
27	Fixed	Extended Radiobeacon Almanac
28–30	—	Undefined
31–36	Tentative	GLONASS messages
37	Tentative	GNSS System Time Offset
38–58	—	Undefined
59	Fixed	Proprietary Message
60–63	Reserved	Multipurpose messages

and IODE (IOD, ephemeris). IODC is a 10-bit parameter and IODE is an 8-bit parameter (the 8 LSBs of IODC). IOD in the SC-104 messages are equal to IODE in the GPS broadcast message.

Message Type 1 repeats these fields for every visible satellite. Message Type 9 uses the same format, except that it only allows up to three satellites per message. The use of Message Type 9 requires a more stable clock in the reference station, since pseudorange corrections for all visible satellites must be broadcast with different reference times.

For carrier-phase DGNSS, Message Types 18 to 21 are used. Message Types 18 and 19 convey the reference station's raw (i.e., uncorrected with the broadcast GPS



* Scale factor

Figure 12.19 RTCM SC-104 Version 2.3 Message 1 format: Words 3 to 7.

ephemerides) carrier-phase and pseudorange measurements, respectively, so that the user can compute the double-differences described in Section 12.3. Message Types 20 and 21 are similar, but convey carrier-phase and pseudorange measurements, respectively, that have been corrected using the GPS broadcast ephemerides.

Message Types 31 to 36 provide messages for use with GLONASS. For support of Galileo, BeiDou and other GNSS constellations, most multi-GNSS systems are using SC-104 Version 3 that is described next.

12.5.2 Version 3.3

The development of the SC-104 Version 3 standard [48, 49] was initially focused on the development of a more efficient DGNSS message format to support RTK carrier-phase operations. The format is radically different from Version 2.3, in part to provide a more efficient parity scheme, designed to protect against bursts errors as well as random bit errors, and in part to overcome limitations of the Version 2.3 format, including an increased efficiency that will allow more timely broadcasts for RTK operations.

Version 3.3 messages are broadcast in variable length frames shown in Figure 12.20. Each frame begins with an 8-bit preamble, followed by 6 reserved bits and

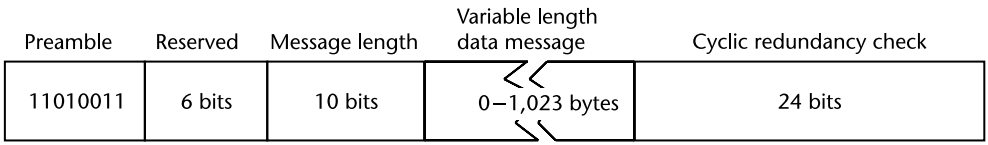


Figure 12.20 RTCM SC-104 Version 3.3 message frame.

a 10-bit message length field. The data message, ranging from 0 to 1,023 bytes, is then broadcast followed by 24 bits of parity for error detection referred to as a cyclic redundancy check (CRC) code. This message format is much more efficient than Version 2.3, which devotes more than 20% of the data link throughput to overhead (e.g., parity). Furthermore, the Version 3.3 parity scheme is much stronger than that used for Version 2.3. The first field in every data message conveys a 12-bit message number, allowing for up to 4,096 message types.

The initial release of Version 3.0 in 2004 included 13 message types, designed primarily to support RTK applications using GPS or GLONASS. These message types, known as Message Types 1001 to 1013 that are still in use within Version 3.3, provide the reference station's pseudorange and carrier-phase measurements for L1 or L1/L2, as well as a wealth of auxiliary information including precise station coordinates, receiver configuration, and antenna characteristics. Version 3.1, released in 2006, added GPS network corrections messages to provide a mobile user RTK information valid over a large area. In addition, Version 3.1 introduced new GPS and GLONASS messages with orbital parameters to assist in rapid acquisition, a text message, and a set of messages reserved for vendors to encapsulate proprietary data. Version 3.2, published in 2013, consolidated five amendments made to Version 3.1 and additionally added Multiple Signal Messages (MSM). The MSM concept [50] was designed so that SC-104 V3 could be readily applied to GNSS constellations beyond GPS and GLONASS, such as Galileo, BeiDou, and QZSS. The latest version, Version 3.3, added MSM support for satellite-based augmentation systems (SBAS) and included numerous other improvements.

Version 3.3 supports not only LADGNSS, but also WADGNSS and PPP through a set of *state-space representation* (SSR) messages. Pseudorange corrections and raw measurements (pseudorange and carrier-phase) needed for RTK are considered to be observation space representation (OSR) messages. SSR messages provide data related to, for example, GNSS satellite clocks and ephemeris. SSR messages were first added to RTCM SC-104 Version 3 in 2011 within a published amendment to Version 3.1. The SSR message set has since been expanded considerably, and is used by several operational PPP systems (see, e.g., [51]).

12.6 DGNSS and PPP Examples

12.6.1 Code-Based DGNSS

12.6.1.1 NDGPS

In the late 1980s, the United States Coast Guard (USCG) began the development of a maritime DGPS (MDGPS) system to satisfy maritime navigation requirements in the United States. In 1989, a radiobeacon located on Montauk Point, New York,

was modified to broadcast DGPS corrections in the RTCM SC-104 message format. By February 1997, 54 radio beacons had been modified to provide DGPS correction coverage for most U.S. coastal areas and inland waterways and the MDGPS service was declared to have achieved FOC. That same year, a decision was made to expand radio beacon DGPS coverage throughout the United States. This program, referred to as Nationwide DGPS (NDGPS), was supported by a partnership of U.S. agencies including the USCG, the U.S. Air Force Air Combat Command (ACC), the Federal Railroad Administration (FRA), the Federal Highway Administration (FHWA), the National Oceanic and Atmospheric Administration (NOAA), the U.S. Army Corps of Engineers (USACOE), and the Office of the Secretary of Transportation (OST) [52]. By 2005, 84 of 136 originally proposed sites were operational providing nearly complete coverage over the United States with two or more sites visible to users in many locations.

The number of operational NDGPS sites remained largely unchanged for a decade until August 2016. On August 4, 2016, due to declining use, 37 sites were shut down out the 83 that were operational at that point in time. Most of the decommissioned sites were at inland locations. 46 sites remain in operation. This section provides a short description of the NDGPS system.

Network Design

The network architecture for NDGPS is shown in Figure 12.21. These systems essentially utilize the code-based local-area DGNSS technique described in Section 12.2.1. The network includes reference stations (RSs) to monitor GPS and generate differential corrections. Each RS consists of two GPS receivers for redundancy.

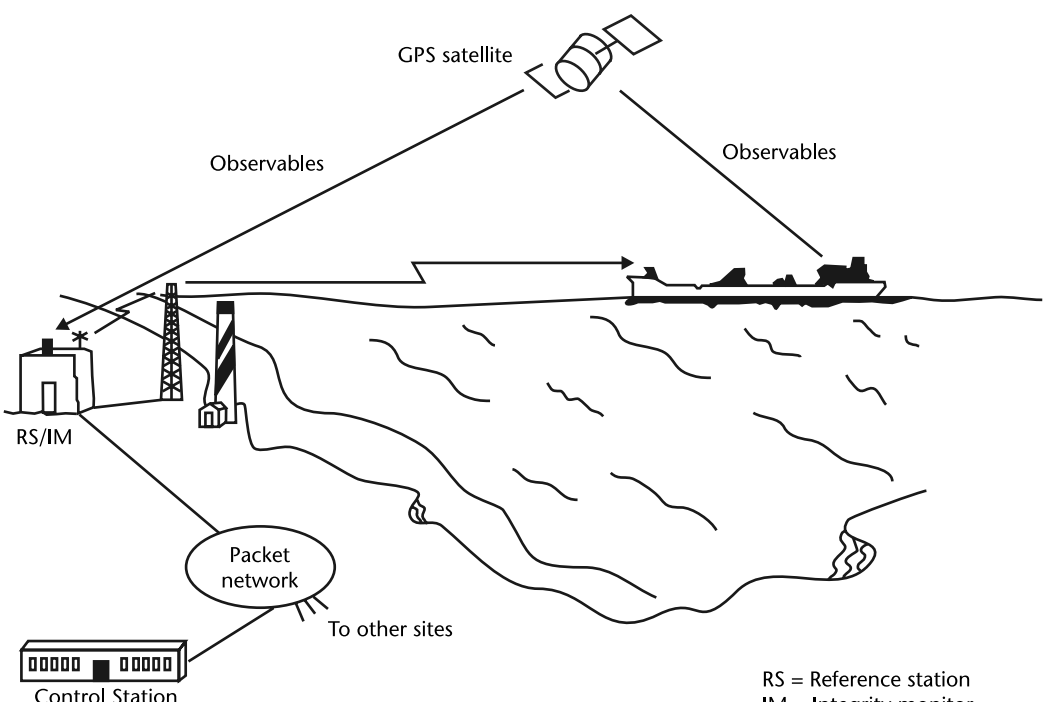


Figure 12.21 NDGPS network architecture.

Integrity monitors (IMs) are collocated with the RSs. All the equipment is generally installed in unmanned equipment sheds with a backup power source (e.g., batteries or a generator). Each IM includes another pair of GPS receivers and also radio beacon receivers to monitor the corrections that the site is itself broadcasting. The IMs compute their positions using GPS and the differential corrections and compare their computed positions with their known (surveyed) positions. If the position exceeds a preset tolerance, problem satellites are expunged from the differential correction calculation and the user is notified that the satellite is unhealthy or the site is shut down to guarantee accurate and reliable information.

A packet network is used so that a central control station in Alexandria, Virginia, which is manned 24 hours a day, 7 days a week, can monitor the status of all the sites. Personnel at the control station, upon observing an equipment failure, can switch in redundant hardware or dispatch a maintenance crew if necessary.

Data Link

Each RS/IM broadcasts digital DGPS corrections in the RTCM SC-104 message format. Version 2.1 message types 3, 5, 6, 7, 9, and 16 are currently supported [53]. These message types are retained in the later RTCM SC-104 Version 2.3 standard, described in Section 12.5.1. The digital data is broadcast in the 285–325-kHz medium frequency (MF) band, which is allocated internationally for radio beacons. A digital modulation technique referred to as minimum shift keying (MSK) (see, e.g., [54]) is employed either directly on the radio beacon center frequencies or on a subcarrier. The use of a subcarrier was originally motivated by the desire to not interfere with direction finding receivers that employed existing radio beacon signals [55]. At present, all marine radio beacons in the United States that are not used for NDGPS have been decommissioned, so backwards compatibility is no longer an issue. Impulsive noise due, for example, to lightning strikes, is prevalent in the MF band at sea because of the excellent conductivity of salt water. This led to a decision to use Type 9 SC-104 messages rather than Type 1 to broadcast pseudorange and range-rate corrections. The use of Type 9 messages provides more frequent preambles for user equipment to resynchronize following a strong impulse. The standards for NDGPS support data rates of 50, 100, or 200 bps. All of the NDGPS sites are currently transmitting at data rates of either 100 or 200 bps.

A large antenna is required to broadcast efficiently at MF because of the large wavelength (1 km). Most of the original NDGPS sites (on the coasts and inland waters) use converted radio beacon broadcast towers with heights ranging from 90 to 150 feet. The NDGPS expansion began with the conversion of 47 obsolete U.S. Air Force Low Frequency Ground Wave Emergency Network (GWEN) sites in one of the largest military to civilian reutilization projects. The GWEN sites were already equipped with 299-foot antennas. A minimum field strength of 75 μV/m for a 100-bps transmission is specified within each transmitters coverage volume [53], which is typically on the order of 250 nmi.

Performance

The specified accuracy of NDGPS systems is 10m, 2drms, within coverage areas [56]. Typical accuracies are much better, typically 1 to 3m. An often-used rule of thumb is 1-m accuracy at the base of a transmitter with errors growing by 1m per

150 km of separation [56]. The specified availability is 99.9% for selected waterways and dual-coverage areas and 98.5% for areas with single-coverage, based upon a 1-month average per site and discounting GPS anomalies [56]. Current coverage is shown in Figure 12.22.

International Harmonization

International standards for maritime DGPS systems, fully compatible with NDGPS, have been developed by the International Maritime Organization (IMO). As of the time of this writing, radio beacon-based DGPS services have been deployed in over 30 nations.

12.6.1.2 Satellite-Based Augmentation System (SBAS)

The International Civil Aviation Organization (ICAO) has developed standards [ICAO] for two types of code-based DGPS systems for civil aircraft navigation applications. This section will describe systems of the first type, which are referred to as *Satellite-Based Augmentation Systems* (SBASs). The following section will describe systems of the second type, which are referred to as *Ground-Based Augmentation System* (GBASs).

An SBAS is a wide-area DGPS system that provides differential GPS corrections and integrity data using geostationary satellites (GEOs) as the communications

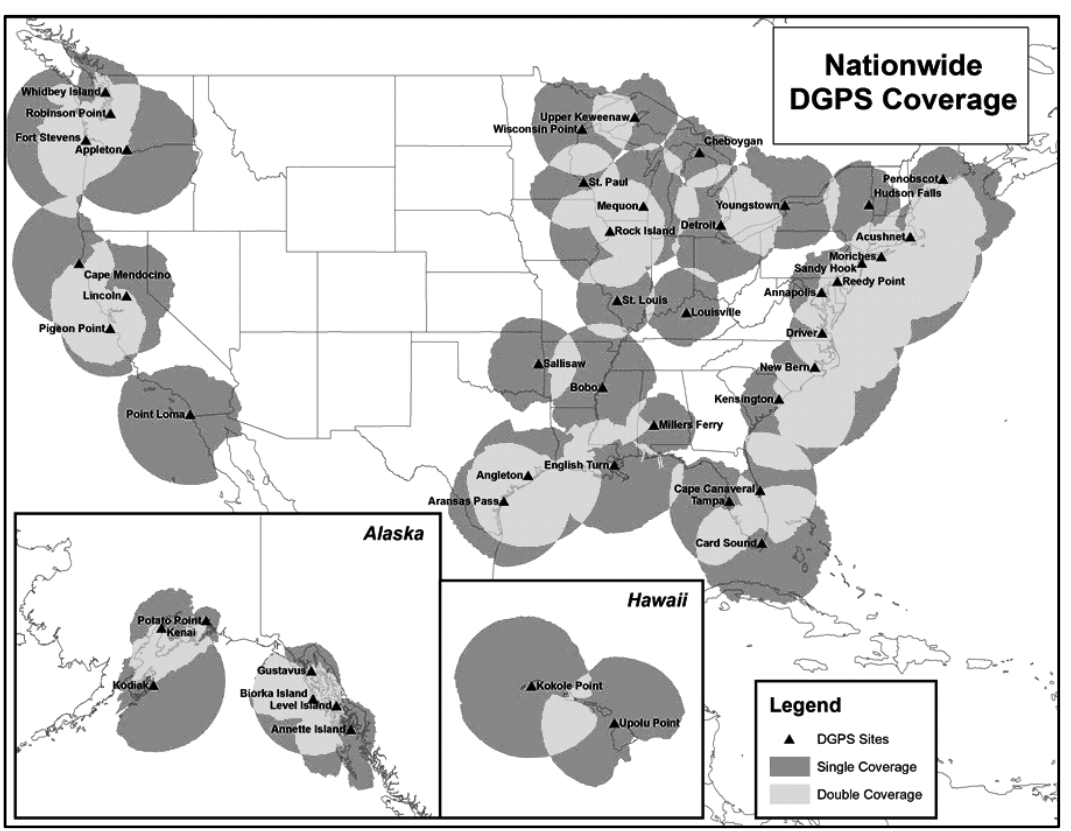


Figure 12.22 NDGPS coverage.

path [57]. A unique feature of SBASs is that they provide DGPS data using a signal broadcast directly at the GPS L1 frequency that can be used for ranging. The goal of SBASs is to meet navigation system requirements for civil aviation from the en-route phase of flight through vertically guided precision approach. A number of SBASs had been implemented or were planned as of the time of this writing [57]. These include the Wide Area Augmentation System (WAAS) within the United States, the European Geostationary Navigation Overlay Service (EGNOS) within Europe, the Multifunctional Transport Satellite (MTSAT)-based Augmentation System (MSAS) within Japan and Southeast Asia, the GPS Aided GEO Augmented Navigation (GAGAN) system in India, the System of Differential Corrections and Monitoring (SDCM) in Russia, the BeiDou Satellite Based Augmentation System (BDSBAS) in China, and the Korean Augmentation Satellite System (KASS) in the Republic of Korea.

History

As discussed in Chapter 11, RAIM or DGNSS are required to provide the necessary levels of integrity to GNSS to support air navigation. In the early 1980s, a concept of providing integrity data for GPS over a GEO communications link using a signal on the GPS L1 frequency emerged. This concept was referred to as a GPS integrity channel (GIC) [58]. In 1989, Inmarsat began test transmissions of GPS-like spread spectrum signals through a geostationary satellite over the Atlantic Ocean to prove the feasibility of using a navigation repeater to transmit pseudorandom-coded spread spectrum ranging signals. The test results indicated that transmitting these signals through geostationary satellites was possible [59]. In the same timeframe, organizations including Inmarsat and RTCA's Special Committee 159 (SC-159) began establishing a signal format for GIC, which later evolved into SBAS. In the 1990s, SBAS programs were well underway within the United States, Europe, and Japan. Inmarsat on their own initiative included navigation transponders on the five Inmarsat-3 satellites that were launched from April 1996 to February 1998. In November 1999, Japan attempted to launch their own SBAS GEO for MSAS, but experienced a setback when the satellite, MTSAT-1, had to be destroyed following a launch failure. In August 2000, the U.S. Federal Aviation Administration (FAA)'s WAAS system, using two of the Inmarsat-3 satellites, Atlantic Ocean Region West (AOR-W) and Pacific Ocean Region (POR), was declared to be continually available for nonsafety applications. In July 2003, WAAS was commissioned for safety-of-life services. The MSAS ground segment is complete and the replacement satellite, MTSAT-1R, was successfully launched in February 2005. EGNOS began its initial operations utilizing three GEOs, two Inmarsat-3 (Atlantic Ocean Region East [AORE], Indian Ocean Region [IOR]) and one European Space Agency (ESA) satellite, Artemis, in July 2005. MSAS was commissioned for safety-of-life, horizontal-only guidance in September 2007. EGNOS commissioned its open service in October 2009 and became certified for safety-of-life service in March 2011. GAGAN began providing safety-of-life, horizontal-only guidance in February 2014 using two GEOs: GSAT-8 and GSAT-10. GAGAN became certified for vertical guidance in April 2015 and a third GEO, GSAT-15, launched in November 2015 serves as an in-orbit spare. SDCM, BDSBAS, and KASS are all in their development phases and anticipate service by 2020.

SBAS Requirements

ICAO requirements for SBAS and GBAS for en-route through Category I precision approach operations are shown in Table 12.3 [60]. It has become apparent that Category I requirements cannot be met by SBAS utilizing only the GPS L1 C/A-code signal. Category I service will require L5-capable GPS satellites and perhaps a second constellation. New classes of vertically guided approaches, referred to as GNSS approach and operations with vertical guidance (APV)-I and II have been defined to enable the full utility of the performance that SBASs currently provide. WAAS, EGNOS, and GAGAN are all pursuing the most stringent member of this new class called Localizer Precision approaches for Vertical guidance with a 200-foot decision height (LPV-200). This SBAS service is nearly equivalent to Category I service as it shares the same decision height.

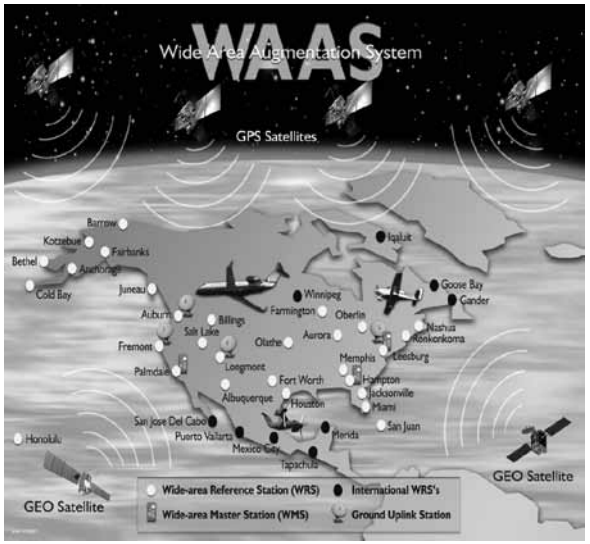
SBAS Architecture and Functionality

All SBAS systems are composed of four subelements: monitoring receivers, central processing facilities, satellite uplink facilities, and one or more geostationary satellites. Unfortunately, the terminology for these subelements is not consistent among the specific implementations. Within the U.S. WAAS, the monitors are referred to as wide-area reference stations (WRSs), the central processing facilities are known as wide area master stations (WMSs) and the uplink facilities as ground uplink stations (GUSs) (see Figure 12.23). Within EGNOS, these elements are referred to as ranging and integrity monitoring stations (RIMS), mission control centers (MCCs), and navigation land Earth stations (NLES), respectively (see Figure 12.24). Within MSAS, the respective terms are ground monitoring stations (GMSs), master control stations (MCSs), and GESs. MSAS contains 6 GMSs and 2 MCSs (see Figure 12.25). For GAGAN, the terms Indian Reference Station (INRES), Indian Master Control Centre (INMCC), and Indian Land Uplink Station (INLUS) are used (see Figure 12.26).

Table 12.3 ICAO GNSS Signal-in-Space Performance Requirements

Operation	*Horizontal/ Vertical Accuracy (95%)*	*Integrity Level*	*Horizontal/ Vertical Alert Limit*	*Time to Alarm*	*Continuity*	*Availability*
En-route	3.7 km/NA	$1–1 \times 10^{-7}$/h	3.7 to 7.4 km/ NA	5 minutes	$1–1 \times 10^{-4}$/h to $1–1 \times 10^{-8}$/h	0.99 to 0.99999
Terminal	0.74 km/NA	$1–1 \times 10^{-7}$/h	1.85 km/NA	15 seconds	$1–1 \times 10^{-4}$/h to $1–1 \times 10^{-8}$/h	0.999 to 0.99999
Nonprecision approach	220m/NA	$1–1 \times 10^{-7}$/h	556m/NA	10 seconds	$1–1 \times 10^{-4}$/h to $1–1 \times 10^{-8}$/h	0.99 to 0.99999
Approach with vertical guidance (APV)-I	16m/20m	$1–2 \times 10^{-7}$/ approach	40m/50m	10 seconds	$1–8 \times 10^{-6}$ in any 15 seconds	0.99 to 0.99999
Approach with vertical guidance (APV)-II	16m/8m	$1–2 \times 10^{-7}$/ approach	40m/20m	6 seconds	$1–8 \times 10^{-6}$ in any 15 seconds	0.99 to 0.99999
Category I	16 m/4 to 6m	$1–2 \times 10^{-7}$/ approach	40m/10 to 35m	6 seconds	$1–8 \times 10^{-6}$ in any 15 seconds	0.99 to 0.99999

Source: [60]. NA = Not applicable.



38 Reference Stations

3 Master Stations

6 Ground Earth Stations

3 Geostationary Satellite Links

2 Operational Control Centers

Figure 12.23 WAAS ground network.

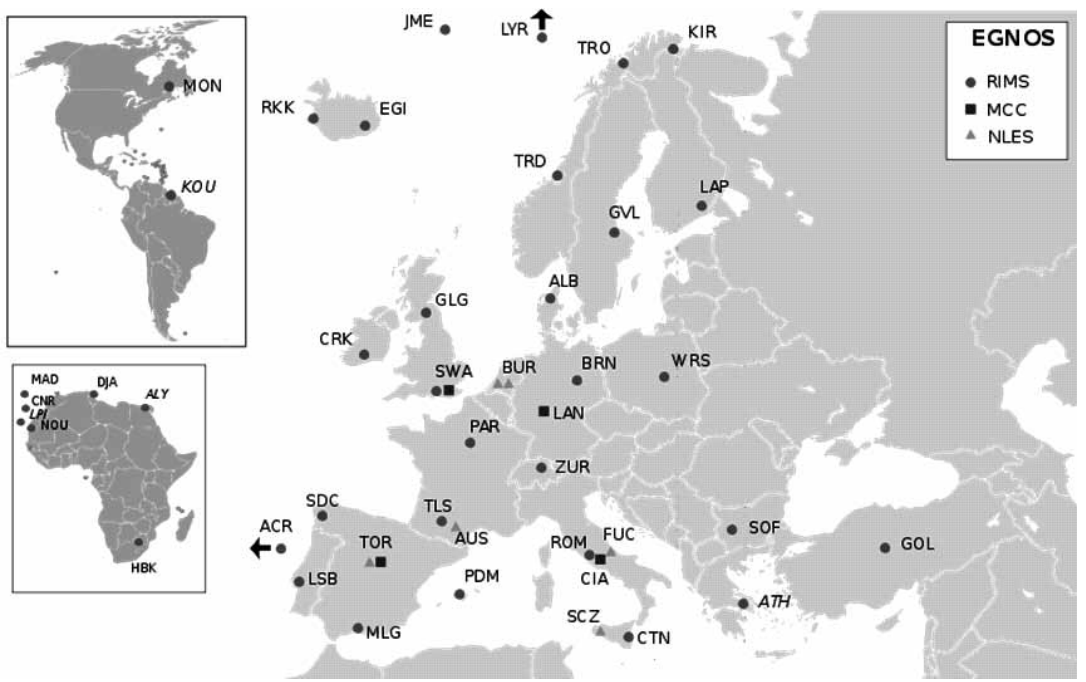


Figure 12.24 EGNOS ground network.

The functionality provided by the current operational SBAS subelements is summarized in Figure 12.27. As observed in Figure 12.27, users receive navigation signals transmitted from GPS satellites. These signals are also received by monitoring networks operated by the SBAS service providers. In the near future, satellites from other core constellations will also be monitored by some SBASs. Each site within the monitoring networks generally includes a number of GPS receivers (for redundancy) that provide L1 C/A code and L2 P(Y) code pseudorange and carrier

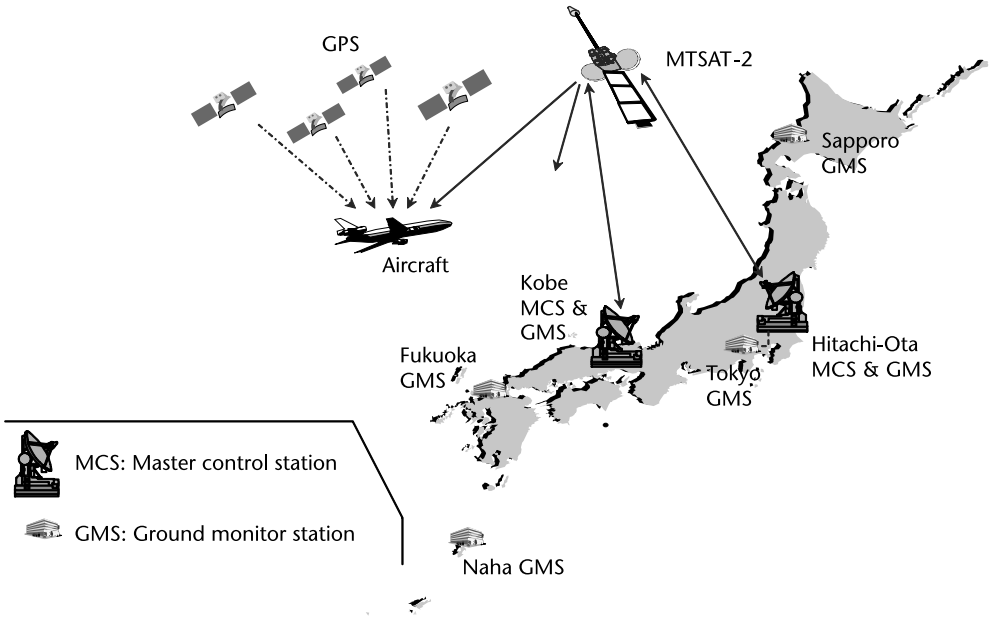


Figure 12.25 MSAS ground network.

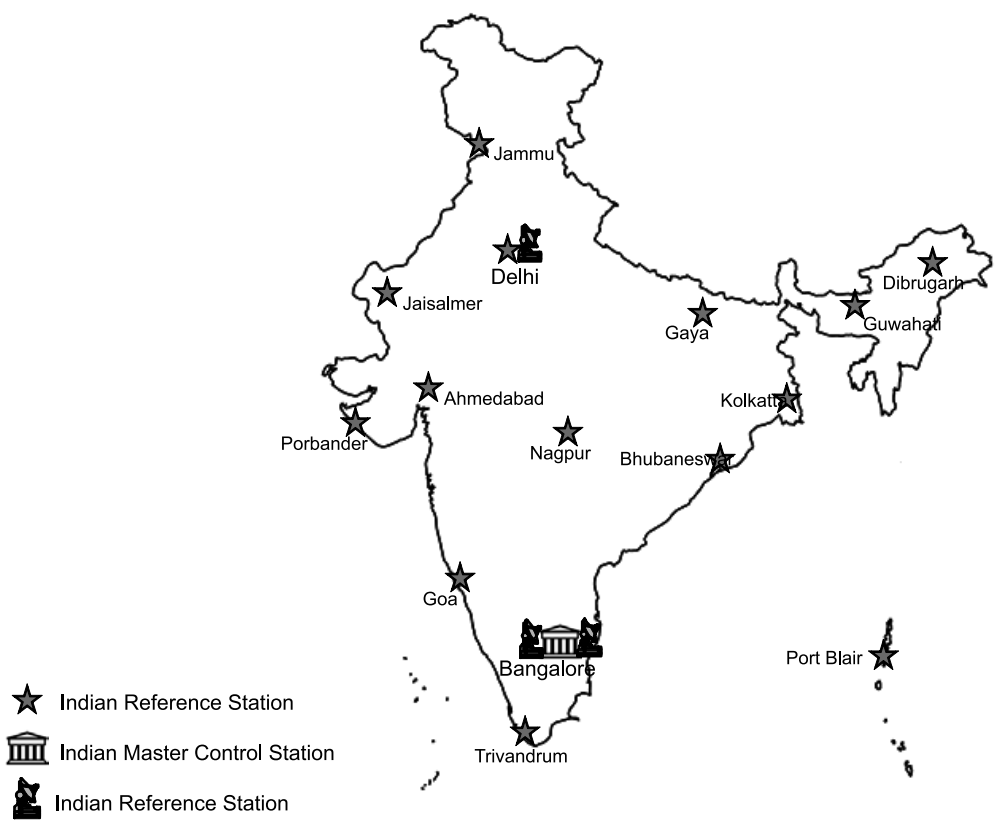


Figure 12.26 GAGAN ground network.

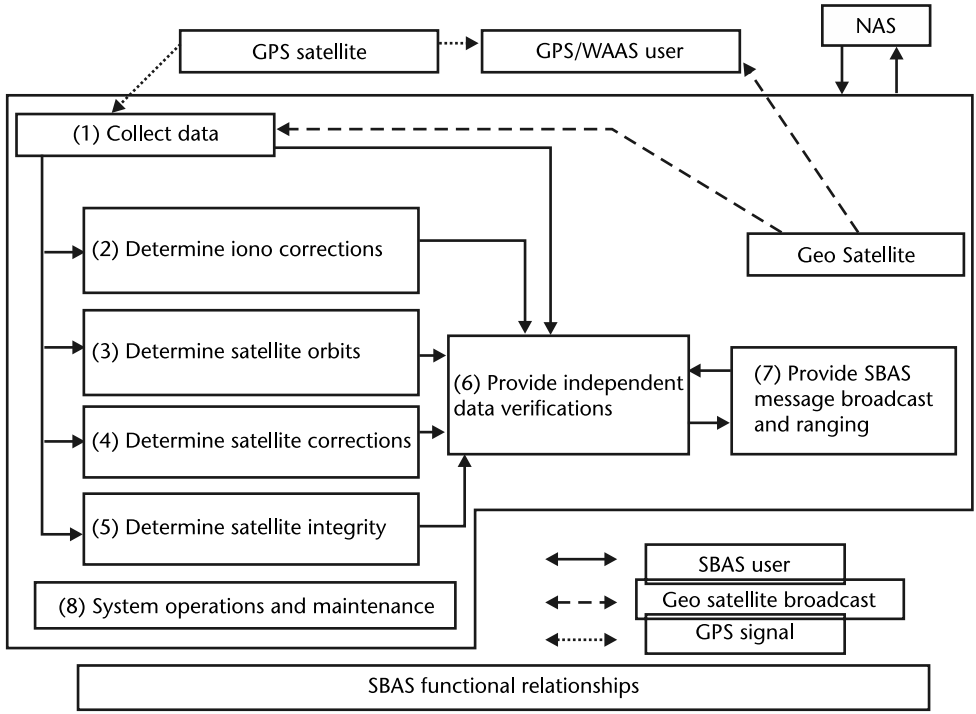


Figure 12.27 SBAS functional overview.

phase data (using semicodeless processing techniques for the L2 measurements; see Chapter 8) to the central processing facilities. At each central processing facility, the data from the entire network are processed to develop estimates of each GPS satellite's true position and clock values, corrections based upon the differences between the network's estimates of these parameters and the values in the broadcast GPS navigation data, as well as an estimate of the vertical ionospheric delay error across the service area. Each central processing facility also checks for problem satellites (e.g., those whose signals are distorted or whose clocks are running erratically) that the SBAS user may be warned to not use. These estimates and integrity information are used to form wide-area differential corrections and integrity messages that are then forwarded to the satellite uplink facilities. At the uplink facilities, the spread spectrum navigation signal is generated and precisely synchronized to a reference time and modulated with the SBAS data. This composite signal is continuously transmitted to a geostationary satellite. Onboard the satellite, this navigation signal is frequency translated within the navigation payload and transmitted to the user on the GPS L1 frequency. The timing of the signal is done in a very precise manner so that the signal appears as though it was generated onboard the satellite as a GPS ranging signal. Redundant central processing and uplink facilities may be used to provide hot standbys in the event of a failure at the primary facility.

SBAS Signal Structure

The signal broadcast via the SBAS geostationary satellite to the SBAS users [61] is designed to minimize standard GPS receiver hardware modifications. The GPS L1 frequency and GPS C/A code type of modulation, including the use of length-1,023

Gold codes at a 1.023-MHz chip rate, are used. In addition, the code phase timing is synchronized to GPS System Time to emulate a GPS satellite and provide a ranging capability. A data rate of 250 bps is used. The data is convolutionally encoded using a rate 1/2 constraint length 7 encoder to generate an overall symbol rate of 500 symbols/s. The SBAS data symbols are synchronized with the 1-kHz GPS C/A code epochs.

The C/A codes used by SBAS belong to the same family of 1,023-bit Gold codes as the 63 PRN codes reserved by the GPS system described in Chapter 3. The SBAS C/A codes were specifically selected to not adversely interfere with GPS signals (see, e.g., [62]). The 39 current SBAS C/A codes and the associated geostationary satellites are shown in Table 12.4. The listing of PRN code assignments is maintained by the United States Air Force GPS Directorate, accessible at www.gps.gov/technical/#prn. The SBAS C/A codes are identified by the PRN number; and the G2 delay in chips and the initial G2 state. The definition of either the G2 delay or initial G2 setting is required for implementation of the generation of the SBAS C/A codes. Like the GPS C/A codes, the PRN number is arbitrary, but for SBAS starts with 120 instead of 1. The actual codes are defined by either the G2 delay or the initial G2 register setting. In the octal notation for the first 10 chips of the SBAS code as shown in the table, the first digit on the left represents a 0 or 1 for the first chip. The last three digits are the octal representation of the remaining 9 chips. For example, the initial G2 setting for PRN 120 is 1001000110. Note that the first 10 SBAS chips are simply the octal inverse of the initial G2 setting.

Some current and all future SBAS satellites will be capable of also transmitting a signal on the GPS L5 frequency. Such signals use or will use PRN codes from the same family as the GPS L5 signals (see Section 3.7.2.2), but without a dataless component. The L5 data rate is 250 bps convolutionally encoded into a 500 symbol/s stream. This service and the corresponding data content are still in the process of being defined.

SBAS Message Format and Contents

The 250-bps data from each SBAS GEO is packed into 1-second blocks of 250 bits, as shown in Figure 12.28. Each block includes an 8-bit preamble (one of three parts of a 24-bit unique word, 01010011 10011010 11000110, that is distributed over three blocks), a 6-bit message type field (allowing for up to 64 message types), a 212-bit payload with unique meaning specifically defined for each message type, and 24 bits of CRC parity for error detection as shown in Figure 12.28. The start of every other 24-bit preamble is synchronous with a 6-second GPS subframe epoch. The preambles and timing information provided in the messages facilitate data acquisition. They also aid the user receiver to perform time synchronization during initial acquisition before GPS satellites are acquired, thus aiding the receiver in subsequent GPS satellite acquisitions.

Table 12.5 lists the message types that have been defined thus far for SBAS. These message types support the basic wide area GPS concepts discussed in Section 12.2.3. Message Types 2 to 5 provide broadcast clock corrections. Message Type 25 provides broadcast orbit corrections. Message Type 26 provides the L1-only user with vertical ionospheric delay values over a grid of locations with predefined latitude and longitude values. Each user receiver calculates the latitude and longitude

Table 12.4 SBAS Ranging C/A Codes

PRN	G2 Delay (Chips)	Initial G2 Setting (Octal)	First 10 SBAS Chips (Octal)	Geostationary Satellite PRN Allocations	Orbital Slot
120	145	1,106	0671	EGNOS (INMARSAT 3F2)	15.5 W
121	175	1,241	0536	EGNOS (INMARSAT 3F5)	25 E
122	52	0267	1,510	Unallocated	—
123	21	0232	1,545	EGNOS (ASTRA 5B)	31.5 E
124	237	1,617	0160	EGNOS (Reserved)	—
125	235	1,076	0701	SDCM (Luch-5A)	16 W
126	886	1,764	0013	EGNOS (INMARSAT 4F2)	25 E
127	657	0717	1,060	GAGAN (GSAT-8)	55 E
128	634	1,532	0245	GAGAN (GSAT-10)	83 E
129	762	1,250	0527	MSAS (MTSAT-2)	145 E
130	355	0341	1,436	ARTEMIS (ARTEMIS-1)	21.5 E
131	1012	0551	1,226	WAAS (Satmex 9)	117 W
132	176	0520	1,257	GAGAN (GSAT-15)	93.5 E
133	603	1,731	0046	WAAS (INMARSAT 4F3)	98 W
134	130	0706	1,071	Unallocated	—
135	359	1,216	0561	WAAS (Intelsat Galaxy 15)	133 W
136	595	0740	1,037	EGNOS (ASTRA 4B)	5 E
137	68	1,007	0770	MSAS (MTSAT-2)	145 E
138	386	0450	1,327	WAAS (ANIK-F1R)	107.3
139	797	0305	1,472	Unallocated	—
140	456	1,653	0124	SDCM (Luch-5B)	95 E
141	499	1,411	0366	SDCM (Luch-4)	167 E
142	883	1,644	0133	Unallocated	—
143	307	1,312	0465	Unallocated	—
144	127	1,060	0717	Unallocated	—
145	211	1,560	0217	Unallocated	—
146	121	0035	1,742	Unallocated	—
147	118	0355	1,422	Unallocated	—
148	163	0355	1,442	Unallocated	—
149	628	1,254	0523	Unallocated	—
150	853	1,041	0736	Unallocated	—
151	484	0142	1,635	Unallocated	—
152	289	1,641	0136	Unallocated	—
153	811	1,504	0273	Unallocated	—
154	202	0751	1,026	Unallocated	—
155	1021	1,774	0003	Unallocated	—
156	463	0107	1,670	Unallocated	—
157	568	1,153	0624	Unallocated	—
158	904	1,542	0235	Unallocated	—

of the intersection points between each GPS signal and the ionosphere, which is modeled as a thin shell at 350 km altitude above the surface of the Earth. The vertical ionospheric delays at these intersection points, referred to as *ionospheric pierce points* (IPPs), are determined for each visible satellite by interpolating the

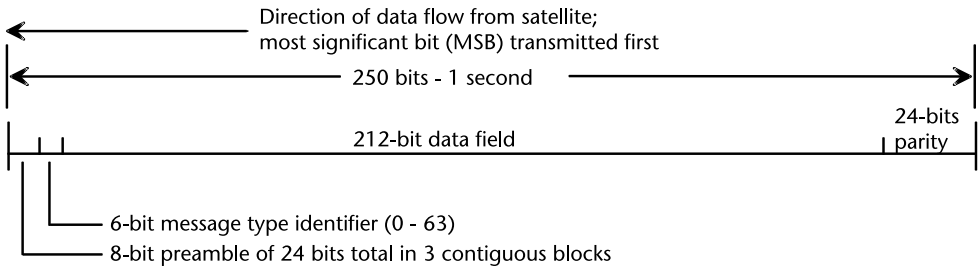


Figure 12.28 SBAS data block format.

Table 12.5 SBAS Message Types

Type	*Contents*
0	Do not use for safety applications (for SBAS testing)
1	PRN Mask assignments, set up to 51 of 210 bits
2 to 5	Fast corrections
6	Integrity information
7	Fast correction degradation factor
8	Reserved for future messages
9	GEO navigation message (*X*, *Y*, *Z*, time, and so forth)
10	Degradation Parameters
11	Reserved for future messages
12	SBAS Network Time/UTC offset parameters
13 to 16	Reserved for future messages
17	GEO satellite almanacs
18	Ionospheric grid point masks
19 to 23	Reserved for future messages
24	Mixed fast corrections/long-term satellite error corrections
25	Long-term satellite error corrections
26	Ionospheric delay corrections
27	SBAS Service Message
28	Clock-Ephemeris Covariance Matrix Message
29 to 61	Reserved for future messages
62	Internal Test Message
63	Null Message

delays from the three or four nearest grid points as discussed later in this section. The reader is referred to [61] for a complete description of the messages and their application.

User Algorithms

SBAS user equipment is modified GPS L1 C/A code receivers. The equipment must be modified to be able to generate and track the SBAS PRN codes described above,

demodulate the higher-rate (250 bps) convolutionally encoded data, and must include modified software to apply the corrections and integrity data.

Application of the clock and ephemeris corrections is straightforward. Message Types 2 to 5 provide range domain clock corrections that are simply added to the receiver's raw pseudorange measurements for all visible satellites. The SBAS data does not include range rate corrections. These are generated within the user equipment itself by differencing successive clock corrections [62, 63]. Message Type 25 provides broadcast satellite position corrections in ECEF x, y, z coordinates. Satellite broadcast position error rate terms and a clock bias term can also be provided, if necessary, in Message Type 25 using a 1-bit velocity code flag.

As mentioned earlier, currently operational SBASs only use L1-only GPS and GEO signals. Ionospheric corrections for visible satellites are determined using an interpolation algorithm using SBAS broadcast vertical ionospheric delay values. Applying the law of sines to Figure 12.29, the user first calculates the angle ψ_{pp}, the Earth's central angle between the user position and pierce point:

$$\psi_{pp} = \frac{\pi}{2} - E - \sin^{-1}\left(\frac{R_E}{R_E + h} \cdot \cos E\right)$$

where R_E is the radius of the Earth, h is the altitude of the IPP, and E is the elevation angle of the satellite from the user position. The user then calculates the latitude, ϕ_{pp}, and longitude, λ_{pp}, of the IPP, as follows:

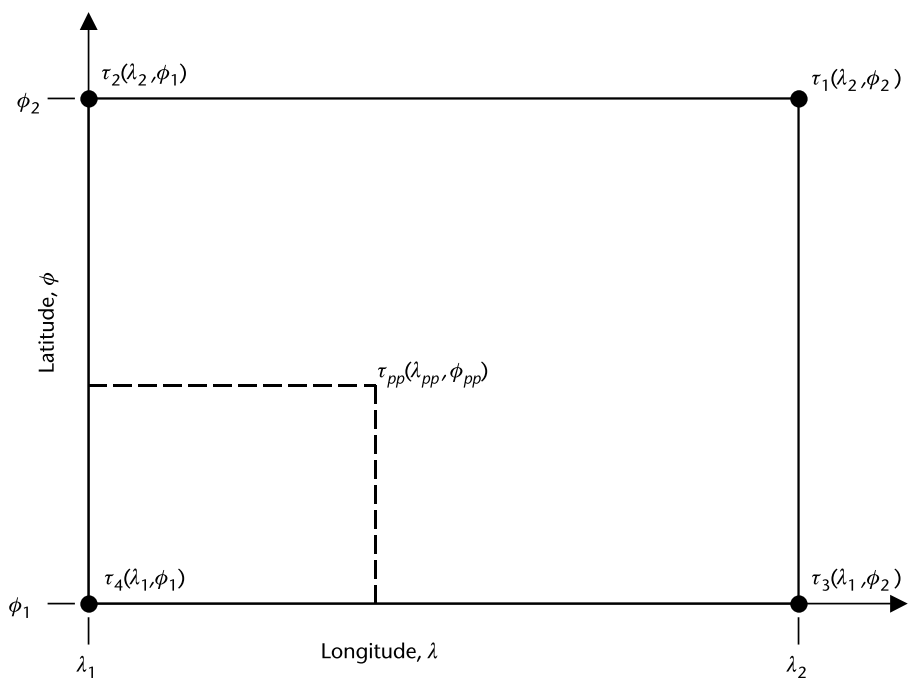


Figure 12.29 Finding the relative IPP position.

$$\phi_{pp} = \sin^{-1}\left(\sin\phi_u \cdot \cos\psi_{pp} + \cos\phi_u \cdot \sin\psi_{pp} \cdot \cos A\right)$$

$$\lambda_{pp} = \lambda_u + \pi - \sin^{-1}\left(\frac{\sin\psi_{pp} \cdot \sin A}{\cos\phi_{pp}}\right)$$

$$\text{if } \phi_u > 70^\circ, \text{ and } \tan\psi_{pp}\cos A > \tan\left(\frac{\pi}{2} - \phi_u\right)$$

$$\text{or if } \phi_u < -70^\circ, \text{ and } \tan\psi_{pp}\cos A < \tan\left(\frac{\pi}{2} - \phi_u\right)$$

$$\lambda_{pp} = \lambda_u + \sin^{-1}\left(\frac{\sin\psi_{pp} \cdot \sin A}{\cos\phi_{pp}}\right) \qquad \text{,otherwise}$$

where the angles λ_u and ϕ_u are the azimuth and elevation angles, respectively, of the satellite from the user's position. Then the receiver determines the most suitable set of predefined grid points in the proximity of the IPP for each visible satellite. If no suitable set is available, then an ionospheric correction is unavailable for that particular satellite. If four suitable surrounding grid points are found, the receiver determines the IPP position relative to those four points, using Figure 12.29, from the following equations:

$$x_{pp} = \frac{\lambda_{pp} - \lambda_1}{\lambda_2 - \lambda_1} \quad y_{pp} = \frac{\phi_{pp} - \phi_1}{\phi_2 - \phi_1} \text{ for IPPs between N85° and S85°}$$

$$\left.\begin{aligned} y_{pp} &= \frac{\left|\phi_{pp}\right| - 85^\circ}{10^\circ} \\ x_{pp} &= \frac{\lambda_{pp} - \lambda_1}{90^\circ} \cdot \left(1 - 2y_{pp}\right) + y_{pp} \end{aligned}\right\} \text{ for IPPs above N85° and below S85°}$$

The interpolation is weighted, with greater weights given to the nearer grid points. The weights are given by

$$W_1 = x_{pp} \cdot y_{pp}$$
$$W_2 = \left(1 - x_{pp}\right) \cdot y_{pp}$$
$$W_3 = \left(1 - x_{pp}\right) \cdot \left(1 - y_{pp}\right)$$
$$W_4 = x_{pp} \cdot \left(1 - y_{pp}\right)$$

Finally, the vertical delay, τ_{pp}, at the IPP is determined by

$$\tau_{pp}\left(\lambda_{pp}, \phi_{pp}\right) = \sum_{i=1}^{4} W_i \cdot \tau_i \tag{12.53}$$

where the τ_i are the vertical delays at the four grid points provided in the Ionospheric Delay Corrections Message (Message Type 26).

If only three of the four suitable surrounding grid points are available, the calculation of the weights is modified slightly, as follows:

$$W_1 = y_{pp}$$
$$W_2 = 1 - x_{pp} - y_{pp}$$
$$W_3 = x_{pp}$$

The same delay formula (12.53) is used, except that the sum is over three weightings. The remaining ionospheric delay calculation accounts for a difference in delay from the vertical, and is a function of the elevation angle to the satellite. To obtain the ionospheric correction, which is added to the pseudorange measurement, the vertical delay $\tau_{pp}(\lambda_{pp}, \phi_{pp})$is multiplied by the *obliquity factor*, F, where

$$F = \left[1 - \left(\frac{R_E \cos E}{R_E + h}\right)^2\right]^{-\frac{1}{2}}$$

To account for tropospheric delays, the user equipment is required to apply a tropospheric delay correction to each raw pseudorange measurement. The UNB3 algorithm presented in Chapter 10 is employed. After applying all specified corrections, SBAS user equipment computes user position using a weighted least squares algorithm.

In addition to application of the SBAS differential corrections, user equipment for safety applications must also compute position error bounds, referred to as the *horizontal protection level* (HPL) or *vertical protection level* (VPL) in the local horizontal and vertical directions, respectively. These protection levels represent the user position errors that will not be exceeded without a timely warning. They are determined using the associated probability levels and time to alerts listed in Table 12.3 under "Integrity Level." The HPL and VPL levels are continually compared to the applicable *horizontal and vertical alert limits* (HAL and VAL) for the current phase of flight and a warning is issued to the pilot if HPL > HAL or VPL > VAL pertaining to that operation. For instance, for APV-I approaches, VAL = 50m, HAL = 40m, and the time to alert is 10 seconds. An SBAS system is designed so that the probability is less than 2×10^{-7} per approach that an aircraft conducting an APV-I approach computes VPL < 50m and HPL < 40m when the true vertical or horizontal position errors are greater than these levels for longer than 6 seconds without a warning being issued.

The user equipment computes HPL and VPL using variances that are broadcast in Message Types 2 to 6 for the SBAS fast and long-term corrections and in Message Type 26 for the SBAS ionospheric corrections. A set of complicated rules are also applied to adjust these variances for latency, missed messages, and other factors [61]. Variances for receiver noise, multipath, and residual tropospheric errors are computed based upon the elevation angles of the visible satellites. The individual

error variances are summed to form overall residual pseudorange error variances for the visible satellite. Finally, the geometry matrix and known weighting matrix for the WLS solution are used to bound the standard deviation of horizontal and vertical position errors. Under the assumption that all errors are Gaussian, multipliers for the horizontal and vertical position error standard deviations are applied to determine HPL and VPL. Although it is well known that the true residual errors are not Gaussian, the variances in the broadcast message and the variances for receiver noise, multipath, and tropospheric errors are inflated by design to represent Gaussian distributions that *overbound* the true errors for the probabilities of interest [64].

SBAS GEOs

At present, the U.S. WAAS uses the Intelsat Galaxy 15 satellite at 133°W (launched in October 2005), the ANIK-F1R satellite at 107.3°W (launched in September 2005), and the Inmarsat-4F3 satellite at 98°W (launched in August 2008) with coverage as shown in Figure 12.30. The coverage contour for each satellite surrounds the user locations where the satellite is visible above 5° in elevation angle. In the near future, WAAS will replace one of these satellites with Eutelsat 9 at 117°W. A functional overview of the Inmarsat-4 navigation payload, which is a representative design for SBAS GEO transponders, is provided in Figure 12.31 [65].

The current EGNOS space segment is composed of the INMARSAT 3F2 satellite at 15.5°W (launched in September 1996) and the SES 5 satellite at 5°E (launched in July 2012). An ASTRA 5B satellite at 31.5°E, launched in March 2014, is in test mode and expected to be entered into service soon. The current EGNOS GEO coverage is shown in Figure 12.32 (the dashed coverage contour is for the ASTRA 5B satellite, which is in test mode).

MSAS currently only has a single GEO, MTSAT-2, which is capable of broadcasting two PRN codes. This is a unique capability of MSAS. Each PRN code signal

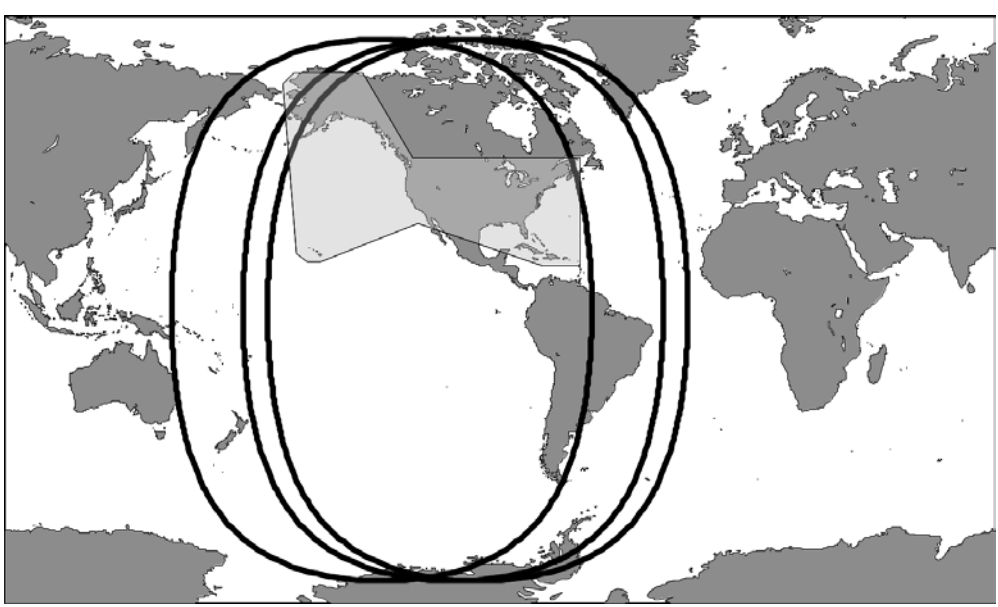


Figure 12.30 Current WAAS GEO coverage and primary service area.

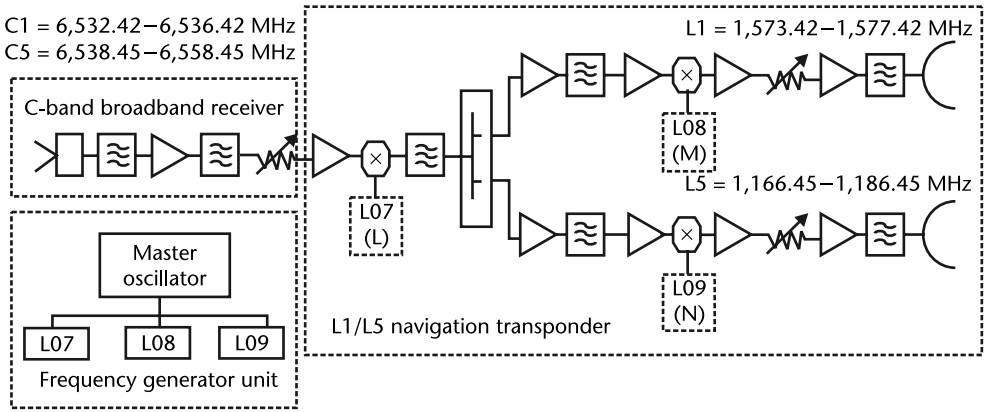


Figure 12.31 Inmarsat-4 navigation payload.

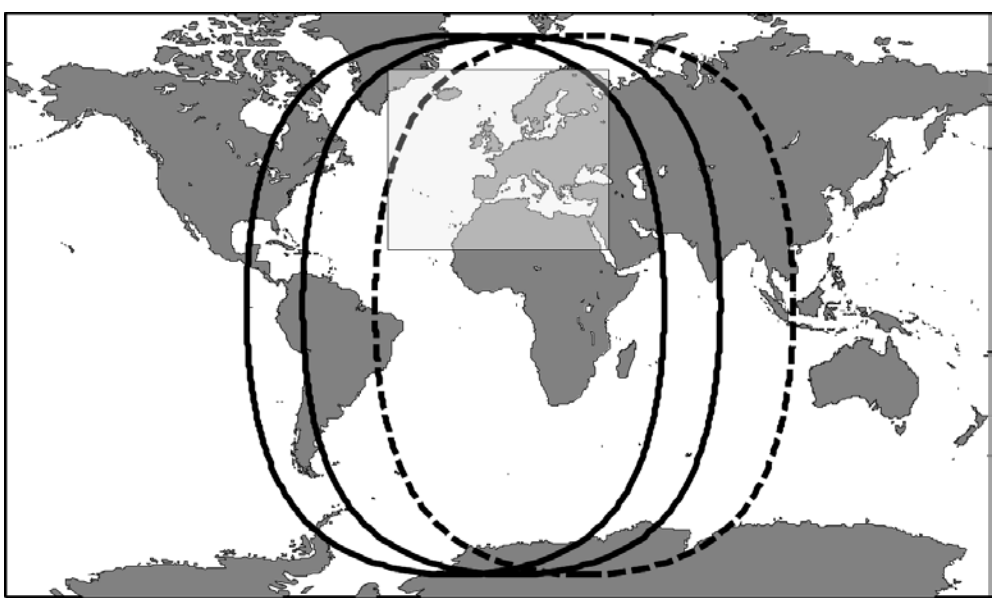


Figure 12.32 Current EGNOS space segment and primary service area.

is generated from a separate MCS. Thus, although the spacecraft has no redundant backup, if either MCS or uplink station experiences a fault, a redundant signal is available to the users. The first GEO used for MSAS, MTSAT-1R, was retired from service in December 2015. The coverage of MTSAT-2 at 145°E and the MSAS primary service area are shown in Figure 12.33.

GAGAN currently uses the GSAT-8 and GSAT-10 satellites at 55°E and 83°E, respectively. A third GEO, GSAT-15, is also equipped with an SBAS transponder and was launched in November 2015. GSAT-15 is located at 93.5°E and is currently not broadcasting SBAS signals but could do so if either GSAT-8 or GSAT-10 experienced an SBAS transponder failure. Figure 12.34 shows the current GEO coverage for GAGAN (with the coverage of the reserve GSAT-15 GEO shown with a dashed line).

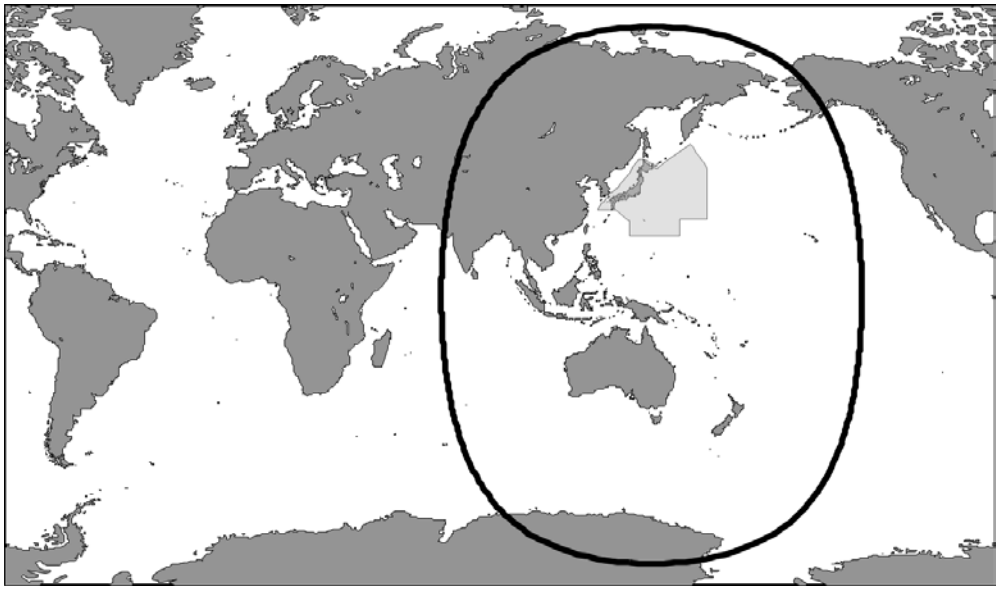


Figure 12.33 MSAS GEO coverage and primary service area.

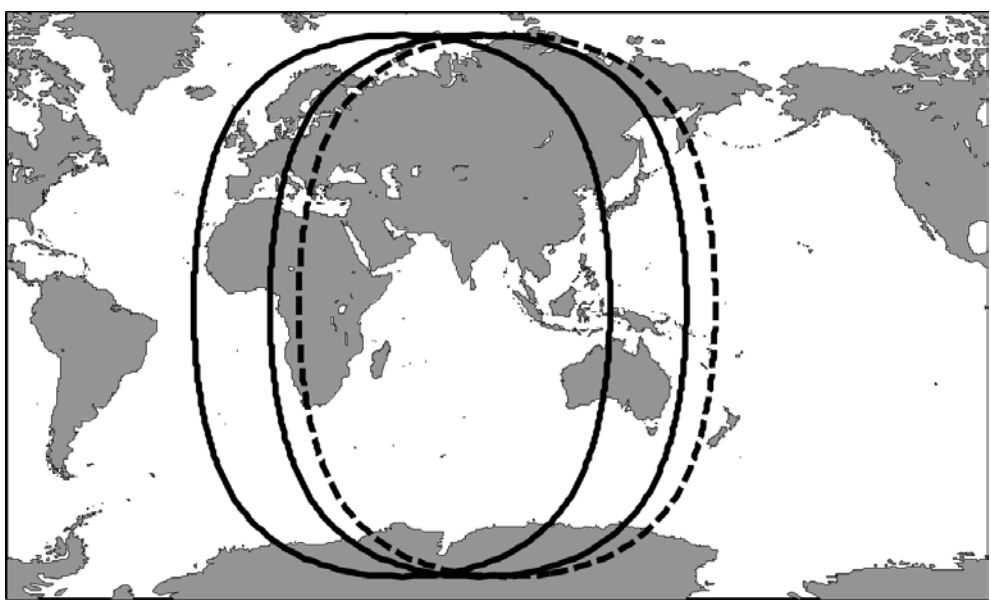


Figure 12.34 GAGAN GEO coverage.

Utilization by Non-aviation Users

Although the SBAS signal format described in this chapter was developed to support aeronautical requirements, the signal may also be used by non-aviation users with a suitable receiver. The vast majority of current SBAS users are not involved with aviation applications; many, if not most, low-cost, GNSS receivers include an SBAS reception capability.

Modernization

As mentioned earlier an L5 capability is being developed for SBAS. The new message format and associated services will be capable of supporting the use of the GPS L5 signal. Further, these messages will also support the use of the other core constellations. Users of this new service will combine signals at the L1 and L5 (or equivalent) frequencies so as to form the ionospheric-free combination. This allows users to directly eliminate the major effects of ionospheric delay. Ionospheric delay has the largest associated uncertainty for the current L1-only user. As a result, users will experience much smaller protection levels. Further, they will not require grid corrections that limit availability to the immediate regions around the reference stations. Users will experience a much reduced drop in availability as they move away from the reference stations.

The ability to use more than one constellation further reduces the protection levels and improves availability. Having extra satellites in the WLS solution greatly reduces the risk of experiencing poor geometries. By eliminating ionospheric delay effects and significantly increasing the number of corrected satellites, SBASs may be able to offer new services such as the ability to automatically land aircraft in certain weather conditions. The new L5 messages will allow for continued vertical service in more ionospheric conditions. The current L1 service can be limited in equatorial regions and during ionospheric storm events. This new service is still being developed and will require a significant effort to be able to utilize the new signals from GPS and from the other GNSS core constellations. A preliminary definition of the service is intended for review in 2018 and the full definition should be ready in 2022.

12.6.1.3 GBAS

In a Ground Based Augmentation System (GBAS), the GPS SPS is augmented with a ground reference station to improve the performance of the navigation services in the local area surrounding an airport. The ultimate objective for GBAS is to support all phases of flight within its area of coverage including the precision approach, landing, departure, and surface movement [66]. At the time of this writing, international standards had been developed by ICAO for GBAS through Category III precision approaches, although only standards through the less demanding Category I precision approaches have been validated and implemented.

As shown in Figure 12.35, GBAS is split into three separate segments: the space segment consisting of the GPS satellites, the ground segment or GBAS ground facility, and the airborne segment. Pseudorange corrections and correction rates are computed at the local reference station and broadcast to the airborne GBAS receiver via a communication link. In the aircraft the corrections and correction rates are applied to the local pseudorange measurements and used to obtain an improved position estimate.

RTCA published Minimum Aviation System Performance Standards (RTCA DO-245A) [66] for the U.S. version of GBAS (formerly referred to as the Local Area Augmentation System [LAAS]) describing the requirements allocation between ground facility and airborne avionics. RTCA also published Minimum Operational Performance Standards (RTCA DO-253C) [67] with airborne equipment

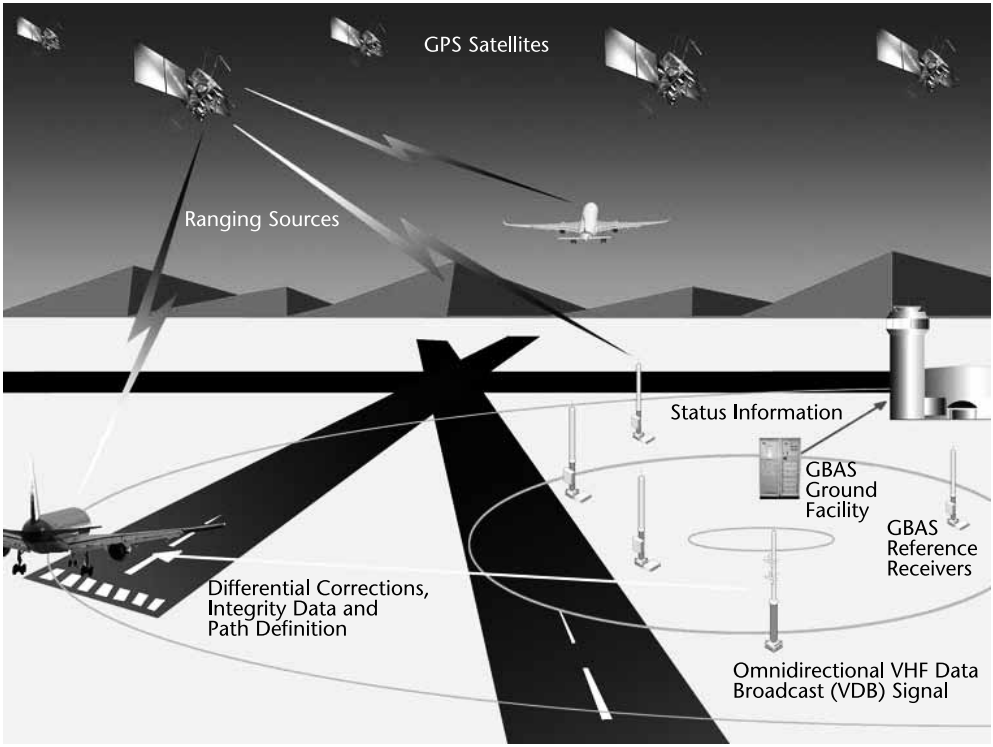


Figure 12.35 GBAS ground facility. (Courtesy of the Federal Aviation Administration.)

requirements and an Interface Control Document (RTCA DO-246D) [68] describing the communication link signal-in-space. The airborne standards and interface control document are being updated to support Category III precision approaches. These updates are anticipated to be completed in 2017. New standards for GBAS to support multiple GNSS constellations and dual-frequency airborne equipment are planned to be developed and validated by 2023.

Pseudorange Correction Computation

Originally, three GBAS alternatives were investigated [69]: single frequency (L1) carrier-smoothed code-phase DGPS [70, 71], kinematic dual-frequency carrier-phase GPS [72], and kinematic single-frequency carrier-phase with integrity beacons [73]. Eventually, carrier-smoothed code-phase DGPS was selected. The ground facility in the specified architecture reduces the noise component on the pseudoranges at each reference receiver (RR) by carrier-smoothing the code pseudorange measurements for each satellite. Carrier-smoothing can be implemented using:

$$\rho_{smooth}(k) = \frac{N-1}{N}\left[\rho_{smooth}(k-1) + \phi(k) - \phi(k-1)\right] + \frac{1}{N}\rho_{meas}(k) \tag{12.54}$$

where k = the time epoch, ϕ = the carrier-phase measurement, N = the number of measurements used for smoothing purposes

The carrier-smoothed pseudoranges are used to compute the pseudorange correction:

$$\Delta\rho_{sc,n,m} = R_{n,m} - \rho_{smooth,n,m} - t_{sv\_gps,n} \tag{12.54}$$

where R = predicted range, n = satellite index, m = RR index, and t_{sv_gps} = correction due to the satellite clock from the decoded GPS Navigation Data.

The broadcast correction can be computed from (12.54) following:

$$\Delta\rho_{corr,n} = \frac{1}{M_n}\sum_{m\in S_n}\left[\Delta\rho_{smooth,n,m} - \frac{1}{N_c}\sum_{n\in S_c}\Delta\rho_{smooth,n,m}\right]$$

where M_n = number of elements in set S_n, S_n = set of RRs with valid measurements for satellite n, N_c = number of elements in set S_c, and S_c = set of valid ranging sourcestracked by all RRs.

After reception and application of the broadcast ground facility corrections the three-dimensional aircraft position is calculated using a weighted least squares or equivalent algorithm [67].

Performance Requirements

The GBAS Approach Service Types (GAST) [67] are a set of GBAS performance and functional requirements that include navigation performance parameters such as accuracy, integrity, continuity, and availability. Furthermore, [60] defines the coverage area within which the GBAS service should be available. Table 12.6, based on [60, 61], assigns values to the navigation performance parameters accuracy and integrity for GAST-C covering Category I (CAT I) precision approaches. GAST-D will include some additional requirements to support operations through Categories II and III (CAT II/III). The GAST-D requirements are omitted from Table 12.6 as their validation and implementation have not yet been completed. The end-state goal for GBAS is to support all categories of precision approach and landings; CAT I, II, IIIa, and IIIb. Each of these landing categories is defined by the decision height (DH) at which the pilot or aircraft must make the decision to either continue or abort the landing. This decision depends on the runway visual range (RVR) at the corresponding decision height. Table 12.7 shows the DH and RVR for each of the categories.

Table 12.6 GBAS Performance Requirements for GAST-C

Accuracy	*Vertical position accuracy, 95%*	4.0m (NSE)
	Lateral position accuracy, 95%	16.0m (NSE)
Integrity	*Vertical alert limit (VAL)*	10m (200 ft HAT*)
	Lateral alert limit (LAL)	10m (200 ft HAT*)
	Time to alert	2 seconds
	Exposure time	15 seconds
	Allowable integrity risk	2×10^{-7}/approach

*Height above touchdown.

Table 12.7 Decision Height and Runway Visual Range for GBAS

Category	*DH*	*RVR*
CAT I	200 ft HAT	>2,400 ft
CAT II	100 ft HAT	> 1,200 ft
CAT IIIa	<100 ft HAT	>700 ft
CAT IIIb	<50 ft HAT	>150 m

*Height Above Touchdown (HAT) zone

Integrity Monitoring

GBAS includes an integrity monitoring function that determines, with a certain level of probability, that the code and carrier phase corrections do not contain misleading information. The integrity monitor function can be subdivided into multiple monitors: a signal quality monitor (SQM) to detect anomalous behavior in the satellite and pseudolite signals, a data quality monitor (DQM) to check if the satellite navigation data contains anomalies, a measurement quality monitor (MQM) to detect anomalies in the measurements such as pseudorange steps, a multiple reference consistency check (MRCC) to check the consistency of the corrections among the ground facility RRs, and a sigma monitor (SM) to check the nominal error characteristics of the ground facility.

Ground Facility Antennas, Airport Pseudolites, and Data Broadcast

The presence of ground multipath at the ground facility could introduce large errors in the airborne position and velocity computations. To mitigate the error due to ground multipath antennas can be designed that limit the multipath error. One example is the Integrated Multipath Limiting Antennas (IMLAs) [74]. To increase the availability of GBAS, at one time airport pseudolites (APLs) were envisioned [75]. APLs transmit a GPS-like signal that can be processed by the RRs and aircraft avionics in a similar fashion as the GPS signals. APLs eventually fell out of favor and were removed from GBAS standards years ago. The communications link used to transmit corrections from the ground facility to the GBAS avionics is referred to as the very high frequency (VHF) data broadcast (VDB).

12.6.2 Carrier-Based

In the past, geodetic positioning required line of sight connections to a network of monumented points in the ground. This geodetic network defined a consistent reference frame and helped control measurement error. Now a network of continuously operating GNSS receivers may replace the traditional geodetic network of monumented points. The network of receivers has an authoritative set of coordinates and also supplies base station carrier phase and code range data for accurate differential processing. Continuously operating networks are popular and numerous examples exist. We shall focus on two, the U.S. CORS and the global IGS system.

12.6.2.1 Continuously Operating Reference Stations (CORS)

The National Geodetic Survey (NGS) of the National Oceanic and Atmospheric Administration (NOAA) manages a CORS system to support non-navigation, post-processing applications of GNSS. GNSS receiver data are collected throughout the country and are archived at the main site in Silver Spring, Maryland, and at a parallel facility in Boulder, Colorado. The U.S. CORS system provides code range and carrier phase data from a nationwide network of GNSS receiver stations through the Internet. The CORS home page is at www.ngs.noaa.gov/CORS/.

The NGS makes use of stations established by other groups rather than by building an independent network of reference stations. In January 2016, the CORS network had over 1,900 operational stations run by 156 partners. CORS typically collect GNSS data 24 hours/day, 7 days/week, and are expected to conform to the guidelines at www.ngs.noaa.gov/CORS/Establish_Operate_CORS.shtml. NGS seeks a CORS station spacing of about 70 km.

The fundamental data of CORS are RINEX format (version 2.11) files containing dual-frequency carrier phase and pseudorange measurements. For many sites, Doppler data are also available. If supported by a receiver, the L1 pseudoranges derived from both C/A code (the C1 pseudorange) and the P(Y) code (the P1 pseudorange) are provided. As GNSS modernization progresses, triple-frequency data are becoming available. The principal translation package that converts the varieties of manufacturers' binary data into RINEX is program TEQC, maintained by UNAVCO. TECQ is documented at www.unavco.org/software/data-processing/teqc/teqc.html.

CORS reference coordinates and velocities are key values needed to use CORS as base stations for carrier-based differential GNSS applications. CORS coordinates and velocities are provided in two distinct reference frames, NAD 83 and ITRF08. The formal datum label for the CORS NAD 83 is NAD 83 (2011), and they are realized with an epoch of 2010.00. Stations in the Pacific are an exception, since they are on differing tectonic plates. These CORS are in the NAD 83 (PA11) or the NAD 83 (MA11) frames.

It should be noted that the CORS ITRF positions and velocities are established in a reference frame denoted IGS08.Users may treat IGS08 and ITRF08 as equivalent for most purposes. All published IGS08 positions and velocities carry a common datum tag IGS08 (2010.00), and are realized with the epoch date of 2010.00. The price of such ITRF global uniformity is that tectonic plate motion, as well as local motion, is expressed in the velocity values, and is seldom negligible for precision applications. More detail on the computation of CORS coordinates can be found at www.ngs.noaa.gov/CORS/coords.shtml. A conversion utility [Horizontal Time Dependent Positioning (HTDP)] between the NAD 83 and ITRF08 reference frames is located at www.ngs.noaa.gov/TOOLS/Htdp/Htdp.html.

Coordinate locations for a CORS antenna are referred to two different station reference points, the antenna reference point (ARP) and the L1 phase center (L1 PC). The ARP is defined as the center of the bottom-most, permanently attached, surface of the antenna. The L1 PC is a notional, electrical location for receipt of the L1 signal. Under most antenna designs the L1 (and L2) phase center varies with the elevation angle to a given GPS satellite. Establishment of an L1 PC origin

is done in conjunction with a companion model of the L1 phase center variation (www.ngs.noaa.gov/ANTCAL/). Due to the abstract character of the L1 PC and its dependence on specific calibration models, NGS considers the ARP as the definitive location for a CORS site. Also note that since June 30, 2012, the preferred antenna models are the absolute models.

Extensive metadata are also available for stations in CORS. This includes availability profiles, detailed data sheets and site logs, maps, photos, and time series of daily coordinate solutions. These metadata answer many questions on stability and reliability.

Other data that are not RINEX receiver data or metadata are available at the CORS site. In particular, both broadcast and precise GPS orbits can be obtained. The broadcast orbits are collected from the IGS global tracking network and do not show the satellite dropouts common to single site collections. The IGS orbits are combined products and include NGS contributions. Further discussion on IGS orbits is continued in the next section. For CORS sites that have a weather sensor, RINEX meteorological files are produced. As described earlier, users may obtain files and diagrams describing receiver antenna phase center offsets and phase center variation.

CORS RINEX data are stored in standardized directory locations. Users may access these files though the standard method, or through user-friendly CORS (UFCORS). Standard access is most readily obtained by clicking on the coverage map found on the CORS home page (www.ngs.noaa.gov/CORS/). Successive clicks on a location of interest will enable a user to zoom down to a specific site with a unique 4 character ID (e.g., GODE). A menu to the left will enable a user to select RINEX data or other metadata. A request for RINEX data will transfer the user from the map interface to the standard download interface. One must then again select the station of interest and the request for RINEX data and then add the year month and day of interest. This will lead the user to the directory holding the RINEX data.

The standard method is convenient for access of the various metadata. But, if one is interested solely in RINEX data, it can be obtained by directFile Transfer Protocol (FTP) access though a Web browser at ftp://geodesy.noaa.gov/cors/rinex/2017/ (for example). Multiple years are stored under the "rinex" directory, multiple days under the "year" directory, the various "site id's" under the "day" directories, and the RINEX files under the "site id" directories. A schematic of the FTP directory structure can be found at: ftp://geodesy.noaa.gov/cors/README.txt. Note that the FTP structure is most convenient for those software products that automatically download data for local processing.

In contrast to the standard access method, the UFCORS interface provides a customized file collection that is automatically compressed and downloaded to a user's computer (www.ngs.noaa.gov/UFCORS/). The user fills in a menu, indicating the desired block of time and the CORS site. Other options include receipt of the data sheet, IGS precise orbits, compression options, and alternative data rates. When selecting an alternative data rate, the collected data can be decimated to accommodate the desired target rate. The UFCORS interface frees the user from knowing specifics about the RINEX file storage system.

This section cannot be closed without addressing some of the support tools available from the CORS and NGS site. Earlier in this section, HTDP was discussed as a utility for conversion of coordinates between reference frames and epochs. A

dynamic map utility linked at the CORS home page allows one to build customized views of CORS coverage. A multifeatured Geodetic Toolkit at www.ngs.noaa.gov/TOOLS/ supports numerous online computations and coordinate conversions.

Special remarks must be made about the Online Positioning User Service (OPUS) tool (www.ngs.noaa.gov/OPUS/). This service allows users to upload anywhere from 15 minutes to 48 hours of dual-frequency RINEX data from a stationary antenna for automated, remote processing at NGS. In the static processing algorithm single baselines are computed and merged from three nearby CORS stations. In the rapid static algorithm the submitted data are processed simultaneously with the nearby CORS station data. In both cases, the results are e-mailed back to the user. Turnaround is typically just a few minutes. At its heart, OPUS uses the CORS as a subsystem in computation of the user's coordinates. This is suggestive of new directions and roles that continuously operating reference stations can take in the future.

12.6.2.2 International GNSS Service (IGS)

The International Association of Geodesy established the IGS in 1993 to support geodetic and geophysical research activities by providing GNSS data and products. The IGS serves a coordination role, sets standards and specifications, and encourages international adherence to its conventions. The IGS operates through a Governing Board and a Central Bureau (its executive arm) and functions through the cooperation of international groups of GNSS satellite tracking networks, data centers, analysis centers, and various working groups. The IGS home page is found at both www.igs.org/ and igscb.jpl.nasa.gov/.

IGS is known foremost as the source of precise GNSS orbits. However, IGS also produces GNSS satellite clock and ground receiver clock solutions, and Earth Orientation Parameter (EOP) products (polar motion, polar motion rate, and length of day). The IGS solutions are combined products that integrate the solutions generated by the individual analysis centers.

IGS products come in 3 varieties, with progressively greater latencies and accuracies. The ultra-rapid products are 48 hours in length. The first 27 hours are observed, and the remaining 21 hours are predicted. Ultrarapid products are produced four times a day, so that one can always utilize the early part of the prediction interval. Accuracies range from 3 to 5 cm (orbit) and 0.15 to 3 ns (clock). The rapid products have 17-hour latency, and are better than 2.5-cm orbit accuracy with 0.075-ns clock accuracy. The final combination products have a 12–18-day latency with slightly better orbit and clock accuracies.

The IGS products are organized by GPS week number and are available for FTP access through any browser. For example, ftp://igscb.jpl.nasa.gov/igscb/product/1881 contains precise orbit, clock, and EOP products for the week of January 24, 2016 (GPS week 1881). The nomenclature is described at: igscb.jpl.nasa.gov/components/dcnav/igscb_product_wwww.html. For example, "igr18810.sp3.Z" refers to the Rapid product, week 1881, day 0 (Sunday, January 24, 2016) orbit in the SP3 format, and compressed with a UNIX-compatible algorithm. Note that slightly different directory trees are used at the 5 sites that archive IGS products.

IGS also archives and disseminates GNSS continuously operating reference station (CORS) data globally. The master index, igscb.jpl.nasa.gov/components/data.

html, shows CORS data grouped by update intervals and archive sites. Directory paths vary between archive sites. For example, SOPAC data is mapped at http://igscb.jpl.nasa.gov/components/dcnav/sopac_rinex.html. Continuing the example, GPS observation CORS data for Sunday, January 24, 2016, can be found at ftp://garner.ucsd.edu/pub/rinex/2016/024/. The "2016" refers to the year, and the "024" refers to the day number for January 24. The file name, "algo0240.16d.Z," refers to ALGO, a site in Ontario, Canada, day number 024, a daily (not hourly) file, year 2016, GPS receiver data, and compressed with a UNIX-compatible algorithm. Not all CORS sites are found at all the IGS servers.

An overview map of the IGS international cooperative GNSS tracking network is found at igs.org/network and at igscb.jpl.nasa.gov/network/maps/allmaps.html. One can immediately note the global distribution of the sites. Some sites upload their data hourly, while others do so daily. The Web page at itrf.ensg.ign.fr/ITRF_solutions/2014/ITRF2014_files.php contains authoritative ITRF14 Cartesian coordinates and velocities for the IGS sites. Weekly network solutions are also available.

The IGS site has a rich set of resources, as would be expected from an international scientific operation. In addition to the items above, one may also find products for tropospheric zenith path delay and global grids of ionospheric Total Electron Content (TEC). Data are available from GPS sensors in LEO satellites. Publications are found at kb.igs.org/hc/en-us, mail archives are at igscb.jpl.nasa.gov/pipermail/igsmail, and analysis conventions are stored at www.iers.org/IERS/EN/Publications/TechnicalNotes/tn36.html.

12.6.3 PPP

As described in Section 12.4, due to the required initial position solution convergence period of a few tens of minutes in PPP, the technique has a specific user base for real-time positioning and navigation, and postprocessed positioning. At the same time, the great advantage of PPP over baseline processing approaches is that its corrections are generated from regional or global GNSS reference networks, which removes baseline limitations. As a result, PPP has become the ubiquitous method of precise positioning and navigation in remote areas of limited economic activity that would otherwise spur the development of continuously operating reference networks. Commercial uses include offshore positioning, precision agriculture, geodetic surveying, airborne mapping; and scientific applications include plate tectonics, seismic monitoring, tsunami warning, and precise orbit determination. Note that modern, conventional PPP was initially developed to significantly reduce the huge computational burden in the daily operational processing of data from large geodetic GNSS networks.

The following sections provide summary listings of current popular PPP services in two categories: free Internet-based PPP data processing services and commercial services. Clearly, there is significant research and commercial activity in the PPP arena as the technology is still maturing, so the presented material should be read with the caution that certainly services may cease operation, be altered, or new services may be introduced in the coming years.

12.6.3.1 Web-Based Services

Web-based services typically consist of a website with, in some cases, a free registration process. Subscribers submit RINEX-formatted GNSS observation files via http or e-mail. The GNSS observations files are processed by the service's PPP measurement processing engine, and the results then sent back via http or e-mail to the client in a range of timescales from near instantaneously to minutes or longer. Positioning results are accompanied by a variety of plots, analysis, and statistics, depending on the service provider. As PPP research advances, additional functionality can be found in some of these online processing services: dual-frequency GPS, single-frequency GPS, static data, kinematic data, additional ionospheric modeling, ambiguity resolution, GLONASS data processing, multiconstellation data processing, and nongeodetic receiver data processing. The services can be public, private, and academic, each sector with its own objectives and target user constituency. Note that other Web-based GNSS data processing services currently exist, but they make use of the network baseline rather than PPP processing technique and therefore are not listed here. A nonexhaustive list (in alphabetical order by operator) of Web-based PPP services is given in Table 12.8.

12.6.3.2 Commercial Services

Numerous commercial PPP services have been introduced over the past almost two decades, and a great deal of research and development activity and mergers and acquisitions has occurred. Service providers are either GNSS original equipment manufacturers (OEM) or precise positioning and navigation service providers from particular industrial sectors (e.g., offshore positioning or precision agriculture).

Table 12.8 Web-Based PPP Services

Web-Based PPP Service	*Operator*
PPP-Wizard	Centre National D'Etudes Spatiales
magicGNSS	GMV
APPS – The Automatic Precise Positioning Service of the Global Differential GPS (GDGPS) System	Jet Propulsion Laboratory, California Institute of Technology
Canadian Spatial Reference System Precise Point Positioning – CSRS-PPP	Natural Resources Canada
GAPS – GPS Analysis and Positioning Software	University of New Brunswick

Table 12.9 Commercial PPP Services

Global, Commercial Real-Time PPP Service	*Operator*
Starfix	Fugro
Atlas	Hemisphere GNSS
StarFire	NAVCOM
CORRECT	NovAtel
CenterPoint RTX	Trimble
Apex and Ultra	Veripos

Services typically consist of multifrequency, multiconstellation geodetic receivers and antennas that include communication satellite-based PPP correction reception capability, as well as PPP user processing engines. The customer pays for the user equipment as well as a subscription to the PPP corrections. Some service providers maintain their own global reference GNSS networks that are used to generate GNSS satellite orbit and clock corrections, and possibly satellite equipment delay correction and ionospheric and tropospheric corrections, whereas other service providers make use of corrections generated by their commercial partners. Most of the services offer a variety of global, real-time positioning/navigation products ranging from DGPS/DGNSS to baseline and network RTK to different versions of PPP. A nonexhaustive list (in alphabetical order by operator) of commercial PPP services is given in Table 12.9.

References

[1] Lapucha, D., and M. Huff, "Multi-Site Real-Time DGPS System Using Starfix Link: Operational Results," *ION GPS-92*, Albuquerque, NM, September 16–18, 1992.

[2] Brown, A., "Extended Differential GPS," *NAVIGATION: Journal of the Institute of Navigation,* Vol. 36, No. 3, Fall 1989.

[3] Kee, C., B. W. Parkinson, and P. Axelrad, "Wide Area Differential GPS," *NAVIGATION: Journal of the Institute of Navigation*, Vol. 38, No. 2, Summer 1991, pp. 123–145.

[4] Ashkenazi, V., C. J. Hill, and J. Nagle, "Wide Area Differential GPS: A Performance Study," *Proc. of the Fifth International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GPS-92)*, Albuquerque, NM, September 16–18, 1992, pp. 589–598.

[5] Montenbruck, O., and E. Gill, *Satellite Orbits: Models, Methods, Applications*, New York: Springer-Verlag, 2000.

[6] Remondi, B., "Using the Global Positioning System (GPS) Phase Observable for Relative Geodesy: Differential GPS 383 Modeling, Processing, and Results," Ph.D. Dissertation, Center for Space Research, University of Austin, Austin, TX, 1984.

[7] Counselman, C., and S. Gourevitch, "Miniature Interferometer Terminals for Earth Surveying: Ambiguity and Multipath with Global Positioning System," *IEEE Trans. on Geoscience and Remote Sensing*, Vol. GE-19, No. 4, October 1981.

[8] Greenspan, R. L., et al., "Accuracy of Relative Positioning by Interferometry with Reconstructed Carrier, GPS Experimental Results," *Proc. of the 3rd International Geodetic Symposium on Satellite Doppler Positioning*, Las Cruces, NM, February 1982.

[9] Teunissen, P. J. G., "Least Squares Estimation of Integer GPS Ambiguities," *IAG General Meeting*, Beijing, China, August 1993.

[10] Hatch, R., "The Synergism of GPS Code and Carrier Measurements," *Proc. of the 3rd International Symposium on Satellite Doppler Positioning*, New Mexico State University, Vol. 2, February 1982.

[11] Abidin, H., "Extrawidelaning for 'On the Fly' Ambiguity Resolution: Simulation of Multipath Effects," *Proc. of the ION Satellite Division's 3th International Meeting, ION GPS-90*, Colorado Springs, CO, September 1990.

[12] Paielli, R. A., et al., "Carrier Phase Differential GPS for Approach and Landing: Algorithms and Preliminary Results," *Proc. of the 6th International Technical Meeting, ION GPS-93*, Salt Lake City, UT, September 1993, pp. 831–840.

[13] van Graas, F., D. W. Diggle, and R. M. Hueschen, "Interferometric GPS Flight Reference/Autoland System: Flight Test Results," *NAVIGATION: Journal of the Institute of Navigation*, Vol. 41, No. 1, Spring 1994, pp. 57–81.

[14] Cohen, C. E., et al., "Real-Time Flight Testing Using Integrity Beacons For GPS Category III Precision Landing," *NAVIGATION: Journal of the Institute of Navigation*, Vol. 41, No. 2, Summer 1994.

[15] Cohen, C. E., et al., "Flight Test Results of Autocoupled Approaches Using GPS and Integrity Beacons," *Proc. of the ION Satellite Division's 7th International Technical Meeting*, Salt Lake City, UT, September 1994, pp. 1145–1153.

[16] Leick, A., *GPS Satellite Surveying*, 3rd ed., New York: John Wiley & Sons, 2004.

[17] van Graas, F., and M. Braasch, "GPS Interferometric Attitude and Heading Determination: Initial Flight Test Results," *NAVIGATION: Journal of the Institute of Navigation*, Vol. 38, No. 4,Winter, 1991-1992, pp. 297–316.

[18] Potter, J., and M. Suman, "Thresholdless Redundancy Management with Arrays of Skewed Instruments," *AGARD Monograph*, No. 224, NATO, Neuilly sur Seine, France, 1979.

[19] Walsh, D., "Real-Time Ambiguity Resolution While on the Move," *Proc. of the ION Satellite Division's 5th International Meeting, ION GPS-92*, Albuquerque, NM, September 1992, pp. 473–481.

[20] Chen, H. C., *Theory of Electromagnetic Waves: A Coordinate-Free Approach*, New York: McGraw-Hill, 1983.

[21] Golub, G. H., and C. F. Van Loan, *Matrix Computations*, 2nd ed., Baltimore, MD: The Johns Hopkins University Press, 1989.

[22] Frei, E., and G. Beutler, "Rapid Static Positioning Based on the Fast Ambiguity Resolution Approach 'FARA': Theory and First Results," *Manuscript Geodaetica*, Vol. 15, 1990.

[23] Hatch, R., "Instantaneous Ambiguity Resolution," *Proc. of the IAG International Symposium 107 on Kinematic Systems in Geodesy, Surveying and Sensing*, New York, September 10–13, 1990.

[24] Erickson, C., "An Analysis of Ambiguity Resolution Techniques for Rapid Static GPS Surveys Using Single Frequency Data," *Proceedings of The Institute of Navigation ION GPS '92*, Albuquerque, NM, September 1992, pp. 453–462.

[25] Teunissen, P. J. G., P. J. De Jonge, and C. C. J. M. Tiberius, "Performance of the LAMDA Method for Fast GPS Ambiguity Resolution," *NAVIGATION: Journal of the Institute of Navigation*, Vol. 44, No.3, Fall 1997, pp. 373–383.

[26] Braasch, M. S., "On the Characterization of Multipath Errors in Satellite-Based Precision Approach and Landing Systems," Ph.D. Dissertation, Department of Electrical and Computer Engineering (Avionics Engineering Center), Ohio University, Athens, OH, 1992.

[27] Wuebbena, G., "The GPS Adjustment Software Package GEONAP—Concepts and Models," *Proc. of the 5th International Geodetic Symposium on Satellite Positioning*, Las Cruces, NM, March 1989.

[28] Kiran, S., "A Wideband Airport Pseudolite Architecture for the Local Area Augmentation System", Ph.D. Dissertation, School of Electrical and Computer Engineering (Avionics Engineering Center), Ohio University, Athens, OH, 2003.

[29] Kuipers, J., *Quaternions and Rotation Sequences*, Princeton, NJ: Princeton University Press, 2002.

[30] Cohen, C. E., "Attitude Determination Using GPS: Development of an All Solid-state Guidance, Navigation, and Control Sensor for Air and Space Vehicles Based on the Global Positioning System," Ph.D. thesis, Stanford University, Stanford, California, December 1992.

[31] Cohen, C. E., "Attitude Determination," in *Global Positioning System: Theory and Applications, Vol. II*, B. Parkinson and J. J. Spilker, Jr., (eds.), Washington, DC: American Institute of Astronautics and Aeronautics, 1996.

[32] Anderle, R. J., "Point Positioning Concept Using Precise Ephemeris," *Proc. of the 1st International Geodetic Symposium on Satellite Doppler*, Las Cruces, NM, 1976, pp. 47–75.

[33] Héroux, P., and J. Kouba, "GPS Precise Point Positioning with a Difference," *Geomatics '95*, Ottawa, Canada, 1995.

[34] Zumberge, J. F., et al., "Precise Point Positioning for the Efficient and Robust Analysis of GPS Data from Large Networks," *Journal of Geophysical Research*, Vol. 102, 1997, pp. 5005–5017.

[35] Kouba, J., and P. Héroux, "Precise Point Positioning Using IGS Orbit and Clock Products," *GPS Solutions*, Vol. 5, No. 2, 2001, pp. 12–28.

[36] Wu, J. T., et al., "Effects of Antenna Orientation on GPS Carrier Phase," *Manuscripta Geodetica*, Vol. 18, 1993, pp. 91–98.

[37] McCarthy, D. D., and G. Petit, (eds.), *IERS Conventions (2003)*, International Earth Rotation and Reference Systems Service Technical Note No. 32, Frankfurt, 2004.

[38] Bisnath, S., and Y. Gao, "Current State of Precise Point Positioning and Future Prospects and Limitations," *International Association of Geodesy Symposia*, Vol. 133, 2009, pp. 615–623.

[39] Laurichesse, D., and F. Mercier, "Integer Ambiguity Resolution on Undifferenced GPS Phase Measurements and Its Application to PPP," *Proc. of ION GNSS 2007*, 2007, pp. 839–848.

[40] Collins, P., "Isolating and Estimating Undifferenced GPS Integer Ambiguities," *Proc. of ION National Technical Meeting 2008*, 2008, pp. 720–732.

[41] Ge, M., et al., "Resolution of GPS Carrier-Phase Ambiguities in Precise Point Positioning (PPP) with Daily Observations," *Journal of Geodesy,* Vol. 82, 2008, pp. 389–399.

[42] Teunissen, P. J., D. Odijk, and B. Zhang, "PPP-RTK: Results of CORS network-Based PPP with Integer Ambiguity Resolution," *Journal of Aeronautics, Astronautics and Aviation, Series A*, Vol. 42, 2010, pp. 223–230.

[43] Melbourne, W. G., "The Case for Ranging in GPS-Based Geodetic Systems," *Proc. of the 1st International Symposium on Precise Positioning with the Global Positioning System*, Vol. 1, U.S. Dept. of Commerce. Rockville, MD, April 15–19, 1985, pp. 373–386.

[44] Wübbena, G., "Software Developments for Geodetic Positioning with GPS Using TI 4100 Code and Carrier Measurements," *Proc. of the 1st International Symposium on Precise Positioning with the Global Positioning System*, Vol. 1, U.S. Dept. of Commerce. Rockville, MD, April 15–19, 1985. pp. 403–412.

[45] Geng, J., et al., "Rapid Re-Convergences to Ambiguity-Fixed Solutions in Precise Point Positioning," *Journal of Geodesy*, Vol. 84, No. 12, 2010, pp. 705–714.

[46] Leandro, R., et al., "RTX Positioning: The Next Generation of Cm-Accurate Real-Time GNSS Positioning," *Proc. of the 24th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2011)*, Portland, OR, September 2011, pp. 1460–1475.

[47] Special Committee 104, *RTCM Recommended Standards for Differential GNSS (Global Navigation Satellite Systems) Service*, Version 2.3 with Amendment 1, Radio Technical Commission for Maritime Services, Alexandria, VA, May 21, 2010.

[48] Special Committee 104, *RTCM Recommended Standards for Differential GNSS (Global Navigation Satellite Systems) Service*, Version 3.3, Radio Technical Commission for Maritime Services, Alexandria, VA, October 7, 2016.

[49] Kalafus, R., "The New RTCM SC-104 Standard for Differential and RTK GNSS Broadcasts," *Proc. of the Institute of Navigation ION GPS/GNSS 2003*, Portland, OR, September 2003, pp. 741–747.

[50] Boriskin, A., D. Kozlov, and G. Zyryanov, "The RTCM Multiple Signal Messages: A New Step in GNSS Data Standardization," *Proc. of The Institute of Navigation ION GNSS 2012*, Nashville, TN, September 2012, pp. 2947–2955.

[51] Laurichesse, D., and A. Blot, "Fast PPP Convergence Using Multi-Constellation and Triple-Frequency Ambiguity Resolution," *Proc. of The Institute of Navigation ION GNSS+ 2016,* Portland, OR, September 2016, pp. 2082–2088.

[52] **Chop, J., et al., "Local Corrections, Disparate Uses: Cooperation Spawns National Differ**ential GPS," *GPS World*, April 2002.

[53] *Broadcast Standard for the USCG DGPS Navigation Service*, COMDTINST M16577.1, United States Coast Guard, Washington, D.C., April 1993.

[54] Proakis, J., *Digital Communications*, 4th ed., New York: McGraw-Hill, 2001.

[55] Enge, P., and K. Olson, "Medium Frequency Broadcast of Differential GPS Data," *IEEE Trans. on Aerospace and Electronic Systems*, July 1990.

[56] DoD/DHS/DHS, *2014 Federal Radionavigation Plan*, U.S. Departments of Defense, Homeland Security, and Transportation, Washington, D.C., May 2015.

[57] Walter, T., and M. B. El-Arini, (eds.), *Global Positioning System: Papers Published in Navigation, Volume VI (SBAS)*, Fairfax, VA: Institute of Navigation, 1999.

[58] Braff, R., and C. Shively, "GPS Integrity Channel," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 32, No. 4, Winter 1985-1986.

[59] Kinal, G., and O. Razumovsky, "Upgrades to the Inmarsat PN Transmission Test Bed," *Proc. of The Institute of Navigation ION GPS '90*, Colorado Springs, CO, September 1990, pp. 315–322.

[60] Navigation Systems Panel (NSP), *Amendment 90 to the International Standards and Recommended Practices, Aeronautical Telecommunications (Annex 10 to the Convention on International Civil Aviation)*, International Civil Aviation Organization, Montreal, Canada, November 2015.

[61] RTCA Special Committee SC-159, *Minimum Operational Performance Standards for Global Positioning System/Satellite-Based Augmentation System Airborne Equipment*, RTCA/DO-229D with Change 1, Washington, D.C., RTCA, February 1, 2013.

[62] Nagle, J., A. J. Van Dierendonck, and Q. Hua, "Inmarsat-3 Navigation Signal C/A-Code Selection and Interference Analysis," *NAVIGATION: The Journal of The Institute of Navigation*, Vol. 39, No. 4, Winter 1992-1993.

[63] Hegarty, C., "Optimizing Differential GPS for a Data Rate Constrained Broadcast Channel," *Proceedings of The Institute of Navigation ION GPS '93*, Salt Lake City, UT, September 1993.

[64] DeCleene, B., "Defining Pseudorange Integrity—Overbounding," *Proc. of The Institute of Navigation ION GPS 2000*, September 2000.

[65] Soddu, C., and O. Razumovsky, "Inmarsat's New Navigation Payload," *GPS World*, November 1, 2001.

[66] Braff, R., "Description of the FAA's Local Area Augmentation System (LAAS)," *NAVIGATION: The Journal of the Institute of Navigation*, Vol. 44, No. 4, Winter 1997-1998.

[66] Special Committee 159, *Minimum Aviation System Performance Standards for Local Area Augmentation System (LAAS)*, RTCA DO-245A, December 2004.

[67] Special Committee 159, *Minimum Operational Performance Standards for GPS Local Area Augmentation System Airborne Equipment*, RTCA DO-253C, December 2008.

[68] Special Committee 159, *GNSS-Based Precision Approach Local Area Augmentation System (LAAS) Signal-in-Space Interface Control Document (ICD)*, RTCA DO-246D, December 2008.

[69] van Graas, F., "GNSS Augmentation for High Precision Navigation Services," *AGARD Lecture Series 207, System Implications and Innovative Applications of Satellite Navigation*, June 1996.

[70] van Graas, F., et al., "Ohio University/FAA Flight Test Demonstration of Local Area Augmentation System (LAAS)," *NAVIGATION: The Journal of the Institute of Navigation*, Vol. 45, No. 2, Summer 1998.

[71] Hundley, W., et al., "FAA-Wilcox Electric Category IIIB Feasibility Demonstration Program – Flight Test Results," *Proc. of the Institute of Navigation GPS-95*, September 12–14, 1995.

[72] Kaufmann, D., "Flight Test Evaluation of the E-Systems Differential GPS Category III Automatic Landing System," NASA Ames Research Center, Moffett Field, CA, September 1995.

[73] Cohen, C., et al., "Autolanding a 737 Using GPS Integrity Beacons," *NAVIGATION: The Journal of the Institute of Navigation*, Vol. 42, No. 3, Fall 1995.

[74] Thornberg, D., et al., "LAAS Integrated Multipath-Limiting Antenna," *NAVIGATION: The Journal of the Institute of Navigation*, Vol. 50, No. 2, Summer 2003.

[75] Bartone C., and F. van Graas, "Ranging Airport Pseudolite for Local Area Augmentation," *IEEE Trans. on Aerospace and Electronic Systems*, Vol. 36, No. 1, January 2000.

CHAPTER 13

Integration of GNSS with Other Sensors and Network Assistance

J. Blake Bullock and Mike King

13.1 Overview

In the previous chapters, we have observed that GNSS receivers can be thought of as discrete-time position/velocity sensors with sampling intervals of approximately 1 second. 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 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, a mathematical estimator [1]. A Kalman filter can provide optimal estimates of the instantaneous state of a linear system perturbed by Gaussian white noise, and update them in near real time or through postprocessing. One of the key attributes of the Kalman filter is that it provides a means of inferring information by the use of indirect measurements. It does not have to read control variable(s) directly, but it can read an indirect measurement (including associated noise) and estimate the control variable(s). In GNSS applications, the control variables as we will see later on in this chapter are position, velocity, time, and possible attitude errors. The indirect measurements are the GNSS pseudorange (PR) and pseudorange rate (PRR) and/or delta range.

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. With network-assistance, however, 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 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.

In military and other applications of GNSS/inertial integration, the inertial navigation system (INS) is generally viewed as the primary sensor, providing a reference trajectory which is immune to jamming and interference, with GNSS providing measurements which can be used to periodically update absolute position and reduce error growth. Such integrations are addressed in Section 13.2. However, in many commercial systems, this role can be reversed. In such systems, a very low cost set of inertial or other sensors or even digital maps can be used to augment GNSS (i.e., fill in the coverage gaps as might occur in an urban canyon). Such systems are addressed in detail in Section 13.3.

This chapter consists of four major sections beyond this overview. In Section 13.2, the motivations for GNSS/inertial integration are detailed. Because of the importance of the INS in this class of integrations, the discussion begins with a review of inertial navigation, including the error behavior of the inertial sensors as well as the full INS. 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. The implementation issues related to a GNSS/inertial integration for land vehicle applications are discussed. 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.

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.

13.2 GNSS/Inertial Integration

Navigation employing GNSS and inertial sensors is a synergistic relationship. The integration of these two types of sensors not only overcomes performance issues found in each individual sensor, but also produces a system whose performance exceeds that of the individual sensors. GNSS provides bounded accuracy, while inertial system accuracy degrades with time. Not only does the GNSS sensor bound the navigation errors, but the GNSS sensor can calibrate the inertial sensor. In navigation systems, GNSS receiver performance issues include susceptibility to interference from external sources, time to first fix (i.e., first position solution), interruption of the satellite signal due to blockage, integrity, and signal reacquisition capability. The issues related to inertial sensors are their poor long-term accuracy without calibration and their cost.

This section first discusses in more detail the relative weaknesses of GNSS (Section 13.2.1) and inertial sensors (Section 13.2.2) as outlined earlier. Next, an introduction to Kalman filtering is provided (Section 13.2.3), followed by a description of a variety of practical GNSS/inertial integrations and their performance features (Sections 13.2.4 through 13.2.6).

13.2.1 GNSS Receiver Performance Issues

One primary concern with using GNSS as a stand-alone source for navigation is signal interruption. Signal interruption can be caused by shading of the GNSS antenna by terrain or manmade structures (e.g., buildings, vehicle structure, and tunnels) or by interference from an external source. An example of signal interruption is shown in Figure 13.1. Each vertical line in this figure indicates a period of shading while driving in an urban environment. The periods of shading (i.e., less than three-satellite availability) are caused by buildings and are denoted by the black lines in the lower portion of Figure 13.1. (This experiment was conducted with a GPS-only receiver when five to six GPS satellites above a 5° mask angle were available for ranging.) When only three usable satellite signals are available, most receivers revert to a two-dimensional navigation mode by utilizing either the last known height or a height obtained from an external source. If the number of usable satellites is less than three, some receivers have the option of not producing a solution or extrapolating the last position and velocity solution forward in what is

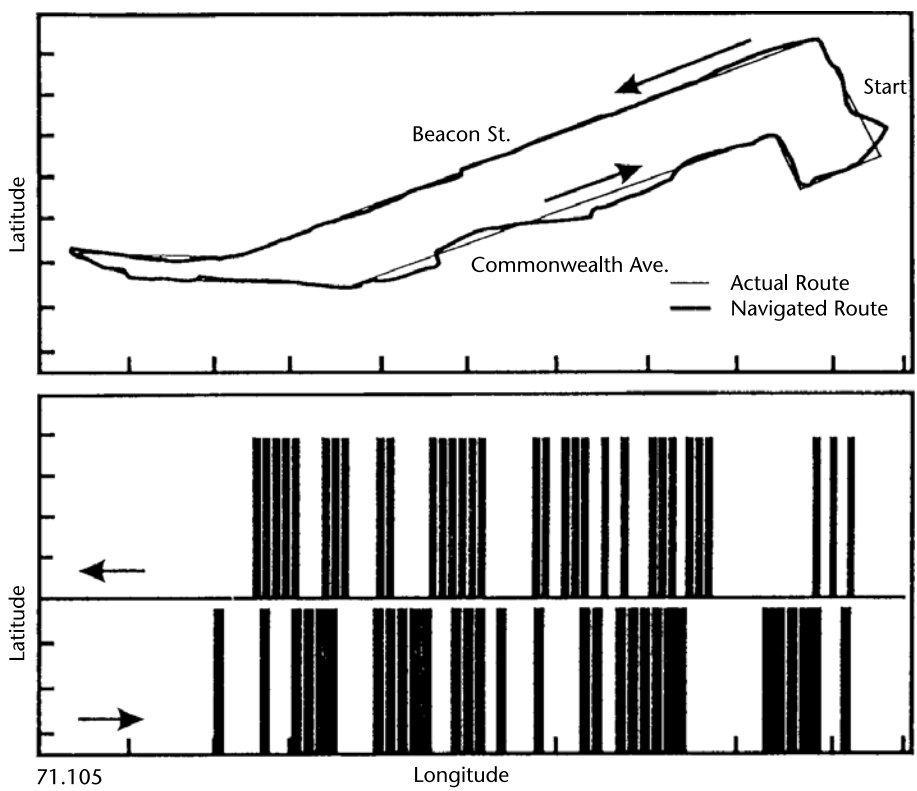


Figure 13.1 Effects of signal blockage on GNSS receiver operation.

called *dead-reckoning* (DR) navigation. INSs can be used as a flywheel to provide navigation during shading outages.

The discrete-time nature of the GNSS solution in some equipment is also of concern in real-time applications, especially those related to vehicle control. As shown in Figure 13.2, if a vehicle's path changes between updates, the extrapolation of the last GNSS measurement produces an error in the estimated and true position. This is particularly true for high-dynamic platforms such as fighter aircraft. In applications where continuous precision navigation is required, inertial sensors can be employed. An alternative solution is the use of a GNSS receiver that provides higher-rate measurement outputs. In principle, if phase lock on the L-band carrier is maintained, a nearly perfect velocity reference is available internal to the GNSS receiver through its carrier phase tracking output, which is maintained at a minimum rate of 50 Hz. Such a reference may require a customization of the receiver's output, and the carrier phase would need to be corrected for the rollover associated with wavelength transitions (i.e., exceeding 360°). The GNSS delta range measurement, discussed as a Kalman filter input later in this chapter, is constructed from a carrier phase difference.

In addition to providing navigation continuity during short GNSS shading outages and between GNSS sensor position outputs, an INS, when calibrated using a Kalman filter (see Section 13.2.3), can be used to improve the GNSS receiver performance in two other ways. First, the information that is maintained by the integration filter can be used to reduce the time to reacquire GNSS signals that have been lost through interference or obscuration; and second, the integration filter can be used to aid the receiver's tracking loops, extending the thresholds for signal tracking. Both techniques have been used since the very first GPS sets were designed [2]. The first enhancement, often referred to as prepositioning, computes an a priori estimate of a signal's code phase and Doppler using the integration filter's estimates of position and velocity, and time and frequency error. If the combined position and timing errors are less than one-half a chip [e.g., roughly 150m for the GPS C/A code, and 15m for the GPS P(Y) code], then nearly instantaneous reacquisition of a lost signal is possible, since the prepositioning limits the tracking error to the linear range of the loop's error detector (see Section 8.7). Similarly, Doppler on the signal to be reacquired can be predicted from the integration filter's estimates

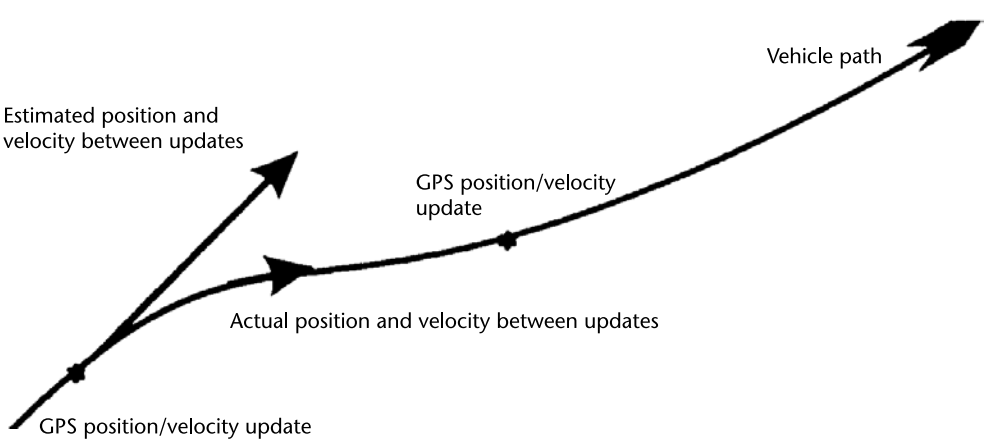


Figure 13.2 Extrapolation of a GNSS navigation solution in a dynamic environment.

of velocity and signal frequency, and if those estimates are within the linear range of the frequency error detector (see Section 8.6.1.3), nearly instantaneous signal acquisition may be possible. For example, if using an arctangent error detector with a 5-ms predetection integration interval (PDI), the combined frequency error can be as large as 50 Hz (see Section 8.6.1.3), which translates to a velocity accuracy of 10 m/s, which is readily achievable using a navigation grade inertial measurement unit (IMU), and potentially achievable with tactical grades [3]. Generally speaking, if the navigation filter is a robust design (i.e., its covariance matrix is consistent with the error in its navigation solution), then the uncertainty associated with the predicted code phase and Doppler can be determined from the filter's covariance matrix. For example, if $\mathbf{P}_4$ represents the 4×4 partition of the filter's covariance matrix corresponding to position and time error, then the error variance associated with a predicted code phase can be computed using:

$$\sigma_{cp}^2 = \mathbf{h}^{\mathrm{T}}\mathbf{P}_4\mathbf{h} \tag{13.1}$$

In (13.1), $\mathbf{h}$ is the filter's measurement gradient vector to the satellite of interest, comprised of the line-of-sight (LOS) unit vector to the satellite of interest (first three elements) and the unity sensitivity of the user's clock phase error (fourth element). Generally, the elements of the covariance matrix $\mathbf{P}_4$ in (13.1) are expressed in units of m^2. In this case, the code phase error variance σ_{cp}^2 will also be expressed in m^2. Given the error variance predicted by (13.1), suitable search ranges can be determined about the predicted code phase. Often, the search region is selected as three sigma, corresponding to $3\sigma_{cp}$. This ensures a high probability (roughly 99% under a jointly Gaussian assumption for the probability distribution) that the signal is within the selected search region. Figure 13.3 illustrates the two-dimensional nature of the search region for a code phase and frequency search. For the selected example, which corresponds to a relatively strong GPS P(Y) signal (42 dB-Hz carrier-to-noise ratio), prepositioning information has limited the search region to

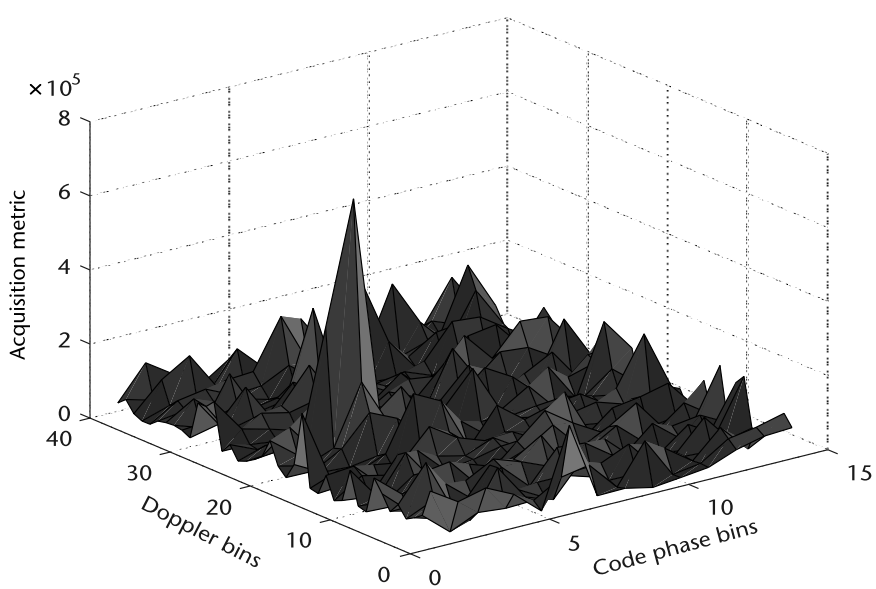


Figure 13.3 Prepositioning supports the acquisition/reacquisition process.

a 1 kHz Doppler uncertainty (which corresponds to roughly 40 frequency bins as illustrated) and a 200-m code uncertainty (corresponding to roughly 15 half-chips as illustrated). These uncertainty regions are symmetric about the expected code phase and Doppler, which corresponds to the center of the search space in the figure. The signal location in the search space is clearly visible, offset relative to its expected location, but nonetheless within the established region.

Prepositioning, as just described, offers the potential for drastically reducing the search space by collapsing both dimensions. If the number of cells remaining (i.e., subsets of the two-dimensional search region one-half chip by the chosen Doppler bin size) are less than or equal to the available number of receiver correlators, then parallel searching can be performed to reacquire the signal. This technique can also be used for initial acquisition of the GNSS signals using an INS that has been calibrated by other means.

As mentioned earlier in this section, the use of INS velocity outputs, corrected by the Kalman filter, can be used to extend signal tracking in adverse signal conditions. Fundamentally, any tracking loop performs three functions: attenuation of the noise in the observables that are passed to the Kalman filter; tracking of the dynamics of the host vehicle in which the receiver is installed; and finally, tracking of the dynamics of the receiver's oscillator. Use of INS aiding effectively removes the second requirement, enabling significant reduction of the tracking loop bandwidths, thus enabling tracking at lower signal-to-noise ratios. It is, in general, the requirement to track the dynamics of the receiver's oscillator that sets a floor on the bandwidth reduction and track extension. As discussed in Section 13.2.8, tracking loop aiding can be performed for both code and carrier tracking. The ability to extend code track can reach J/S levels approaching 75 dB, when used in combination with data aiding, with a higher quality IMU (e.g., high-end tactical grade) and a higher-quality reference oscillator within the GNSS receiver [4, 5].

An area of concern in the use of GNSS, especially in commercial aircraft applications, is integrity (see Section 11.4). An anomalous GNSS satellite signal most likely will result in the calculation of an erroneous position. The use of inertial components allows the GNSS pseudorange measurement to be compared against statistical limits (typically 3-sigma deviation) and reject those measurements that are beyond the limits. The components of the INS (i.e., gyros and accelerometers) can fail as well. Historically, the use of redundant INS or gyros and/or accelerometers has been used to increase reliability.

13.2.2 Review of Inertial Navigation Systems

13.2.2.1 Classes of Inertial Systems

Before addressing the sensors utilized in all inertial systems, a few remarks about the distinction between the two essential classes of inertial systems are needed. INSs can be broadly classified as either *gimbaled* or *strapdown* [6, 7]. The basic distinction between the two lies in the method by which the coordinate frame utilized for navigation is maintained: in gimbaled systems, the frame is mechanized physically by preserving a platform which is generally either the navigation frame itself or a frame related to the navigation frame by a known transformation (e.g., the azimuth in a *wander azimuth* mechanization of a gimbaled system). The platform is usually

kept *local-level* (i.e., level with respect to the horizon), where the accelerometers are able to directly sense the horizontal components of host vehicle acceleration. However, use of a *space-stable* gimbaled orientation (e.g., as was used for the Space Shuttle's inertial system) is an example of a gimbaled system which is not locally level. To summarize, in a gimbaled inertial system, the sensors are maintained in a preferred orientation, and generally isolated from the vehicle's changes in attitude. In a strapdown mechanization, on the other hand, the instruments are fixed in the vehicle (e.g., along the nose of an aircraft, out the left wing, and with third axis completing the set). The navigation frame is maintained mathematically, not physically, by the calculation of a transformation between the vehicle's body frame (where the instruments reside) and the navigation frame: this transformation is most commonly referred to as a direction cosine matrix, but its mechanization is usually as a quaternion or rotation vector [6, 7] for improved efficiency.

The relative advantages and disadvantages of the two types of systems are fairly well known. The gimbaled systems tend to be more expensive, owing to the additional hardware required for maintaining the physical platform, while the computational requirements for the strapdown system (largely for maintenance of the direction cosine matrix) are higher. Historically speaking, gimbaled systems were used almost exclusively decades ago in navigation systems where accuracy was a significant driver, while strapdown systems were relegated to applications with very short flight times (e.g., a missile interceptor problem). However, advances in microprocessor and inertial sensor technology have changed this trend, making strapdown inertial systems the selection in most applications. Microprocessor improvements have made the high-rate computation of the direction cosine matrix relatively easy, and the advent of optical gyros (i.e., ring-laser and fiber optic) have produced designs without the significant acceleration sensitivity of their mechanical counterparts. This is quite important since the strapdown sensors see the full vehicle dynamics, which leads to additional errors relative to their gimbaled counterparts in high dynamic applications. Strapdown inertial navigation systems will be the focus of the remainder of the discussion of inertial navigation, as it has the widest use in GNSS/INS systems. Consult [7] for a more thorough treatment of strapdown navigation.

13.2.2.2 Inertial Navigation System Sensors

Returning now to the inertial sensors, there are two types, *gyroscopes* and *accelerometers*. The output of a gyroscope is a signal proportional to angular movement about its input axis ($\Delta\theta$) and the output of an accelerometer is a signal proportional to the change in velocity sensed along its input axis (Δv). Both gyroscopes and accelerometers can also be designed to sense more than one axis of angular velocity or acceleration: each can therefore be referred to as either Single Degree of Freedom (SDOF) or Two Degree of Freedom (TDOF) sensors. A three-axis IMU would then require three SDOF gyroscopes and three SDOF accelerometers to determine position and velocity in three dimensions.

Errors in the gyro sensed inertial angular velocity can generally be expressed by (13.2), which summarizes the impact of misalignments, scale factor error, bias, and acceleration sensitivity:

$$\boldsymbol{\delta\omega} = \mathbf{M}_{gyr}\boldsymbol{\omega} + \mathbf{b}_{gyr} + \mathbf{Ga} \tag{13.2}$$

where $\boldsymbol{\delta\omega}$ is the three-dimensional angular velocity error vector; $\boldsymbol{\omega}$ is the three-dimensional angular velocity vector; $\mathbf{b}_{gyr}$ is the three-dimensional gyro bias; $\mathbf{a}$ is the three-dimensional vector of accelerations in sensor axes; $\mathbf{M}_{gyr}$ is a matrix of gyro scale factor errors and misalignments; and $\mathbf{G}$ is a matrix of acceleration sensitive error effects.

The diagonal elements of $\mathbf{M}_{gyr}$ represent the scale factor errors, while the off-diagonal elements represent misalignments. The misalignments are composed of the misalignments of each of the individual sensors relative to the IMU case and the misalignments of the case installed on its host vehicle. The sensor axis misalignments are generally controlled and specified by the vendor and represent six uncorrelated error components. The installation misalignments can be represented as three orthogonal rotations and often are much larger than the internal misalignments of the sensors relative to the case. Scale factor asymmetry can be significant and warrant inclusion in any high fidelity model: asymmetry models are generally sensor specific and can be proportional to the absolute value of angular velocity for some designs or proportional to angular velocity squared for other designs. The gyro bias components may be represented as bounded, time-correlated random processes with a defined correlation time (i.e., a Markov process) or an initial bias level driven by white noise (i.e., a random walk), or a bias level driven by a second bias component (i.e., a random ramp). There also are temperature sensitivities of each of these error sources which are generally compensated by the vendor to error levels which are not significant. Additional sources of error that can be significant include g^2 sensitivity (an error term which is proportional to products of accelerations across axes), vibration rectification error (a bias component proportional to vibration), anisoinertial effects proportional to products of angular velocity components across axes, and angular acceleration sensitivities.

Errors in the accelerometer sensed specific force can be summarized by (13.3), which summarizes the impact of misalignment and scale factor error and bias:

$$\delta\mathbf{f} = \mathbf{M}_{acc}\mathbf{f} + \mathbf{b}_{acc} \tag{13.3}$$

where $\delta\mathbf{f}$ is the three-dimensional specific force error vector; $\mathbf{f}$ is the three-dimensional specific force vector; $\mathbf{b}_{acc}$ is the three-dimensional accelerometer bias; and $\mathbf{M}_{acc}$ is a matrix of accelerometer scale factor errors and misalignments.

Note the use of specific force, in place of acceleration, since the sensor cannot separate inertial and gravitational acceleration components. The diagonal elements of $\mathbf{M}_{acc}$ represent the scale factor errors, while its off-diagonal elements represent misalignments. The misalignments are comprised of the misalignments of each of the individual sensors relative to the IMU case and the misalignments of the case installed on its host vehicle. The sensor axis misalignments are generally controlled and specified by the vendor and represent six uncorrelated error components. The installation misalignments can be represented as three orthogonal rotations and often are much larger than the internal misalignments of the sensors relative to the case. Scale factor asymmetry can be significant and warrant inclusion in any high fidelity model: asymmetry models are generally sensor specific, proportional

to the absolute value of specific force for some designs, or proportional to specific force squared for other designs. The accelerometer bias components may be represented as bounded, time-correlated random processes with a defined correlation time (i.e., a Markov process) or an initial bias level driven by white noise (i.e., a random walk), or a bias level driven by a second bias component (i.e., a random ramp). There also are temperature sensitivities of each of these error sources which are generally compensated by the vendor to error levels which are not significant. Additional sources of error which can be significant include g^2 sensitivity (an error term which is proportional to products of accelerations across axes), vibration rectification error (a bias component proportional to vibration), angular velocity cross-product effects scaled by error offsets of each sensitive element relative to the instrument center, and angular acceleration sensitivities, again scaled by error offsets of the sensitive element.

The upper curves of Figure 13.4 show the performance of three classes of inertial sensors. (Note that CEP is an indicator of delivery accuracy. It is the radius of a circle in which 50% of the projectiles are expected to fall within the given radius. See Section 11.2.3.) When these systems are integrated with GNSS, the lower curve dictates the performance of the integrated GNSS/inertial (GNSSI) system. Therefore, during operation of a navigation system when both GNSS and inertial components are operational, the inertial navigation errors are bounded by the accuracy of the GNSS solution.

One significant contribution the GNSS receiver makes to the operation of the inertial subsystem is the calibration of the inertial sensors (see Figure 13.5). (Note that MRE is another indicator of delivery accuracy. Mean radial error is the mean of the miss distance of all projectiles.) Inertial instruments are specified to meet a turn-on to turn-on drift requirement. (Each time a gyro is powered up, its initial drift rate differs.) The major errors are gyro and accelerometer bias, which are typically six of the states within an inertial or GNSSI Kalman filter, which is discussed later in the section.

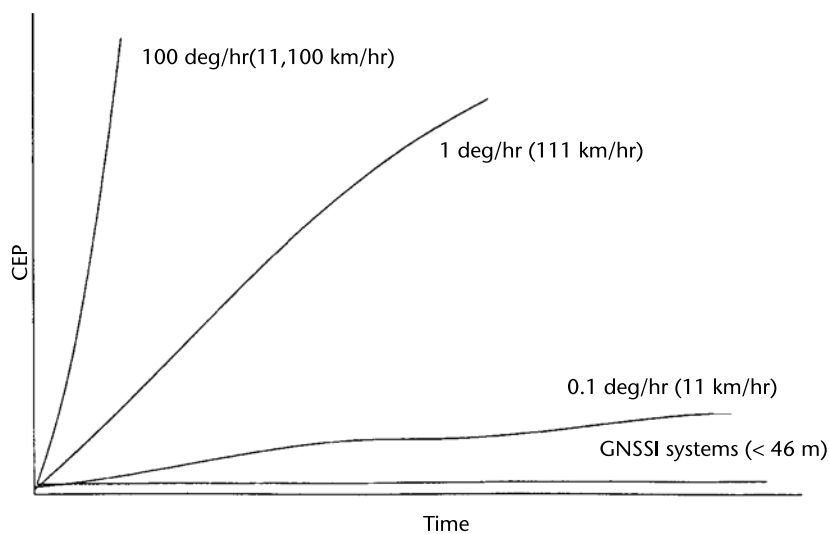


Figure 13.4 Comparison of navigation accuracies.

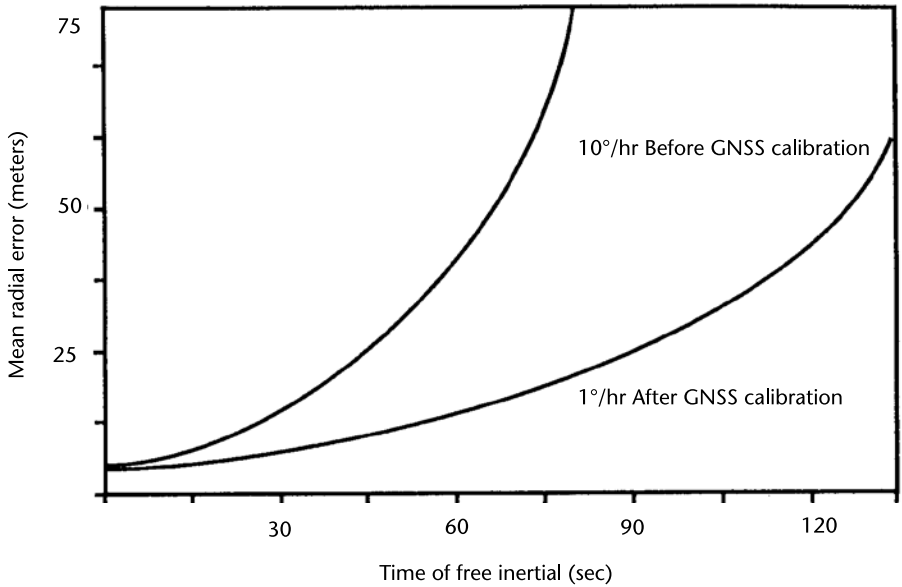


Figure 13.5 Inertial navigation before and after GNSS calibration.

13.2.2.3 Inertial Navigation System Error Behavior

Inertial navigation systems can be further classified as tactical, navigation, or strategic grades, as summarized in [3]. Levels of gyro and accelerometer bias and scale factor error levels can be used to distinguish the grade: for the tactical grade, gyro bias levels generally exceed 1°/hr, and can reach more than 1,500°/hr for commercial designs, while scale factor errors generally exceed 100 PPM and can reach more than 1000 PPM. For navigation grades, gyro bias levels generally exceed 0.001°/hr, and can reach 1°/hr, while scale factor errors generally exceed 10 PPM and can reach more than 100 PPM. Finally, the expected gyro performance for strategic grades is better than their navigation-grade counterparts. For the tactical grade, accelerometer bias levels generally exceed 1 mg, and can reach more than 10 mg for commercial applications, while scale factor errors generally exceed 100 PPM and can reach more than 1000 PPM. For navigation grades, accelerometer bias levels generally exceed 1 μg, and can exceed 100 μg, while scale factor errors generally exceed 1 PPM and can reach more than 10 PPM. Finally, the expected accelerometer performance for strategic grades is better than their navigation-grade counterparts.

INS error performance has traditionally been characterized using a nautical mile per hour rating. Navigation-grade INS are expected to be in the neighborhood of 1 nm/hr, with corresponding levels at 0.02 and 20 nm/hr for strategic and tactical grades, respectively. While such ratings have merit for an INS operating alone, they are generally not useful when characterizing integrated GNSS performance. The traditional nautical mile per hour rating corresponds to the position error drift corresponding to the effective level axis gyro bias levels; for example, a 1°/hr effective "east" gyro bias produces a Schuler oscillation (84-minute period INS error dynamic reviewed later in this section) about a bias level in the north component of INS velocity error. This bias level is the gyro bias, converted to radians per hour, scaled by the radius of the Earth, which produces the ramp in position error over

time periods on the order of hours (longer-term INS error dynamics give rise to error effects with roughly a 24-hour period, which reduces the longer-term effective position drift rate). A 1-nm/hr rating corresponds roughly to a gyro bias of 0.016°/hr [3], which matches the scaling by the Earth radius. GNSS outages are generally not hours in duration, and the short-term error propagation of an INS can exceed that implied by the nautical mile per hour rating, as discussed later in this section.

13.2.2.4 Inertial Navigation System Error Dynamics

INS error dynamics have been analyzed for decades, and many textbooks have been dedicated to the subject [8–10]. The treatment herein will therefore simply review the inherent error dynamics without deriving them. All derivations rely upon a fundamental assumption that the INS attitude errors are sufficiently small such that they can be treated as three-dimensional vectors and that sines of attitude error components can be replaced by the attitude error in radians, while their cosines can be replaced by unity. The error dynamics can then be placed in the following general form:

$$d\mathbf{x}/dt = \mathbf{F}\mathbf{x} + \mathbf{b} \tag{13.4}$$

where $d\mathbf{x}/dt$ represents the time rate of change of the error vector $\mathbf{x}$; $\mathbf{x}$ is a nine-dimensional vector that represents three components of INS position error, followed by three components of INS velocity error, followed by three components of INS attitude error; $\mathbf{F}$ is a nine-by-nine matrix which expresses the (unforced) error dynamics of the INS; and $\mathbf{b}$ is a nine-dimensional forcing function which represents the effects of gyro and accelerometer error (including gravity modeling error).

In representing (13.4), the error vector $\mathbf{x}$ will be expressed in a navigation frame, denoted by the superscript *N*. This frame could be selected as a number of local-level frames (see Section 2.2.3), for example, East North Up (ENU), North East Down (NED), or a frame in which the local level frame is not fixed in its azimuth direction (e.g., wander azimuth or free azimuth coordinates [7]). Alternatively, because of the emphasis on GNSS integration, an Earth Centered Earth Fixed (ECEF) frame (see Section 2.2.2) is frequently selected. The GNSS satellite position and velocity calculations are optimized for an ECEF frame, and so there are computational savings associated with representing the INS error dynamics in this frame. This frame will be the basis of the more detailed equations that follow. Returning to (13.4), the dynamics matrix $\mathbf{F}$ will dictate how initial position, velocity, and attitude errors propagate over time, and so enables determination of the frequency content of the underlying error dynamics (i.e., Schuler and Earth rate dynamics referenced above) and the instability associated with the vertical channel, to be addressed later. The forcing vector $\mathbf{b}$ has the following form in a navigation frame:

$$\mathbf{b}^{NT} = \left[0^T \mathbf{b}_v^T \mathbf{b}_\varphi^T\right] \tag{13.5}$$

where T denotes vector transpose; 0 denotes a zero vector; $\mathbf{b}_v$ denotes a vector forcing the velocity error equations; and $\mathbf{b}_\varphi$ denotes a vector forcing the attitude error equations.

Because gyro and accelerometer errors live in the Sensor (S) frame in a strapdown mechanization, $\mathbf{b}_v$ and $\mathbf{b}_\varphi$ can be expressed in the following form:

$$\mathbf{b}_v = {}_N\mathbf{C}_S\delta\mathbf{f} + \delta\mathbf{g}^N \tag{13.6}$$

$$\mathbf{b}_\varphi = {}_N\mathbf{C}_S\boldsymbol{\delta\omega} \tag{13.7}$$

where ${}_N\mathbf{C}_S = {}_N\mathbf{C}_B\mathbf{C}_S$, and ${}_N\mathbf{C}_S$ denotes the direction cosine matrix from the S (sensor) frame to the N (navigation) frame; ${}_B\mathbf{C}_S$ denotes the fixed transformation between B (body) and sensor frames; and $\delta\mathbf{g}^N$ represents gravity modeling error, best represented in the N frame.

Before presenting the general form of the **F** matrix and reviewing what it reveals about the frequency content of the underlying INS error dynamics, it is worth noting that even gyro and accelerometer biases which are relatively constant can produce time-varying error effects in the navigation frame as the orientation between body and navigation frames changes (i.e., the modulating effect of the direction cosine matrix ${}_N\mathbf{C}_B$). The general form for **F** matrix, again with the navigation error vector represented in a navigation frame, is most easily represented and understood in terms of its nine 3×3 partitions, given by the following equations:

$$\begin{aligned}
&\mathbf{F}_{11} = [0], \mathbf{F}_{12} = \mathbf{I}, \mathbf{F}_{13} = [0] \\
&\mathbf{F}_{21} = \mathbf{F}_{vp}, \mathbf{F}_{22} = \mathbf{F}_{vv}, \mathbf{F}_{23} = [\mathbf{f}\mathrm{X}] \\
&\mathbf{F}_{31} = [0], \mathbf{F}_{32} = [0], \mathbf{F}_{33} = \mathbf{F}_{\varphi\varphi}
\end{aligned} \tag{13.8}$$

where [0] represents a 3×3 matrix of zeroes; **I** represents the 3×3 identity matrix; and [**f**X] represents the 3×3 skew symmetric matrix constructed using the components of the specific force vector.

Before presenting the equations for the remaining partitions of **F**, two comments are worth making: first, the $\mathbf{F}_{12}$ partition implies that the INS ECEF position frame components are simply driven by the corresponding INS ECEF velocity error components, a simplification resulting from the use of ECEF coordinates (the $\mathbf{F}_{11}$ partition is nonzero for local level frame coordinate options). Second, the $\mathbf{F}_{23}$ partition, which couples the INS attitude error into the rate of change of INS velocity error, is largely what enables calibration of the gyro biases through processing of GNSS range rate information. Here, unfortunately, the use of ECEF coordinates obscures the message about the underlying error dynamics: if expressed in a local level frame, it becomes clear that gravitational acceleration alone enables estimation of the attitude errors about local axes, while some kind of maneuver in a horizontal plane is necessary to observe heading error. Such sensitivities are not as apparent when using an ECEF representation.

The $\mathbf{F}_{vv}$ partition represents a Coriolis coupling between the velocity components related to the rotational rate between ECEF and inertial frames:

$$\mathbf{F}_{vv} = [\Omega\mathbf{x}] \tag{13.9}$$

where $\boldsymbol{\Omega}$ is the vector representing earth rate in an ECEF frame.

The $\mathbf{F}_{vp}$ partition represents the impact which INS position error has upon the gravity modeling error component of acceleration error, and gives rise to the vertical channel instability, producing an exponentially divergent altitude error in all inertial navigators using a gravity model with an altitude dependence. Stated an alternate way, the employed gravity model is used to correct the vertical acceleration component of the transformed accelerometer outputs. A positive altitude error results in a gravity compensation error which is too small, leading to a positive vertical acceleration error which, through two integrations, leads to an even greater altitude error.

$$\mathbf{F}_{vp} = -\left(g_0/R_e\right)\left[\mathbf{I} - 3\mathbf{u}_h\mathbf{u}_h^T\right] \tag{13.10}$$

where g_0 is the gravitational acceleration at zero altitude; R_e is the Earth's equatorial radius; and $\mathbf{u}_h$ denotes a unit vector in the vertical direction expressed in ECEF coordinates.

Finally, the $\mathbf{F}_{\varphi\varphi}$ partition of the $\mathbf{F}$ matrix also simplifies in the ECEF frame:

$$\mathbf{F}_{\varphi\varphi} = \left[\Omega x\right] \tag{13.11}$$

The full 9×9 $\mathbf{F}$ matrix is now complete. Recall it represents the unforced error dynamics of the INS: as such, it is of interest to consider its frequency content, as referenced earlier. If the INS is nonaccelerating, the Laplace transform of the unforced dynamic equation can be used to examine the underlying frequency content of the errors [8]:

$$s\mathbf{X}(s) = \mathbf{F}\mathbf{X}(s) + \mathbf{W}(s) \tag{13.12}$$

$$\mathbf{X}(s) = \left(s\mathbf{I} - \mathbf{F}\right)^{-1}\mathbf{W}(s) \tag{13.13}$$

where $()^{-1}$ denotes matrix inversion.

The determinant of the matrix inverted in (13.13) gives the desired characteristic equation, which, once factored, indicates the presence of an Earth rate (i.e., 24-hour period) oscillation, an oscillation at the undamped Schuler frequency referenced earlier (84-minute period), and a damped Schuler oscillation at a Foucalt frequency (Earth rate scaled by the sine of latitude). Noting the frequency content of the errors motivates consideration of approximations to the INS error equations, which leads to useful insights, particularly in understanding error behavior over relatively short (i.e., relative to the Schuler dynamics) periods of GNSS loss. Consult [7] for a derivation of the "medium-term" and "short-term" approximations to the INS error dynamics. The short-term dynamics apply to outages of 10 minutes or less, and lead to the following simplifications of the impact of gyro bias and accelerometer bias on INS position error growth:

$$\delta p = b_{acc}\Delta t^2/2 \tag{13.14}$$

$$\delta p = g b_{gyr} \Delta t^3 / 6 \tag{13.15}$$

where b_{acc} and b_{gyr} represent accelerometer and gyro biases, g represents gravitational acceleration, and Δt represents the duration of the GNSS outage.

Equations (13.14) and (13.15) can provide insights into error growth between GNSS updates. Note these error effects can greatly exceed the nautical mile per hour rating of the INS which characterizes the error behavior over multiple Schuler periods: a residual (i.e., uncalibrated) accelerometer bias produces a quadratic position error growth in the short term, while a residual gyro bias produces a cubic error growth. This completes our discussion of INS error dynamics, and the design of a Kalman integration filter can now be reviewed in this light.

13.2.3 The Kalman Filter as System Integrator

13.2.3.1 Review of Kalman Filtering

The Kalman filter readily satisfies the requirements for GNSS integration with an INS: as was just discussed, the error equations associated with the INS error propagation fit into a linear state space model which matches the dynamic model assumed by the Kalman filter. Similarly, the periodic GNSS pseudorange and pseudorange rate or delta range measurement updates available from the receiver are well represented by the measurement model of the Kalman filter. Before reviewing the filter at an equation level, some high-level remarks are appropriate. Under certain conditions relating to the underlying error dynamics [11], it is an optimal estimator in the sense that it minimizes the mean square error in its state estimates. In addition, the Kalman filter is unique among filters in that it carries an estimate of its accuracy in the form of a covariance matrix. As will be discussed later, this feature permits a level of robustness relative to erroneous measurement updates. However, inadequate statistical modeling can significantly degrade filter performance and even lead to divergence of the filter, which will also be addressed.

The Kalman filter algorithm is presented in Figure 13.6. The filter is a mathematical algorithm to produce estimates of the *state vector* $\mathbf{x}$ at discrete epochs of time (indexed by subscript k) using a vector of noisy measurements $\mathbf{z}$ with (possibly time-varying) covariance $\mathbf{R}$ that is assumed to be available at each epoch. In general, the state vector $\mathbf{x}$ is the set of variables of interest (e.g., INS errors in some integrated navigation systems) and $\hat{\mathbf{x}}$ denotes the filter's estimates of the state vector. For INS state vector variables, the filter's dynamic model follows that of the INS error dynamics, as summarized in (13.4). As a minimum, unless the INS linearized error dynamics as expressed by (13.4) are approximated (using, e.g., the short-term or medium-term representations), a minimum of nine states are required to adequately represent them. Selection of additional states is a function of the desire for real-time calibration of the INS when GNSS is available, and the dynamics of the host: a hierarchy of possible state selection is reviewed in the next section. Since the Kalman filter algorithm necessarily operates in discrete time, the dynamics matrix $\mathbf{F}$ can be assumed piece-wise constant and well approximated over discrete time intervals by the corresponding Φ matrix:

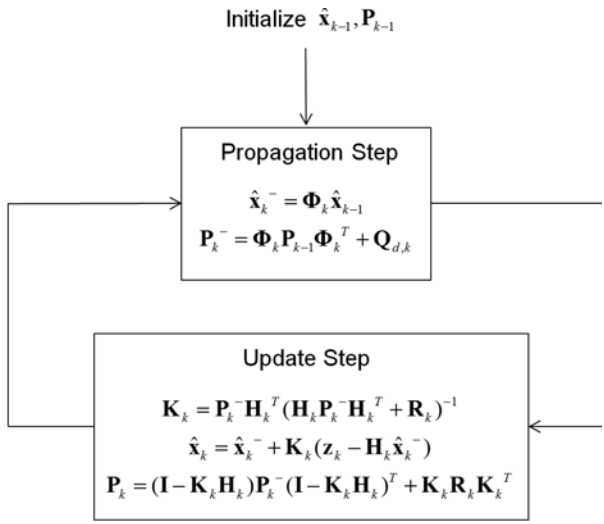


Figure 13.6 Kalman filter processing architecture

$$\Phi = \mathbf{I} + \mathbf{F}\Delta t + \mathbf{F}^2 \Delta t^2 / 2 + \ldots \tag{13.16}$$

where $\mathbf{I}$ represents the $n \times n$ identity matrix; n is the dimension of the state vector; and Δt represents the propagation interval.

Depending on the severity of the host's dynamics, an approximation to Φ of first or second order in Δt is generally adequate. As an alternative to including more terms in the expansion of Φ, the size of Δt can be reduced, resulting in multiple propagation steps of the filter for each measurement update step in Figure 13.6.

In addition to its use in state propagation, the Φ matrix is used to propagate the covariance matrix $\mathbf{P}$, as indicated in Figure 13.6. The covariance matrix, following initialization, tracks the uncertainty of the filter in each of its estimates: for example, a value of 4.0 m^2 in the first diagonal element of the covariance matrix implies that the filter believes it has estimated the x component of position error (in an ECEF frame) to 2m, one sigma. In addition to the term that propagates the covariance matrix through the modeled INS error dynamics embedded in Φ, the discrete process noise ($\mathbf{Q}_d$) matrix is added to the propagation equation as a means of representing the increase in state uncertainties induced by unmodeled dynamics. The unmodeled dynamics include the effects of error states that have not been included in the state vector, in addition to sources of error which are best modeled as truly random effects. Random effects generally include the velocity random walk associated with each accelerometer and the angle random walk associated with each gyro, at least for tactical grade IMUs. Finally, the process noise matrix is used to model state dynamic components, which are represented as random variables, for example, a gyro bias that is represented as a random walk. Note that the $\mathbf{Q}_d$ matrix represented in Figure 13.6 is the discrete process noise, which can be determined from its continuous time counterpart in a manner analogous to the relationship between $\mathbf{F}$ and Φ:

$$\mathbf{Q} = \int \Phi \mathbf{Q} \Phi^{\mathrm{T}} dt \tag{13.17}$$

where the integral above is over the propagation interval Δt, and $\mathbf{Q}$ is the process noise covariance that appears in the differential equation describing the covariance matrix evolution in time.

In many applications of the Kalman filter in GNSS/INS integration, (13.17) is simply but adequately approximated as $\mathbf{Q}_d = \mathbf{Q}\Delta t$. Finally, the elements of the $\mathbf{Q}$ matrix are often viewed as tuning parameters in designing the Kalman filter, discussed further when a specific filter design is reviewed.

As illustrated in the Measurement Update block in Figure 13.6, the filter measurement model includes an $\mathbf{H}$ matrix and an $\mathbf{R}$ matrix. The $\mathbf{H}$ matrix, already introduced in the discussion of PDOP (see Section 11.2.1), represents the linearized relationship between the GNSS measurements and the state vector. As such, it includes the unit LOS vector between the estimated host and satellite positions and the sensitivities to any other modeled states associated with GNSS. As a minimum, those states include the time and frequency errors associated with the GNSS receiver oscillator. For the GNSS PR measurement, the unit LOS vector represents the observability of position error, and the oscillator's time offset from GNSS system time is directly observed, as indicated here:

$$\mathbf{h}_i = \left[\mathbf{u}_i^T \ldots 1\ 0\right] \tag{13.18}$$

where $\mathbf{h}_i$ represents the ith row of $\mathbf{H}$ for the PR measurement, corresponding to the ith satellite in code track, $\mathbf{u}_i$ is the unit vector referenced above expressed in an ECEF frame, and 1 represents the clock time error observability.

Note that, for numerical stability considerations, the timing error of the GNSS clock is best represented in meters to avoid the use of the speed of light in $\mathbf{H}$, which would produce a dynamic range of eight orders of magnitude. Also note that the clock time and frequency error are assumed to be the last two states modeled by the filter in (13.18) and (13.19).

For the GNSS PRR or Doppler measurement, the unit LOS vector represents the observability of velocity error, and the oscillator's frequency offset from the frequency standard maintained by each GNSS control segment is directly observed, as in (13.19):

$$\mathbf{h}_i = \left[\mathbf{0}^T \mathbf{u}_i^T \ldots 0\ 1\right] \tag{13.19}$$

where $\mathbf{0}$ represents a three-dimensional vector of zero elements, $\mathbf{h}_i$ represents the ith row of $\mathbf{H}$ for the PRR measurement, corresponding to the ith satellite in at least frequency track, and 1 represents the frequency error observability in meters/second.

Inclusion of additional states is discussed later when a typical filter design is presented, as are the treatment options for the GNSS delta range measurement. The measurement error covariance matrix is denoted as $\mathbf{R}$ and models the measurement error. The Kalman filter assumes that the measurement error characterized by $\mathbf{R}$ is white in its update equations (i.e., $\mathbf{z}$ is completely uncorrelated from sample to sample). Such an assumption is rarely the case in dealing with GNSS: even nearly white measurement noise, which is incident upon the GNSS receiver antenna, is bandlimited in the front end and then further correlated by the finite bandwidth of the tracking loop. Nonetheless, there is generally a component of the assumed measurement error variance assigned to each measurement which is a function of

the estimated C/N_0, which is assumed to be white. Given a measurement update rate of once per second, this implies tracking bandwidths which are a fraction of 1 Hz, generally a good assumption for carrier tracking (producing the PRR measurements), but often violated for code tracking (producing the PR measurements). Other components which contribute to the measurement error variance often dominate the noise component and include uncompensated ionospheric and tropospheric delay, residual satellite position and timing error, and multipath effects. These error components tend to be near bias-like, with time constants that persist over multiple updates of the filter, and so do not behave like white noise. These error components therefore deserve special attention and will be addressed when specific Kalman filter designs are reviewed.

The measurement vector **z** that appears in the measurement update step represents a linearization of the measurements available from the receiver. In the case of the PR measurement, a prior estimate of the range to each satellite is required for the linearization:

$$z_i = PR_i^m - R_i \tag{13.20}$$

where PR_i^m denotes the measured PR, and R_i denotes the estimated range that is required for linearization.

Such a range estimate is typically determined from the inertial position (preferably corrected by the prior history of GNSS measurements) and the satellite ephemeris. Generally speaking, position errors approaching a kilometer are acceptable for linearization, which should be satisfied except for very long periods of inertial growth prior to GNSS acquisition or very long outages following initial convergence. Note that it is not necessary to subtract the current estimate of the timing error in the GNSS oscillator in (13.20), as it enters the equation linearly.

The measurement vector **z** corresponding to the PRR measurement is more nearly linear:

$$z_i = PRR_i^m - \mathbf{u}_i^{\mathrm{T}}\left(\mathbf{v} - \mathbf{v}_{si}\right) \tag{13.21}$$

where **v** is the current best estimate of velocity and $\mathbf{v}_{si}$ is the satellite's velocity for the PRR of interest.

Given the preferred ECEF frame mechanization of the GNSSI navigator, **v** and $\mathbf{v}_{si}$ are best represented in an ECEF, as the unit LOS vector must be. Note that the velocity error enters (13.21) linearly. Nonlinear effects in forming the measurement **z** arise from the product of position error (as it impacts the LOS vector calculation) and velocity error and are generally only significant for situations of very large error levels for each (i.e., nearly a km and several meters per second, respectively). The PRR measurement is available from the receiver when in either frequency or phase track, so is more robust than the more accurate delta range measurement, discussed later.

Before leaving this Kalman filter review, the general issue of filter robustness needs addressing: as indicated in the Measurement update block of Figure 13.6, each vector of measurements processed by the Kalman filter is first adjusted by subtracting the filter's expected value for this vector ($\mathbf{H}_k\hat{\mathbf{x}}_k^-$). This difference is commonly referred to as the *measurement residual* or *innovation sequence*, and its

relative magnitude can be used to make judgements about the robustness of the filter's measurement model. Many filter designs will compare the magnitude of the residual with its expected variance (given by $\mathbf{H}^T\mathbf{P}\mathbf{H} + \mathbf{R}$) [11]. Should the residual be excessively large with respect to its expected variance, one or more of the current measurements can be rejected, or at least deweighted since it is highly unlikely that it the residual can be induced by the measurement error (as modeled by the filter). While such a test is well motivated, its blind application can lead to divergence of the filter. In general, a residual that is excessive in a statistical sense can imply that either the measurement or the propagated solution is inconsistent with the model assumed by the filter: in the case of an erroneous propagated solution, rejection of GNSS measurements is removing the only means the filter has to correct the (excessive) error in its propagated solution. There is no uniformly applicable approach to solving the problem of distinguishing whether the measurement or the propagated solution is in error: however, an understanding of the possible failure mechanism signatures can lead to an effective strategy. In the case of GNSS, an erroneous measurement could be caused by a failure to declare loss of track in an environment of low signal-to-noise ratio or an excessive multipath condition. Such error conditions should not persist indefinitely. Errors associated with the propagated solution could correspond to excessive error conditions associated with the IMU (e.g., an excessive scale factor or bias) or the GNSS receiver oscillator (e.g., a micro-jump in the reference frequency). Such error conditions produce errors in the propagated solution that will persist until the GNSS measurements remove them and are not associated with a specific satellite. Thus, a strategy is suggested for discriminating between "failures" associated with the propagated solution and the GNSS measurements: successive, multiple excessive residual magnitudes across many satellites would be indicative of IMU or clock errors, while single, momentary rejections isolated to fewer satellites should lead to measurement edits. Certainly, significant tuning of the filter thresholds in these situations is required for best performance. In the situations where the propagated solution is suspect, a reset of position and velocity to a GNSS-only solution is often the best strategy. Simultaneously with this reset, the error estimates associated with the IMU errors are sometimes reset to zero, and/or their error variances boosted to uncalibrated levels to force the filter to recompute them as dictated by the future GNSS measurement history.

13.2.3.2 Hierarchy of Kalman Filter Designs

Given the preceding general discussion of GNSS/INS filter design issues, a specific design can now be reviewed in greater detail: prior to this, the general hierarchy of designs is reviewed in Table 13.1. Certainly, this cannot include all filter designs, but represents those which have been used in different applications of GPS/INS integration over the last several decades. GNSS/INS designs are similar, but to process multiple constellations may include additional states, for example, for GNSS system time differences (see Section 11.2.5). It should be noted at the outset that the classes of applications identified in the table are qualitative (e.g., there are not precise boundaries between low to moderate dynamics and high dynamics). That transition is determined as much by the type of maneuvers as their magnitude (i.e., a relatively slower maneuver may enable estimation of scale factors and misalignments associated with the instruments). Similarly, the GPS Anti-Jam (AJ) Application is largely

Table 13.1 Hierarchy of GPS/INS Kalman Filter Designs

State Size Ranges	*Description*	*Application*
17	Minimal GPS, IMU bias calibration	Minimal GPS outages, Low to moderate dynamics
23 to 32	Improved INS calibration, IMU scale factors, IMU misalignments, Gyro G-sensitivities	High dynamics
36 to 44	Improved GPS clock model, aiding error estimation	GPS low signal application, high dynamics

a function of the relative criticality, which receiver aiding plays in the AJ solution. For example, designs that incorporate a control reception pattern antenna (CRPA) (see Section 13.2.7) to steer electronic nulls in the direction of jamming sources may come close to satisfying an AJ requirement without significant attention being given to receiver aiding (see Section 13.2.8).

The 17-state filter is perhaps most often used for GPS/INS integrations. It provides a level of IMU calibration, as gyro and accelerometer biases are included as filter states. Only two GPS states are part of the design, corresponding to the time and frequency error associated with the GPS receiver oscillator. Because of its fairly wide use, it is selected as an example for a more detailed description in the next sections. When higher dynamics of the platform are expected, additional states can be added to represent the IMU scale factor errors (a total of six states, bringing the total to 23), as well as the installation (orthogonal) misalignments, adding another six states, bringing the total to 29. Further, for applications using nonoptical (i.e., ring laser or fiber optic) gyro technology, three additional states can be added to represent the gyro g-sensitivities, bringing the total to 32.

For high AJ applications where the advantages associated with receiver aiding are prominent, additional states can offer advantages; included among these are a clock frequency rate and three frequency g-sensitivities, bringing the total to 36. Finally, for best aiding performance, the latency error associated with the INS aiding can be estimated, as well as any error associated with the lever arm between the IMU and phase center of the GPS antenna assembly, bringing the total to forty. Additional states can be added beyond these; for example, in supersonic and hypersonic applications, the error coefficients associated with products of specific force components can be significant (three additional states). For GPS integrations where only a single frequency is available and differential correction data is not available, bias states associated with residual ionospheric delay can be introduced (the number of additional states correspond to the number of parallel tracking channels in the receiver, but could be up to 12).

13.2.4 GNSSI Integration Methods

One approach to GNSS/INS integration, referred to as *loosely coupled*, is generally used when raw measurement data is not available from the GNSS receiver. An approach which makes use of the raw measurement data has been the focus of the treatment to this point and is preferred to the loosely coupled approach. When loosely coupled, the GNSS receiver will output its position and velocity solution.

Such a solution could be obtained using a Kalman filter within the GNSS receiver (which could also make use of inertial information), as indicated in Figure 13.7, or preferably through the use of least squares (LS) or weighted least squares (WLS) for ease of external integration with the INS. If the receiver outputs to the external Kalman integration filter have already been filtered by its internal Kalman filter, time correlations are introduced into the position and velocity solutions that are passed out, which, for best operation, should be modeled by the external Kalman filter. These correlations are removed if an LS or WLS solution is output; however, the components of position and velocity error are nonetheless correlated by the satellite geometry. More rigorous treatments of these correlation issues are beyond the scope of this introductory discussion and are dealt with elsewhere [12].

Returning to Figure 13.7, a loosely coupled architecture is illustrated that includes an IMU, a navigation processor that contains an external Kalman filter, and a strapdown navigation algorithm. The navigation processor, as shown in Figure 13.7, accepts the GNSS position and velocity from the GNSS receiver, and $\Delta\theta$ and Δv from the inertial unit.

Although used in many early applications, the potential for the two separate Kalman filters requires special attention: as a minimum, the second filter must be retuned to the presence of correlated errors in its measurements. Mission scenarios for this configuration must be thoroughly simulated to ensure the best performance of the retuned filter.

Today, most GNSSI systems are *tightly integrated*, as shown in Figure 13.8. This configuration is also referred to as *tightly coupled*. In tightly integrated systems, the Kalman filter in the GNSS receiver is eliminated or bypassed and PR and PRR or delta range data from the GNSS channel processor is sent directly to the navigation processor. In this configuration, unmodeled errors resulting from the GNSS receiver's Kalman filter are eliminated and the system designer is allowed to set gains as a function of the GNSS error characteristics.

In the tight integration of a GNSSI system, as in most inertial systems, the Kalman filter estimates the error in the strapdown navigator and uses the estimated error state vector **x** to correct the output of the navigation equations, as shown

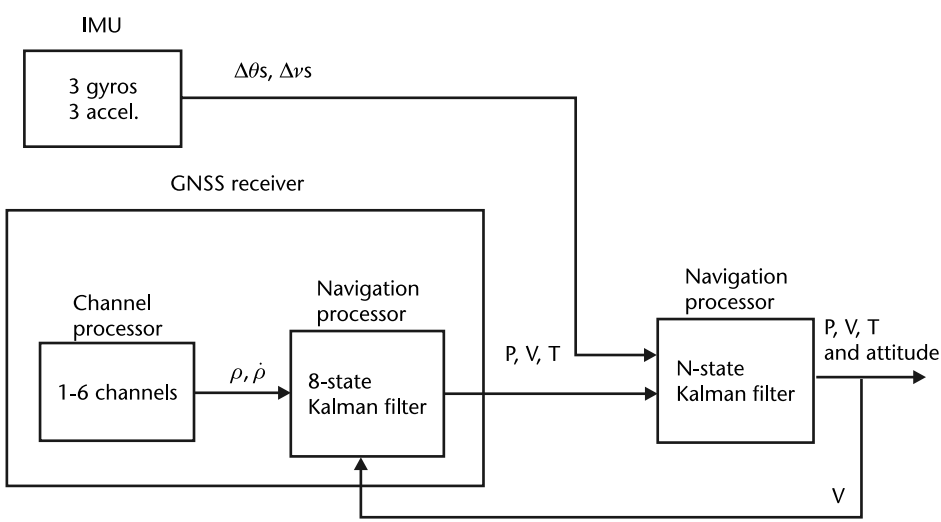


Figure 13.7 Loosely coupled GNSSI system.

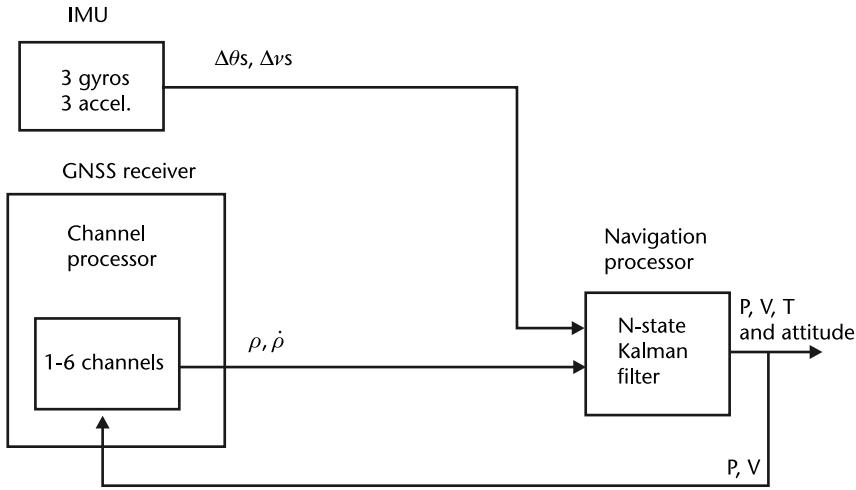


Figure 13.8 Tightly coupled GNSSI system.

in Figure 13.8. Also typical in the tight integration of a GNSSI, especially in applications where AJ enhancement of the GNSS receiver is needed, is some form of tracking loop aiding. This subject deserves special attention, and is addressed in Section 13.2.8, after the performance potential associated with adaptive arrays is discussed.

13.2.5 Typical GPS/INS Kalman Filter Design

Given the preceding, more general discussions of Kalman filter design for GPS/INS integration as a background, the specifics of a particular design can now be reviewed in detail. The 17-state option identified in Table 13.1 is selected for the review.

13.2.5.1 Filter Dynamic Model

The filter's dynamic model begins with the (9×9) **F** matrix specified in (13.8) through (13.11). The next six states correspond to the accelerometer and gyro bias states, respectively, bringing the total to 15, and the last two states represent the receiver oscillator's time and frequency error. Three additional nonzero partitions need to be added to the new 17×17 **F** matrix to represent the dynamics of the added states. The following partitions are needed: the velocity errors are driven by the accelerometer biases, producing the $\mathbf{F}_{va}$ partition:

$$\mathbf{F}_{va} = {}_{E}\mathbf{C}_{S} \tag{13.22}$$

where ${}_{E}\mathbf{C}_{S}$ represents the direction cosine matrix between ECEF and Sensor frames.

The transformation into the sensor frame is necessary since the accelerometer biases are defined in the Sensor frame. The attitude errors are driven by the gyro biases as described by the $\mathbf{F}_{\varphi\omega}$ partition:

$$\mathbf{F}_{\varphi\omega} = {}_{E}\mathbf{C}_{S} \tag{13.23}$$

Finally, the last new partition represents the oscillator's frequency error driving its time error, producing a unity value for that element in **F** (i.e., the seventeenth column in row 16).

Calculation of the state transition matrix from the **F** matrix is straightforward, with the only issues being the rate at which the propagation needs to occur, and the order of the expansion as in (13.16). Note that these two issues are not independent, as pointed out previously; use of smaller value for the propagation step improves the accuracy associated with a first-order expansion of the state transition matrix. Since a relatively low dynamic application is targeted by this filter design, a first-order expansion using a 1-second propagation interval should be adequate. The only potential issues that should be confirmed via simulation relate to how quickly some of the elements of **F** could change over 1 second, notably the specific force vector elements and the direction cosine matrix between ECEF and Sensor frames. The rate of change of the specific force is limited by the maximum jerk which the host can achieve, which are expected to be low for the applications best suited by this filter design, as should the maximum attitude rates that determine the validity of the piece-wise constant ${}_{E}\mathbf{C}_{S}$. Should either of these implicit assumptions appear questionable, a higher rate for the propagation interval may be appropriate. Alternatively, the filter's process noise covariance matrix could include increments which adjust for approximations in the state transition matrix. Such increments are discussed in the next section.

Process Noise Covariance Matrix Selection

As discussed previously, when the filter's covariance propagation was reviewed, proper selection of the process noise variances can be critical to achieving best performance of the filter. As will be shown in this section, selection of proper levels for these variances to represent unmodeled sources of error is not straightforward, so often becomes a tuning parameter for achieving acceptable performance. Use of levels that are too large generally produces measurement overweighting, which reduces the ability of the filter to reduce measurement errors. Selection of levels that are too low (i.e., which underestimate the effects of unmodeled effects) can lead to measurement underweighting, possible erroneously induced measurement editing, and even filter divergence. For the selected 17-state filter design case, certain elements of the process noise covariance matrix can be set directly; process noise variance levels corresponding to angle random walks (added to the propagated covariance matrix for the attitude error states) and process noise levels for velocity random walks (added to the propagated covariance matrix elements for the velocity error states) are generally set from corresponding vendor specification levels.

The unmodeled sources of error of significance for our design include:

- Gyro and accelerometer scale factor and misalignment errors, including asymmetry;
- Random error components of gyro and accelerometer bias;
- Gyro and accelerometer temperature sensitive drifts;
- Vibration-induced gyro and accelerometer bias components;
- Gravity modeling error.

Each of these sources of error can contribute to the process noise covariance. For the scale factors and misalignments, the variance associated with a velocity or attitude increment can be computed as:

$$\sigma^2_{\Delta \mathrm{vi}} = \sigma^2_{\mathrm{aSFi}} \Delta \mathrm{v}^2_{\mathrm{i}} + \sigma^2_{\mathrm{amisij}} \Delta \mathrm{v}^2_{\mathrm{j}} + \sigma^2_{\mathrm{amisij}} \Delta \mathrm{v}^2_{\mathrm{k}} \tag{13.24}$$

$$\sigma^2_{\Delta \theta \mathrm{i}} = \sigma^2_{\mathrm{gSFi}} \Delta \theta^2_{\mathrm{i}} + \sigma^2_{\mathrm{gmisij}} \Delta \theta^2_{\mathrm{j}} + \sigma^2_{\mathrm{gmisij}} \Delta \theta^2_{\mathrm{k}} \tag{13.25}$$

where i, j, k denote the sensor axes, Δv and $\Delta \theta$ represent the velocity and angular increments, respectively, and σ_{aSF} and σ_{amis} and σ_{gSF} and σ_{gmis} represent the one-sigma levels for the scale factor and misalignment for the accelerometers and gyros, respectively.

While (13.24) and (13.25) are correct representations of the variance change, they remain so only as long as the captured Δv and $\Delta \theta$ changes represent the complete host maneuver. This matching cannot be guaranteed, as the Kalman filter propagation step size will not match the maneuver duration, which cannot be predicted by the filter. Stated in strictly mathematical terms, the sum of the propagation step size variance increments given by (13.24) and (13.25) will not equal the variance of the sum of the delta-velocity and delta-theta increments over the maneuver. To illustrate this point, assume that the host is changing heading at a rate of 10°/sec and completes a 90° course change in 9 seconds. Also assume a 2,000-PPM gyro scale factor error about the vertical axis. Application of (13.25) produces the following estimate of the variance increase of heading error at the end of the maneuver:

$$\sigma^2_{\mathrm{A}\theta\mathrm{est}} = 9\left(2000\text{e-}6\right)^2 \left(10\right)^2 = 1.8\,\mathrm{deg}^2 \tag{13.26}$$

The actual variance increase is given by:

$$\sigma^2_{\Delta\theta} = \left(2000\text{e-}6\right)^2 \left(90\right)^2 = 16.2\,\mathrm{deg}^2 \tag{13.27}$$

Thus, blind application of the variance matching over the propagation step does not give the correct estimate of the variance increase; in fact, the prediction is optimistic, and the optimism will worsen the longer the maneuver is present. Adjustments to the filter are required to mitigate this condition, and several approaches can be taken. The simplest is to scale up the variance increase by a fudge factor or tuning parameter such that the variance change for the expected worse-case maneuver is matched. In the example above, if a 90° heading change is indeed the worst-case, a scaling of nine would produce the desired agreement. This would make the filter conservative for other maneuvers, a welcome change from the disparity produced by the blind application of (13.24) and (13.25) in each filter time step. Other approaches to improving the approximation are based upon running sums of the delta-v and delta-theta increments which better match the variance change, that is,

$$\left(\Delta\theta_{\mathrm{k}} + \Delta\theta_{\mathrm{sum}}\right)^2 = \Delta\theta^2_{\mathrm{k}} + 2\Delta\theta_{\mathrm{k}}\Delta\theta_{\mathrm{sum}} + \Delta\theta^2_{\mathrm{sum}} \tag{13.28}$$

where $\Delta\theta_k$ is the current angle increment and $\Delta\theta_{sum}$ is the accumulated sum of the angle increments during the current maneuver.

Both random and temperature sensitive error components of gyro and accelerometer bias are frequently represented by an additive process noise component, with its variance increase (linear and proportional to corresponding continuous process noise element) chosen to represent the random drift of the bias. The gyro and accelerometer temperature sensitivities are compensated internal to the IMU, so it is only the residual effect which is modeled by the process noise increment.

Vibration of the host platform induces shifts in the effective gyro and accelerometer biases: such shift levels, if significant, are characterized by vendors over worst-case environments. To first order, the worst-case level can be considered a three-sigma condition, and a one-sigma level can be root-sum-squared with the nominal one-sigma bias level to produce a level of conservatism in the design. The solution for the gyro bias will then correspond to the sum of the bias levels for constant vibration levels. Should the bias levels change dramatically, more sophisticated approaches that explicitly model the sensitivity over a range of measured vibration magnitudes and frequencies can be considered. Such approaches are considered in [13].

Finally, gravity model errors, although expected to be small (on the order of tens of micro-gs in magnitude), can become noticeable if they persist long enough as effective biases: this is largely a function of the host speed over the surface of the earth, since the gravity modeling error is spatially correlated. For a correlation distance of 10 nm, for example, and a vehicle speed of 100 m/s, this bias level persists for roughly three minutes. A 20 micro-g gravity model error integrates to a velocity error of roughly 0.04 m/s, which is roughly at the noise level of the carrier measurements. However, slower speeds could produce larger error growth and so their effects should be included as process noise (with appropriate scaling to produce sufficient accumulated error over the filter's propagation step). Before moving on to the measurement model used in the filter case study, it is important to note that, although the guidelines given in this subsection for process noise computation are well motivated and technically quite sound, it is nonetheless required, in general, to run any candidate reduced state filter design against a higher order truth model in a realistic simulation that can at least envelope the expected range of host dynamic and GNSS coverage conditions to ensure filter robustness.

13.2.5.2 Filter Measurement Model

As referenced earlier, there are two general sets of measurement data generated by the receiver in conventional (i.e., excluding integrated tracking and navigation architectures considered in Section 13.2.7.3) receiver architectures. The PR measurement, introduced earlier, is an output of code tracking, and is available for processing by the Kalman filter even after carrier tracking has been lost, since the code loop is often aided using corrected inertial information (discussed in Section 13.2.7) in expected weak signal tracking environments. The second set of measurements is derived from tracking the carrier and includes the PRR or Doppler measurement and/or the delta range or integrated Doppler measurement. Both sets require at least frequency track, but are more accurate when phase locked on the carrier.

Code-Based Measurements

The treatment of the PR measurement provided in Section 13.2.3.1 does not need to be repeated here, the expressions for the residual formation (13.21) and measurement gradient vector **h** (13.18) can be applied directly in this design. In the case of the measurement gradient vector, the unity element corresponds to the sixteenth element of the state vector. The emphasis here will therefore be on the proper calculation of a measurement noise variance to assign to each measurement. As referenced earlier, this is not straightforward, since the measurement error can be dominated by sources of error that are time correlated and so violate a fundamental assumption of the filter. In particular, for commercial designs which do not support track on a second (e.g., the L2 frequency for GPS), and which cannot receive differential corrections, residual ionospheric delay will dominate the PR error budget, and have a correlation time sufficiently long to appear bias-like in the filter's 1-second updates. Several approaches exist to mitigate this problem, including augmenting the state vector in an attempt to estimate the residual delay or introduction of the biases as consider states [14], which are beyond the current focus of the seventeen state filter design: hence we will assume track on the second frequency and/or receipt of differential correction data in the design.

Even with the assumption of compensation for ionospheric delay, there are nonetheless multiple sources of error that can comprise the PR measurement error that do not adhere to the filter's implicit assumption of white (uncorrelated) noise, including:

- Residual tropospheric delay (if not compensated by differential);
- Residual satellite position and timing error (if not compensated by differential);
- Tracking loop error, including the effects of multipath.

Each of the sources of error above are correlated over many filter (1-Hz) updates, ranging from tens of seconds for tracking loop error, to several minutes for residual tropospheric delay, to periods approaching an hour for residual satellite position error. The nature and range of magnitudes of each of these error sources is discussed in Chapter 10 and so need not be repeated here. While it seems intuitive to simply assign an error variance to each of the above error sources, and root-sum-square each magnitude with an assigned noise variance which is a function of the receiver's estimated signal-to-noise ratio, as in (13.29), such an approach can lead to optimism in the filter's estimates.

$$R_{PR} = \sigma_n^2 \left(C/N_0\right) + \sigma_{\text{tropo}}^2 + \sigma_{\text{pos}}^2 + \sigma_{\text{sat}}^2 \tag{13.29}$$

where C/N_0 is the estimated signal to noise ratio from the receiver.

The cause of the potential optimism is quite simply the fact that the filter assumes that the lump sum assigned error variance is uncorrelated noise of that magnitude, and so, as a function of how well it has estimated the INS error in the propagated solution, it will believe that it can average away the measurement error from second to second. It cannot due to the unmodeled nature of the error. This problem can be solved by the addition of consider states to represent the correlated

error as in [14] or through the addition of scale factors on each of the error source magnitudes. These scale factors boost up the assumed error variances to avoid the potential for optimism. As a worst-case guideline, an error source that is correlated over 100 filter steps should be scaled by a factor of 10 (at its sigma level) to avoid optimism, since the filter can, at best, average as the square root of the number of samples. However, as was the case with the selection of process noise variance for the filter, measurement noise variance parameters should be tuned for best performance using a higher-fidelity simulation exercised over an envelope of dynamic and GNSS coverage scenarios that includes representative models for each of the sources of error.

Carrier-Based Measurements

The prior discussion of carrier based measurements in Section 13.2.3.1 focused on the more straightforward PRR (or Doppler) measurement, as it more readily fit into the Kalman filter framework. The discussion here will focus on the more accurate delta range measurement, which requires the receiver to be in phase lock for full accuracy. When phase lock is lost, the less accurate Doppler measurement is processed as long as frequency lock is maintained. The delta range is derived by the receiver by differencing its estimate of carrier phase (the outputs of each loop's numerically controlled oscillator) over a predetermined interval, nominally 1 second. The measurement residual is found by subtracting the change in the estimated pseudorange, as in:

$$DR_{res} = DR^{m} - \left(R_{est} + \delta\varphi_{est}\right)_{k} - \left(R_{est} + \delta\varphi_{est}\right)_{k-1} = \mathbf{h}_{k}^{T}\mathbf{x}_{k} - \mathbf{h}_{k-1}^{T}\mathbf{x}_{k-1} \tag{13.30}$$

where R_{est} denotes the best estimate of range, and $\delta\varphi_{est}$ denotes the best estimate of clock phase error at the indicated times.

Most designs use a computed state transition matrix over the delta range integration interval (t_{k-1}, t_k) to collapse the residual to an estimate of the state error at a single point in time:

$$\mathbf{h}_{DR} = \mathbf{h}_{k}^{T} - \mathbf{h}_{k-1}^{T}\Phi\left(t_{k-1} - t_{k}\right) \tag{13.31}$$

Note that the state transition matrix indicated in (13.31) corresponds to the inverse of the transition matrix over the delta range integration interval (i.e., it propagates the state from the end of the interval to the start). While the modeling approach embodied in (13.31) seems reasonable and is often used, the propagation step adds the uncertainty associated with the state dynamics to the delta range measurement uncertainty. Stated another way, this formulation correlates process and measurement noise, which violates a fundamental assumption of the Kalman filter. Nonetheless, many designs proceed with this formulation by root-sum-squaring the clock phase and frequency noise contributions to the assumed delta range measurement noise variance. For some designs, since the delta range measurement can be accurate to a very small fraction of a carrier wavelength, this has the potential to degrade the achievable accuracy. For those situations, consideration of alternate formulations, including the delayed state model [15], could be considered. The delayed state formulation essentially augments the state vector to include estimates

of the state at each end of the delta range interval, doubling the state size in the worst case. The increased state size collapses to the original state vector as the last available delta range measurement is processed. Further consideration of this formulation is beyond the scope of discussion for the selected 17-state option. Finally, in contrast to the PR measurement, the residual correlated errors in the delta range measurement, including multipath on the carrier, satellite frequency, and velocity error, and variations in the atmospheric delay over the delta range integration interval, can be expected to be much smaller than the delta range measurement noise and so can be neglected.

Measurement Residual Editing

As referenced earlier in Section 13.2.3.1, measurement residuals can be compared with the predicted error variance ($\mathbf{h}^T\mathbf{P}\mathbf{h} + \mathbf{R}$) as a means of predicting and potentially screening against excessive measurements errors, as could occur near thresholds of signal track, or in situations where the signal has taken a reflected path to the GNSS receiver. Also as discussed previously, such measurement editing can lead to divergence of the filter if the excessive residual is induced by error in the propagated inertial solution, rather than measurement error. Stated more simplistically, the propagated inertial is lying to the filter, and GNSS is telling the truth and so should not be ignored. For the 17-state filter design, a means for distinguishing between these two error conditions is highly desirable. The selected approach is based upon two separate but related observations: if the INS position or velocity error is excessive, it should produce significant bias error in the residual in the short term and so lead to multiple, successive measurement rejections from different satellites. Although similar behavior could be observed if several tracking channels were near threshold conditions, that should only occur at a relatively low estimated signal-to-noise ratio. The second tool at our disposal is the ability to do a consistency test on the GNSS measurements themselves when the GNSS measurement set is overdetermined. The magnitude of the residual of the overdetermined LS or, preferably, WLS solution (the very same statistic used in RAIM; see Section 11.4.3.1) can be compared to a threshold to determine the consistency of the overdetermined set. Should the GNSS measurement set prove to be consistent, the GNSS measurement data can be trusted, and successive rejections across multiple satellites, in combination with large residual bias levels, indicates that the GNSS measurements should not be edited. In fact, depending upon the relative magnitudes of the confidence statistic and the residual bias relative to the established thresholds, the smartest strategy may be to reset the current GNSS/INS solution to the overdetermined GNSS solution leading to a reset of the position and velocity to the GNSS solutions. A decision then needs to be made what should be done with the estimates of INS attitude error and gyro and accelerometer bias estimates, in addition to the clock error state estimates. As a minimum, their covariance levels should be increased to permit the Kalman filter to develop new estimates; at question is whether or not the former estimates should be carried forward or similarly reset to zero. The foregoing discussion can lead to an effective strategy for use of the residual test: however, such tests must be evaluated using a high-fidelity simulation with failures injected to assist in optimizing the performance of the editor.

13.2.6 Kalman Filter Implementation Considerations

Two issues worthy of addressing at this level are data synchronization and the numerical stability of the Kalman filter. In general, the outputs of the IMU and the GNSS measurements will not be synchronously generated. The Kalman filter design which was reviewed assumed the time of the GNSS measurement and that of the inertial system propagation were identical. Imperfect synchronization of the measurement data produces an unmodeled error effect when the measurement residual is formed. Should the expected synchronization error be random from sample to samples and small relative to the assigned measurement noise, it can generally be neglected or modeled as an increase in the noise variance assigned to each measurement. However, even a relatively small biased synchronization error may require special attention. In the worst case, depending upon its magnitude, it can be modeled as an error state and estimated or introduced as a consider state in the filter to force a level of conservatism. Independent of the approach which is taken by the filter, the design should do everything possible to ensure synchronous measurement data. The key elements of an accurate synchronization are discussed next.

The two key elements required for an accurate synchronization are the timing associated with the inertial data, and the buffering of the inertial data to permit interpolation and/or extrapolation. Timing of the inertial data is accomplished by having the GNSS receiver transmit a 1-PPS signal to the navigation processor. This signal, tied to a high-level interrupt, forces the inertial clock to the next second. The inertial clock is a software clock that is incremented by each inertial measurement received by the navigation processor (typically at a rate of 100 Hz to 800 Hz). The inertial clock is thus resynchronized to GNSS receiver clock time once per second. To initialize the inertial clock, the GNSS receiver must implement a specific message that will inform the navigation processor of the GNSS receiver time at the next interrupt. This must be accomplished well before the receipt of the interrupt to give the navigation processor time to respond to the interrupt and the message and prepare to set the inertial clock before the next interrupt is received. Since the GNSS receiver and the inertial are asynchronous, a circular queue, called a history queue, contains 1 or 2 seconds of inertial position data. By examining the time of the GNSS measurement, the latest inertial position whose time tag is less than that of the GNSS measurement can be extracted from the queue. Using the next queue entry, the data is then interpolated to the time of the GNSS measurement.

Finally, as is fairly well known in Kalman filter circles, the filter is susceptible to numerical stability issues, which, in the worst case, can lead to loss of positive definiteness of its covariance matrix [16]. Experience with Kalman filtering has led to the conclusion that a wide range of magnitudes associated with measurement processing, coupled with the conventional covariance update using the matrix $[\mathbf{I} - \mathbf{k}^T\mathbf{h}]$ are the likely sources of the instability. Two relatively simple steps can be taken: as referenced above, making sure that time and frequency are represented using the same set of units as position and velocity, avoids use of the speed of light in the filter $\mathbf{h}$ matrix, reducing the potential dynamic range required. Second, use of the Joseph formulation (see the covariance update equations in Figure 13.6) replaces the simple matrix subtraction with a matrix equation that closely approximates a matrix root-sum-square. Given these two adjustments, simulations should be performed which exercise the covariance update over the maximum number of

iterations expected, with tests for positive definiteness performed. Should problems be observed, additional measures should be taken. The first is resymmetrizing the covariance after each update by simply averaging the corresponding off-diagonal elements [i.e., P(k, j) and P(j, k)]. The second is giving consideration to increasing the precision of the calculations (i.e., using double precision in lieu of single precision). Finally, the use of covariance factorization [16] can be used and effectively maintains the covariance as its square root, further increasing the effective precision associated with the calculations.

13.2.7 Integration with Controlled Reception Pattern Antenna

This section discusses the integration of a controlled reception pattern antenna (CRPA), originally discussed in Section 9.2.3.2, with a GNSS/inertial system. The gain pattern of the CRPA antenna as compared to a standard fixed reception pattern antenna (FRPA) when a source of interference is present is illustrated in Figure 13.9.

The CRPA minimizes gain towards the interference source adaptively by utilizing an array of N antenna elements, as shown in Figure 13.10, with $N = 7$. The signal from each element is weighted and combined to minimize the incoming power. Since the GNSS signals are more than 30 dB below the receiver's noise floor, any significant incident power can be attributed to jamming or unintentional interference. The Degrees of Freedom (DOF) of the CRPA is the number of antenna elements minus one. The CRPA can generate independent nulls up to its DOF, implying that nulls can be generated in the direction of that many jammers. Another aspect of the CRPA is that it allows for the gain towards the GPS satellites can be increased, maintaining the GNSS signal strength with nulling. Null steering and beam steering antennas have been successfully used to mitigate the effects of interference and multipath for GNSS applications for a number of years. Null steering antennas are currently used on a number of military platforms. Drawbacks of the use of CRPAs include high cost (relative to FRPAs), and weight/size issues. There have been a number of programs trying to address these concerns and reduce the size of the antenna from 14 inches down to 5.5 inches [17–19]. Many current

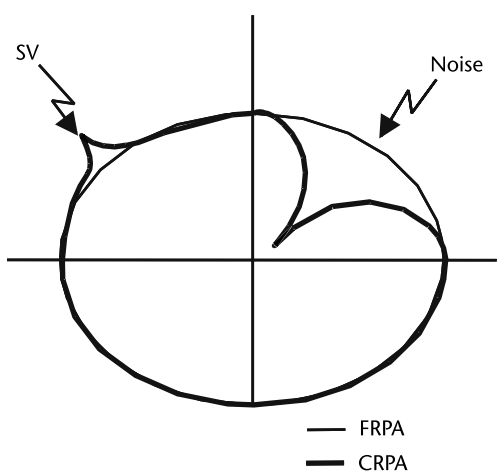


Figure 13.9 Antenna patterns of FRPA and CRPA.

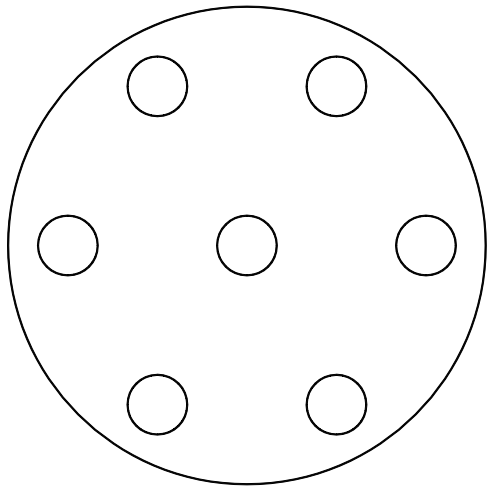


Figure 13.10 Layout of the seven-element CRPA.

CRPA applications implement null steering without beam steering. The reason is that beam steering requires knowledge of the platform attitude, which is not always available or easily accessed, and the processing burden imposed by beam steering.

A diagram of the antenna electronics (AE), which in airborne installations is usually housed within the aircraft rather than in the antenna, is shown in Figure 13.11. The electronics consists of circuitry to control the weighting of the signal from each element, a combiner used to combine the weighed signal from each antenna element, a microprocessor (occasionally referred to as an antenna controller), a combiner to reconstruct the GNSS signal, and optionally a downconverter and a power detector to measure the amount of jamming coming into the receiver if not available from the GNSS receiver. The microprocessor used within the antenna controller executes an iterative algorithm that computes the weight applied to each element that will minimizes the incoming power from the antennas. AE used to implement beam forming additionally must incorporate platform attitude and

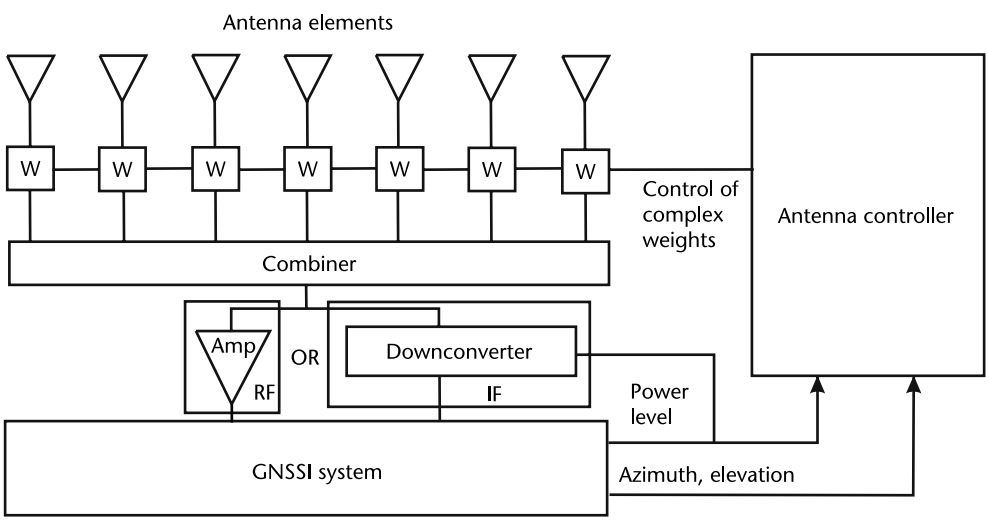


Figure 13.11 CRPA block diagram.

satellite location information into the adaptive algorithm, in combination with the measured voltages from each antenna element, to optimize the weighting applied to each element. Electronics used for currently available CRPAs also include the first downconversion and automatic gain control (AGC) electronics of the GNSS receiver (see Section 8.3). This allows the power detector of the AGC circuitry to be included in the antenna electronics.

In some implementations, a time delay line with M taps is added to each of the N antenna elements. The combiner then weights the $M \times N$ delay line taps. This technique, referred to as *space-time adaptive processing* (STAP), can improve nulling performance when operating against CW jammers by exploiting their time correlation. For an N element array with M time taps per RF channel, the CRPA DOF against CW jammers can be increased to $MN - 1$ [20].

As mentioned earlier, to implement beam steering, the AE must know the LOS direction to the satellites being utilized by the GNSS receiver. This is accomplished by means of a serial interface between the navigation processor and antennal electronics. Satellite azimuth and elevation relative to the antenna and usually the vehicle are periodically sent to the antenna electronics to use in optimizing the gain towards the satellites.

13.2.8 Inertial Aiding of the Tracking Loops

As introduced in Section 13.2.1, this aiding can occur at both carrier and code loop levels. Aiding the code loop is most commonly implemented. Aiding the phase lock loop within the receiver is much more difficult. The difficulty is obviously driven by the relatively tight requirements, from a navigation perspective, for maintaining phase lock on the carrier. Phase lock generally requires that tracking loop error is less than a fraction of the carrier cycle. For example, allowance of 90° of phase error (one-quarter cycle) translates to roughly 5 cm of navigation error. Analysis performed in Section 13.2.8.1 indicates that this translates to a very tight GNSS/INS velocity accuracy requirement. This requirement can be attained, but only with very careful estimation and control of certain IMU error sources and with IMU data extrapolation to achieve the needed update rate for phase lock. In addition, special care is necessary in the installation of the GNSS antenna on the vehicle, relative to the IMU, to avoid contamination by flexible body motion between the two. In fact, for best operation of carrier phase aiding, consideration should be given to minimizing the physical separation between the INS and the GNSS antenna in the host vehicle. Notwithstanding the difficulty of aiding phase lock within the GNSS receiver, aiding frequency lock is relatively easy to do. Further discussion of this alternative appears in Section 13.2.8.1.

Since aiding the code loop is commonly done, let us explain its nature at a conceptual level with reference to Figure 13.12. Note that the code loop nonlinearity is neglected in this simplified model (the detector is represented by a gain of unity), and the numerically controlled oscillator (NCO) within the receiver is represented as an integrator. Also note that the code loop filter is represented simply as a gain, K_c, and a continuous time model is shown. First, to explain the action of an aided code loop with reference to Figure 13.12, the range delay, ρ, minus the loop's estimate, ρ_{est}, measures the range delay tracking error $\delta\rho$, which is computed perfectly by the detector with an additive noise error, n. The loop bandwidth is proportional

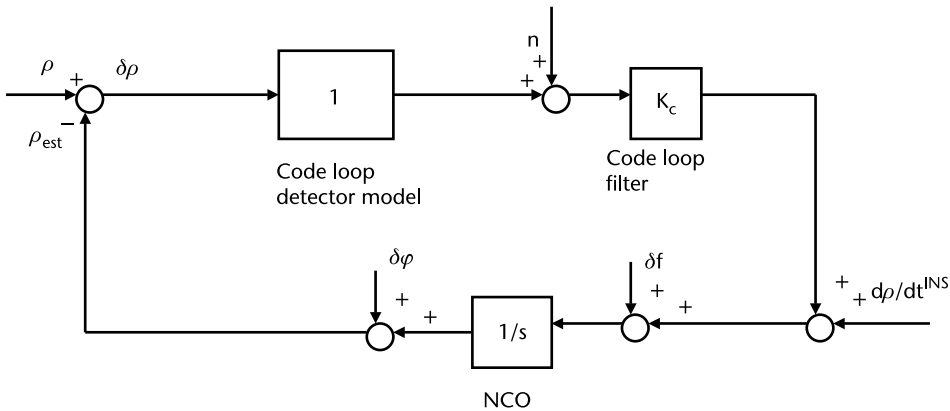


Figure 13.12 Simplified (linear) model for an aided code loop.

to the code loop gain K_c. The INS velocity is subtracted from the satellite's velocity (in a common coordinate frame) and then projected along the LOS to the satellite which is tracked by the loop to construct $d\rho/dt^{INS}$. The INS aiding signal is added to the output from the code loop filter to drive the NCO. As mentioned previously, oscillator imperfections lead to a frequency error, δf, which also drives the NCO, in addition to an additive clock phase error $\delta\varphi$. The very simple form for this model makes certain observations intuitive. Lowering the bandwidth (reducing K_c) reduces the effect of noise, n, or interference on the loop, and places more weight on the INS aiding. As a limiting case, setting K_c to zero drives the range delay estimate entirely with inertial aiding. Even in the case of perfect INS information, this is unwise, since the frequency error of the local oscillator will integrate to a range delay error, which cannot be removed by the zero-bandwidth loop. This range delay error will grow without any corrective action by the loop, and, eventually, force the loop to lose lock. Thus, the clock instability sets a floor for the aided bandwidth. One additional observation can be made using this simple model: if the INS aiding signal is expressed as the sum of the true range rate plus a range rate error induced by INS velocity error, it can be shown that the tracking loop error $\delta\rho$ is a function only of the INS errors: thus, the aiding makes the loop's performance insensitive to the actual motion (i.e., velocity and acceleration) of the host, replacing it with the dynamics of the INS errors.

13.2.8.1 Carrier Loop Aiding

As previously mentioned, aiding a phase lock loop with inertial velocity is quite difficult, due to the small GNSS wavelength (e.g., 19 cm for 1,575.42 MHz). A simplified, linear continuous time model for an aided carrier loop can be constructed in a manner very similar to that used for the aided code loop. In Figure 13.12, the range delay ρ and related quantities (i.e., ρ_{est} and $\delta\rho$) are replaced by their counterparts θ, θ_{est}, and $\delta\theta$, respectively. The code loop filter K_c is replaced by the carrier loop filter (also a gain K_θ in this simple model), and the rate of change of the range delay $d\rho/dt^{INS}$ is replaced by $d\theta/dt^{INS}$. The resultant model for an aided carrier loop can be used to derive (13.32), expressed in terms of Laplace (continuous time) transforms:

$$\delta\Theta(s) = \left[s/(s+K_\theta)\right]\Theta(s) - \left[s/(s+K_\theta)\right]\Theta^{INS}(s) \tag{13.32}$$

where Θ^{INS} represents a carrier phase estimate constructed from the INS velocity following initialization. Note that $\Theta^{INS}(s)$ is simply a mathematical construct introduced in the equation derivation: it is not calculated in the carrier phase aiding process. The INS constructed carrier phase estimate can be expanded as:

$$\Theta^{INS}(s) = \Theta(s) + \delta\Theta^{INS}(s) \tag{13.33}$$

Substituting into (13.32), we see the aided tracking loop error is independent of $\Theta(s)$, the actual carrier phase history, and dependent only upon the INS error (we have neglected the effects of noise and clock error in starting with (13.32) to reach a conclusion about the required INS velocity accuracy).

$$\delta\Theta(s) = -\left[s/(s+K_\theta)\right]\delta\Theta^{INS}(s) \tag{13.33}$$

However, $\delta\Theta^{INS}(s)$ can be related to the satellite LOS component of INS velocity error using:

$$\delta\Theta^{INS}(s) = \mathbf{u}^{\mathrm{T}}\delta\mathbf{v}^{\mathrm{INS}}(s)/s \tag{13.34}$$

Finally, we can express the carrier phase error of an aided loop in terms of the INS velocity error:

$$\delta\Theta(s) = -\left[1/(s+K_\theta)\right]\mathbf{u}^{\mathrm{T}}\delta\mathbf{v}^{\mathrm{INS}}(s) \tag{13.35}$$

From (13.35), the carrier phase error in steady state [determined by setting s to 0 in (13.35)] is the LOS INS velocity error component divided by K_θ. Equivalently, the aided carrier phase error is the LOS INS velocity error times the time constant of this simple loop model (the time constant is just the inverse of the gain K_θ in this first order loop model). Thus, to limit carrier phase error to 90° (assuming a time constant of 10 seconds is used), requires a LOS velocity error in steady state of no greater than 5.0 mm/s, a very tight requirement indeed. As the aided loop time constant is increased (and the corresponding loop bandwidth is reduced to further attenuate the effects of jamming), the INS velocity requirement becomes more difficult to meet. Corresponding requirements for peak transient velocity errors are less stringent: for example, a velocity error component as large as 5 cm/s, if it persists for less than 1 second, may not induce loss of track, depending upon the tracking state when the velocity transient occurred.

This very tight requirement for INS velocity error implies that certain error sources are carefully controlled, including the nonstatic component of accelerometer bias (the static component is generally cancelled by the platform misalignments generated during initial alignment), accelerometer scale factor and misalignments, and even the quantization level associated with the delta velocity derived from each accelerometer. For example, consider a residual accelerometer scale factor of 100 parts-per-million (ppm). Assume that the host vehicle is a high-performance fighter aircraft doing a highly dynamic maneuver, producing a 5g acceleration along its

lateral axis for 5 seconds. This single error source integrates to a velocity error of 2.5 cm/s, which could jeopardize carrier phase aiding with a bandwidth as narrow as that considered in our simplified analysis. Recall that it was mentioned earlier that oscillator instability also limited the potential bandwidth reduction which can be generally be achieved when receiver aiding. For the dynamic example, it is possible that the *g*-sensitivity of the local oscillator (see Section 8.9.6) will limit the utility of carrier phase aiding to as great an extent as the identified INS error sources. This point will be addressed in more detail in Section 13.2.7.8.

Common output rates of delta angle and delta velocity information from an IMU range from 10 to 100 Hz. These output rates may be unacceptable for carrier phase aiding, and can lead to large transient errors in the aiding source under worst-case dynamics. This transient error can be reduced using an extrapolation algorithm. For example, a constant-jerk model could be hypothesized for the delta velocity history, and the coefficient of the jerk term periodically determined from sets of delta-velocities output from the IMU; the model would then be used to generate modeled delta velocities to supply to the carrier loop at a higher rate. Notwithstanding these technical challenges, carrier phase aiding is possible and can be used to extend phase track in stressful signal environments [21].

Given the difficulties associated with aiding the phase lock loop, it is attractive to consider aid of the frequency tracking loop as a fallback position. Frequency track, as discussed in Section 8.6.1.3, is more tolerant of dynamic and interference induced errors than phase track. A typical error detector (see Section 8.6.1.3) used for frequency track can tolerate up to 50 Hz of frequency error. It is the use of frequency track that enables many commercial GNSS receivers to maintain track under foliage. Obviously, maintaining an INS velocity aiding error less than 10 m/s (corresponding to the 50-Hz limit at L1) is relatively easy to do and will guarantee frequency lock as long as excessive frequency error is not induced by the receiver's oscillator. Enhancements on the order of 10 dB in AJ performance are expected.

13.2.8.2 Code Loop Aiding

As mentioned in Section 13.2.1, code loop aiding is the most commonly exercised option. To gain additional insight into the operation of an aided code loop, let us return to Figure 13.12, and consider the decomposition of the aided range delay estimate, ρ_{est}, in terms of an INS component and a GNSS component:

$$P_{est}(s) = \left[K_c/(s+K_c)\right]P_{rcvr}(s) + \left[s/(s+K_c)\right]P_{INS}(s) \tag{13.36}$$

Equation (13.36) is an expression for a classic complementary filter in the frequency (i.e., Laplace) domain, in that it represents the combination of a lowpass filter operation on receiver information with a highpass filter operation on INS information. Thus, as the bandwidth of the receiver is reduced (i.e., K_c is reduced, or the loop's time constant is increased), the aided loop is constructing an estimate of the range delay based largely upon simply integrating the INS velocity from the estimated range delay when the loop was unaided. Thus, in the limit, as K_c approaches 0, the loop's estimated range delay is completely determined from the INS behavior since the onset of aiding. This observation should assist in understanding some of

the problems that are encountered when attempting to process the estimated range delay in a conventional Kalman filter design. These problems are discussed in [22].

Consider the aided code loop, including the Kalman filter operation, depicted in Figure 13.13, referred to in the discussion that follows as a partitioned design. The estimated range delay, ρ_{est}, is used to close the code loop, with its filter represented as the gain K_c as before; it is also used as a code phase measurement input to the Kalman filter. The Kalman filter generates an estimate of the INS velocity error δv_{est}, which is used to correct the INS velocity. The known satellite velocity v_s is then subtracted from the corrected INS velocity and projected along the LOS (represented by the unit vector **u**) to the satellite tracked by this loop. Based upon the complementary filter model derived for the aided code loop model, the utility of the Kalman filter correction when aided can be questioned. The aided configuration can become unstable as the bandwidth is lowered below the effective bandwidth of the Kalman filter [22, 23]: this is also driven by the fact that there are two loop filters. The first, the code loop itself, is using a very low gain (K_c) closure; the second loop closure is through the Kalman filter. The Kalman filter, expecting to receive measurements corrupted by uncorrelated measurement error, is processing measurements whose error is strongly correlated in time. This is a classical filter modeling problem and contributes to the potential for instability.

There are a number of approaches that can be used to stabilize the aided code loop [23]. Two of the more straightforward approaches include simply turning off the Kalman filter corrections to the INS while the loop is aided and reducing the effective bandwidth of the Kalman filter (i.e., reducing its gains) to be less than the lowest bandwidth that the code loop itself (determined by the lowest value used for K_c) can achieve. The referenced analysis [23], which represents the Kalman filter as a fixed gain Butterworth filter (to enable conventional stability analysis), motivates the frequency-domain interpretation in Figure 13.14. Stability problems generally arise when the Kalman filter effective bandwidth exceeds the code loop bandwidth, as illustrated in Figure 13.14.

The aided code loop depicted in Figure 13.13 is referred to as a partitioned design because the tracking loop and navigation filter are considered separate functions: the bandwidth of the tracking loop can be varied as a function of sensed

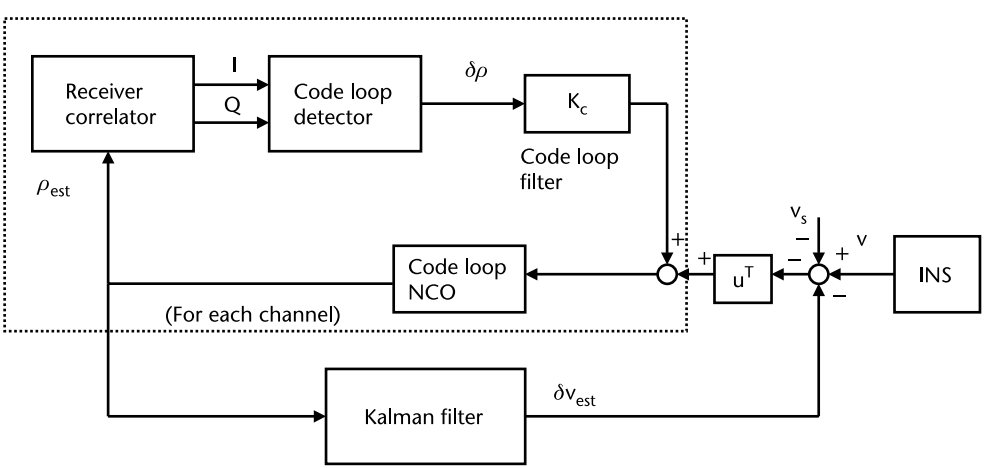


Figure 13.13 Partitioned tracker/navigator block diagram.

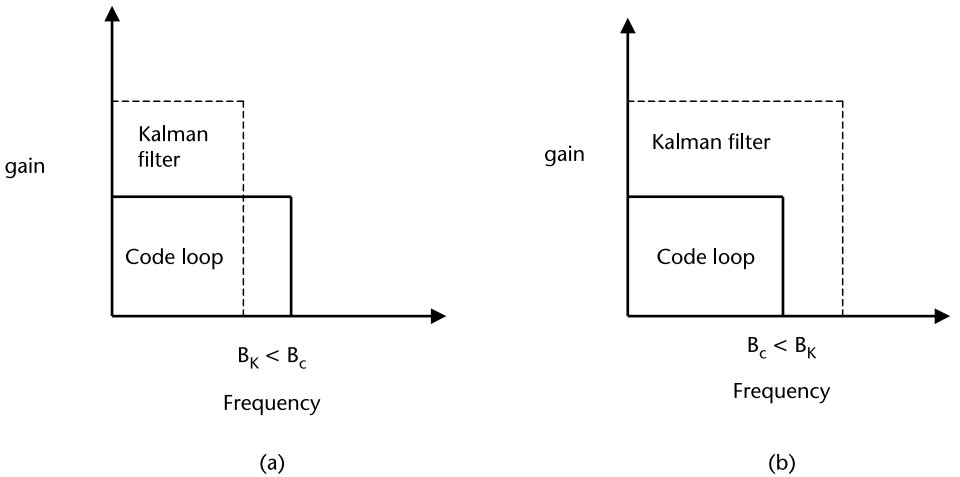


Figure 13.14 Aided code loop frequency-domain perspective.

signal-to-noise ratio, but is independent of the Kalman filter operation. In the next section, the navigation and tracking functions will be considered a single, integrated function, which will lead to a receiver aiding formulation which has been referred to as ultratight integration.

13.2.8.3 Integrated Tracking and Navigation

Figure 13.15 provides a block diagram of the integrated tracker/navigator, also referred to in the literature as *ultratight* or *deeply integrated*. The very first recognition of the benefits of this level of integration occurred in [24]. In that paper, the essential observation is made that the optimal estimators for navigation and signal tracking differ only in their coordinates (i.e., that a best estimator for position, velocity, and clock phase and frequency error should be equivalent to a best estimator for the set of satellite code phases and Dopplers). This essential observation does not depend upon the inertial augmentation of GNSS. Applications of this high-level concept have therefore arisen in commercial applications of GNSS [25, 26], where the INS is absent. These applications have extended the original concept to include direct use of I and Q signal correlations from the GNSS receiver, in place of the

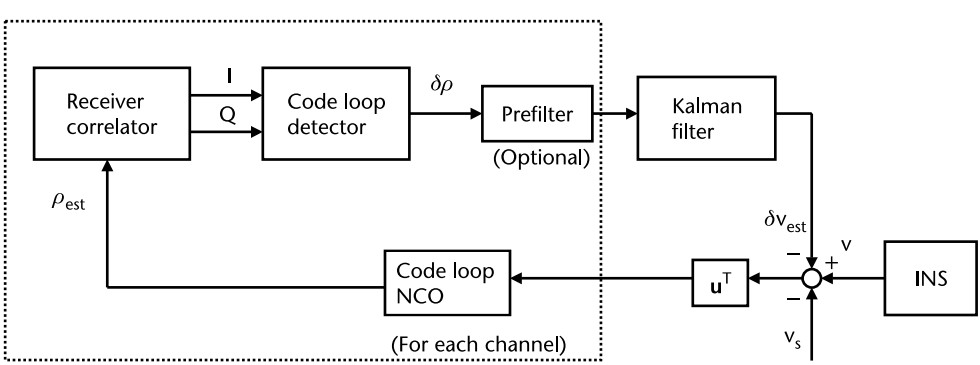


Figure 13.15 Integrated tracker/navigator block diagram.

assumed detector outputs in [24]. Such implementations are sometimes referred to as *vector tracking*, since the individual tracking loops within the GNSS receiver are no longer independent: they are coupled through their response to the position and velocity of the host vehicle and the common clock errors.

Returning to Figure 13.15, it can be seen that this architecture removes the conventional code tracking loop, replacing it by a single loop that is closed through the Kalman filter. A by-product of this new architecture is a solution for the stability problem: without the separate loop which produces an unmodeled measurement error correlation for the Kalman filter, a well-designed filter will not cause stability problems in this aided configuration. Note the optional prefilter. The receiver correlator outputs in-phase (I) and quadrature (Q) correlations at a rate ranging from a few milliseconds to up to 20 ms. This is obviously an extremely high rate for Kalman filter execution: one solution to this problem is to average the outputs of the detector up to a more typical processing rate for a Kalman filter (e.g., once per second). Recent applications of ultratight coupling have made use of reduced order Kalman prefilters to feed the Kalman tracking and navigation filter (the centralized filter) in a federated filter architecture [5, 27, 28]. Alternatively, a multirate mechanization for the Kalman filter can be used, where the state propagation and update occur at the highest rate at which the code loop detector output is generated, but gain calculation, covariance propagation and update (where the bulk of the Kalman computations occur) are performed at a more typical lower rate (e.g., 1 Hz).

Simulations are used to compare the performance of the integrated or tightly coupled and partitioned designs in [24]. Although the improvements in its response to increasing noise levels are not significant (the first simulation case considered in [24]), substantial improvements are realized when significant dynamics are combined with near-threshold noise levels (the second simulation case considered). The results are somewhat intuitive: in the first case, since both designs are able to adapt to an increasing noise level by lowering the effective aided receiver bandwidth, their performance is quite similar. In the second simulation case, it is in fact the recognition of the receiver oscillator's g-sensitivity by the integrated design that leads to the substantial performance improvement. Recall from Section 13.2.1 that tracking the dynamics of the local oscillator is a requirement that sets the floor on the aided bandwidth. As the host vehicle performs highly dynamic maneuvers near threshold tracking conditions, the ultratightly coupled design increases the aided bandwidth just enough to maintain lock. Even though the Kalman filter of the partitioned design similarly correctly models the clock g-sensitivity (its model is identical to that of the integrated design), its tracking loop does not adapt in recognition of this error source. The ultratightly coupled design thus affords another dimension of bandwidth adaptivity. More generally, the improvements of the integrated design can be understood by observing that its bandwidth adapts to everything which is modeled by the Kalman filter, including INS quality and clock dynamics. The maturity of the simulations used in [24] for the comparative evaluations was questionable, in that the receiver motion and satellite geometry were limited to a plane; however, more thorough and detailed evaluations have been reported more recently [29] and confirm the fundamental conclusions of this very early paper. Simulations with higher-quality GNSS oscillators [4, 5] indicate the higher potential associated with designs which more fully exploit direct use of I and Q information with data aiding and Kalman prefilter designs.

Given the potential performance improvements reported in [24], it is natural to ask why it has taken so long for the ultratightly coupled design to gain more acceptance. The reason for its delay in recognition as a worthy design approach may in part be cultural: not many individuals are skilled in both the art of Kalman filtering and receiver design. A more technical reason for the lack of acceptance are some of the significant modeling issues for the ultratightly coupled design, two of which can be addressed here. The first technical issue is the modeling of the code loop nonlinearity by the Kalman filter; the second is loss of lock detection. The code loop model embedded in the Kalman filter is quite important, especially as the loop thresholds are approached. Ignoring the nonlinear nature of the detector generally leads to performance degradations. A quasi-linear or describing function-based [30] approach is preferred, where the representation of the detector gain and/or the associated assigned error variance to the code phase measurement depend upon the input signal-to-noise ratio. As the signal-to-noise ratio is lowered, the quasi-linear gain approach calculates a probability that the detector may be operating outside of its linear range [denoted as p_l in (13.37)], and weights the gain in this region (often zero) by the probability in computing a quasi-linear gain:

$$K_q = (1 - p_l)K_l + p_l K_n \tag{13.37}$$

where K_l is the detector gain in the linear range and K_n is the detector gain in the nonlinear range of the detector. The probabilities are evaluated using the uncertainty, embedded in the filter's covariance matrix, projected along the LOS to the satellite which is tracked. Thus, as loss-of-lock conditions are approached, the integrated design recognizes the limited utility of each code phase measurement: in the limit as the effective detector gain becomes zero, it is using only INS information to close the code loop.

Finally, loss-of-lock becomes difficult for either the partitioned or integrated designs as threshold conditions are approached. This is fundamentally because all parameters that can be used to assess lock (see Section 8.13.2) are unreliable. Sophisticated approaches based upon hypothesis testing and parallel filter operations can be considered. Such approaches, for the one or more receiver channels close to threshold, consider the lock state unknown and process the receiver outputs with parallel filters, one assuming the channel (or channels) is (are) in lock, the other assuming that lock has been lost. This can obviously become computationally intractable very quickly, especially as most of the channels are near thresholds, and passing in and out of a lock state. Use of the quasi-linear model for the code (or other tracking loop) detector as described above can make the design highly resistant to missed loss-of-lock detection, as the loop gain becomes 0 as that condition is approached. Thus, appropriate modeling of the code (or carrier) loop nonlinearity can reduce the criticality of loss-of-lock detection.

13.3 Sensor Integration in Land Vehicle Systems

This section examines integrated positioning systems found in land vehicle systems including automotive applications. Low-cost sensors and methods used to augment GNSS solutions are presented, and example systems are discussed.

13.3.1 Introduction

Ever since GPS was first conceived, it was envisaged that receivers would be used for positioning in motor vehicles. By the early 1990s, GPS receiver technology had advanced to the point where GPS products functioned reliably in automotive environments and costs had dropped to a point where widespread use was possible. Now GNSS receivers are used in automotive systems for locating vehicles, tracking vehicles, controlling vehicles, and providing navigation assistance to drivers, and for Advanced Driver Assistance Systems (ADAS).

For many land vehicle applications, GNSS positioning has adequate accuracy and coverage. For example, vehicle tracking systems used for asset management or delivery typically do not need positioning inside tunnels and parking garages, so GNSS without any augmentation provides sufficient coverage. Precision monitoring and control systems on heavy equipment or farm implements use carrier phase tracking and differential GNSS techniques, but additional sensors are not typically required unless vehicle platform attitude is important for the application.

Vehicle navigation systems are available on most vehicle models in the market today. The purpose of these systems is quite simply to help a driver get to a destination as quickly and/or efficiently as possible. A generic vehicle navigation system architecture is depicted in Figure 13.16. Major components include a user interface to enter a destination, a GNSS receiver to determine the absolute position of the vehicle, possible auxiliary sensors for augmenting the positioning solution, access to a digital map database for planning routes and determining maneuvers, and means to present the directions to the driver by voice, graphics, or both via the user interface. Access to digital map data is essential for route planning and guidance and when available in the vehicle, may also be used to improve the positioning as will be discussed in this chapter. GNSS is used for positioning in every vehicle navigation system on the market. Differential GNSS corrections may be provided and applied to improve the positioning accuracy of the solution.

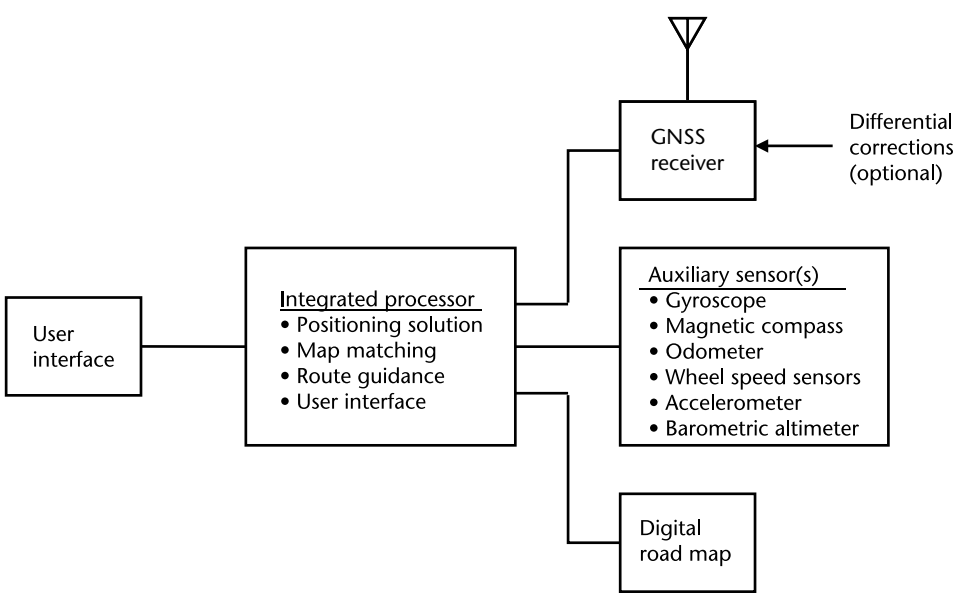


Figure 13.16 Generic vehicle navigation system architecture.

There are many diverse applications that involve vehicle tracking, most of which use GNSS for positioning. In vehicle tracking applications, the position of the vehicle is determined and then is sent via wireless data connection to a centralized monitoring facility or fleet dispatcher. A typical vehicle tracking system architecture is shown in Figure 13.17. Like the navigation system, the tracking system has a GNSS receiver, auxiliary sensors and a computer processor to control the components and calculate the optimized position solution. In addition, there is a wireless data radio for communicating the vehicle position data and possible status to the central monitor. At the central monitor, the vehicle position and other attributes may be displayed or overlaid on a digital map. The digital map can also be used to lookup the nearest street address, a process known as *reverse geocoding*.

There are many wireless technologies that may be used as the data radio including cellular data networks, satellite links, and private radio networks. Some systems track the vehicles on a continuous basis with position reports broadcast at certain intervals, while other systems are designed to record data to be uploaded periodically or on demand. Enterprises that own or operate fleets of vehicles (e.g., taxis, delivery trucks, service vehicles) use vehicle tracking systems to monitor the usage of the vehicles and improve efficiency in logistics through optimum dispatching. Public safety departments (police, fire, ambulance) use vehicle tracking to reduce call response time and to locate workers in the case of distress calls.

Individual vehicles can be located in emergency situations using GNSS and wireless communications. These emergency messaging systems, also known as telematics systems, are offered by many automobile manufacturers today. A generic emergency messaging system architecture is shown in Figure 13.18. Typically, these systems use a cellular phone for wireless data communications because of the dual purpose voice and data capabilities, extensive coverage throughout most developed countries, and relative low cost. These devices are connected to vehicle systems and/or to the vehicle bus and can notify a service provider automatically when an airbag is deployed or some other crash sensor is triggered. The user interface includes

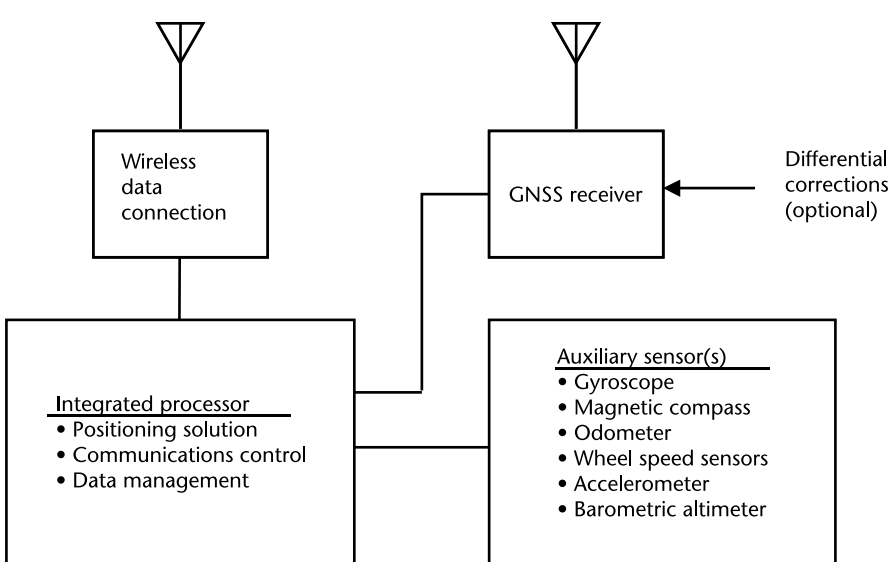


Figure 13.17 Generic vehicle tracking system architecture.

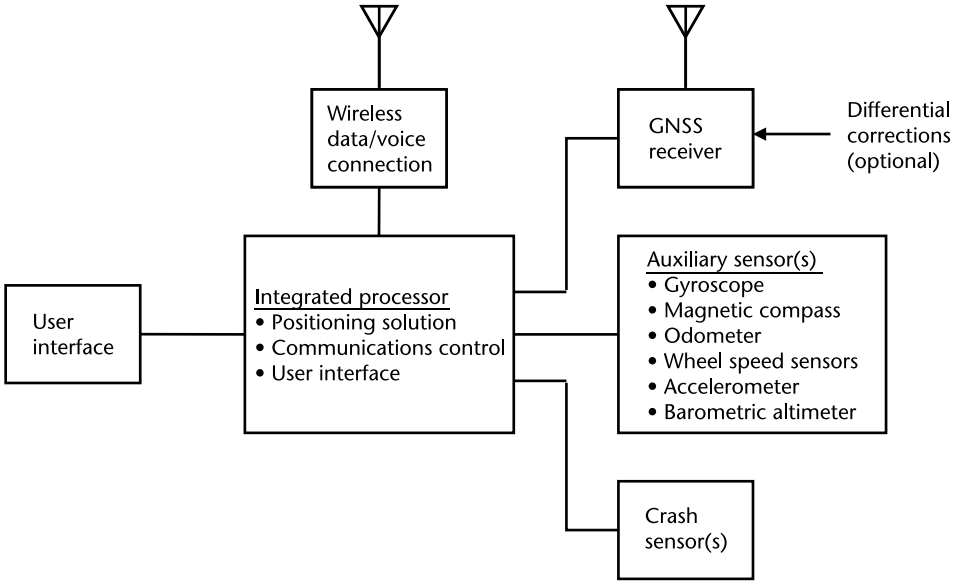


Figure 13.18 Generic emergency messaging system architecture.

one or more buttons to activate the system, a hands-free voice call capability, and may also include a display to indicate status. The GNSS position of the vehicle is sent via the cellular data connection so that emergency services or other assistance can be sent to the exact location of the vehicle. These devices are also used for roadside assistance, theft tracking, and direction assistance and navigation.

In navigation, tracking, and emergency location, the availability of an accurate GNSS position fix is essential. In all of these applications, an L1-only GNSS receiver with 12 or more channels is typically used. The receiver should have rapid signal reacquisition to minimize the effects of urban canyon signal blockage from buildings and structures. The removal of SA had a large impact on the accuracy of low-cost GPS sensors. Adding support for GLONASS and other GNSS constellations provides greater coverage in areas where satellite visibility is severely degraded since more satellites are available in the sky. Using multiple GNSS constellations also further improves the accuracy by providing greater redundancy in these areas with high signal blockage. Differential GNSS is used to improve the accuracy further; however, in the presence of multipath, the multipath error typically dominates all errors that DGNSS can mitigate. Most modern GNSS receivers include support for SBAS signals for ready access to differential correction data. As discussed in Section 12.6.1.2, SBAS including WAAS is a free service and adds little cost to the GNSS receiver. The improvement in accuracy due to SBAS is modest, but meaningful in open areas and integrity information prevents the use of erroneous satellite data. In rare cases, a separate radio may still be used to receive differential corrections such as the Radio Technical Commission for Maritime (RTCM) corrections (see Section 12.5) broadcast by the Nationwide DGPS service in the United States (see Section 12.6.1.1).

GNSS signal blockage in urban canyons and in parking garages can still severely impact the availability of GNSS positions. Figure 13.19 shows the results of a GPS drive test in downtown Phoenix, a moderate urban canyon environment.

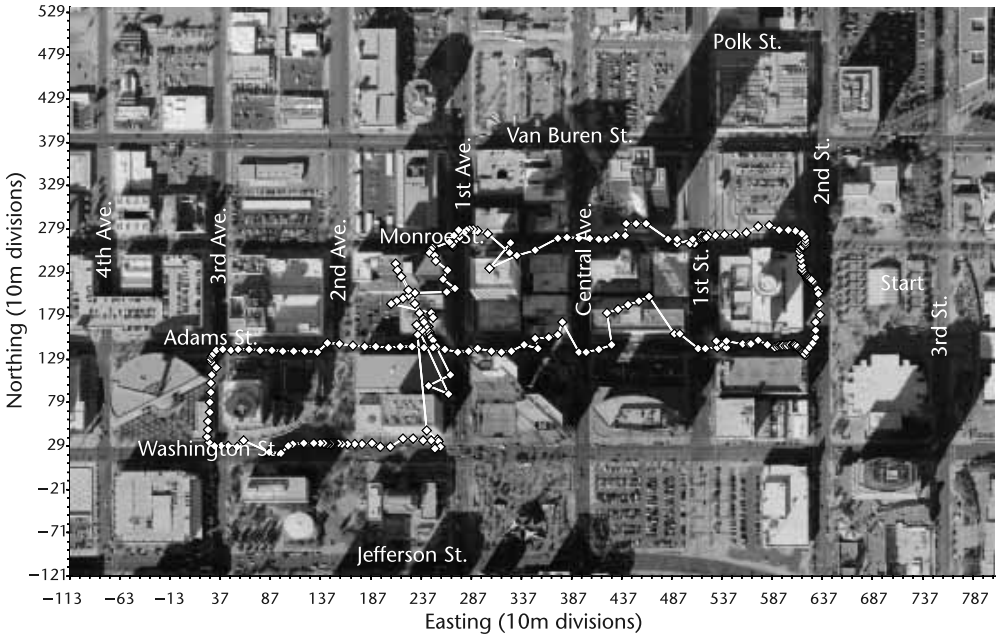


Figure 13.19 GPS performance in moderate urban canyon (Phoenix).

Figure 13.20 shows the results of a GPS drive test in downtown Chicago, a severe urban canyon environment due to the taller and more numerous buildings. The GPS receiver used is a L1-only 12-channel C/A code receiver and the positioning is determined by least squares with no filtering applied in the position domain. Some level of filtering and the use of a high-sensitivity receiver design (whose enhanced acquisition capabilities are discussed in Chapter 8) can be expected to improve the performance. As can be seen, there are several position jumps and gaps, which are caused by signal blockage and reflection due to the tall buildings. In the moderate urban canyon, the jumps are as large as half a block, or 50 to 70m, and there are at least a few position fixes in each block. In the severe urban canyon, the jumps reach 500m and sometimes the receiver goes a block or more without a position fix. Clearly, it is highly desirable to augment the performance of GNSS with additional sensors and filtering methods. Integration of one or more of the auxiliary sensors listed in Figure 13.16 should ensure complete position coverage, and also improve navigation accuracy and reduce susceptibility to gross positioning errors—these issues are discussed further in the following sections.

There is only one factor as important in system design as performance: cost. The overall cost of the system impacts market adoption and then once the systems are made in high volume, every dollar saved in system cost represents a large improvement in profitability. The total annual volume of navigation and telematics systems is in the millions of devices and the cost of the GNSS components including antenna implementation has dropped to a few dollars per unit, and with highly integrated chipsets, the antennaa system is the dominant cost component. There is a natural reluctance amongst equipment manufacturers to include expensive augmentation sensors. Systems integrators are finding ways to use lower grade (and

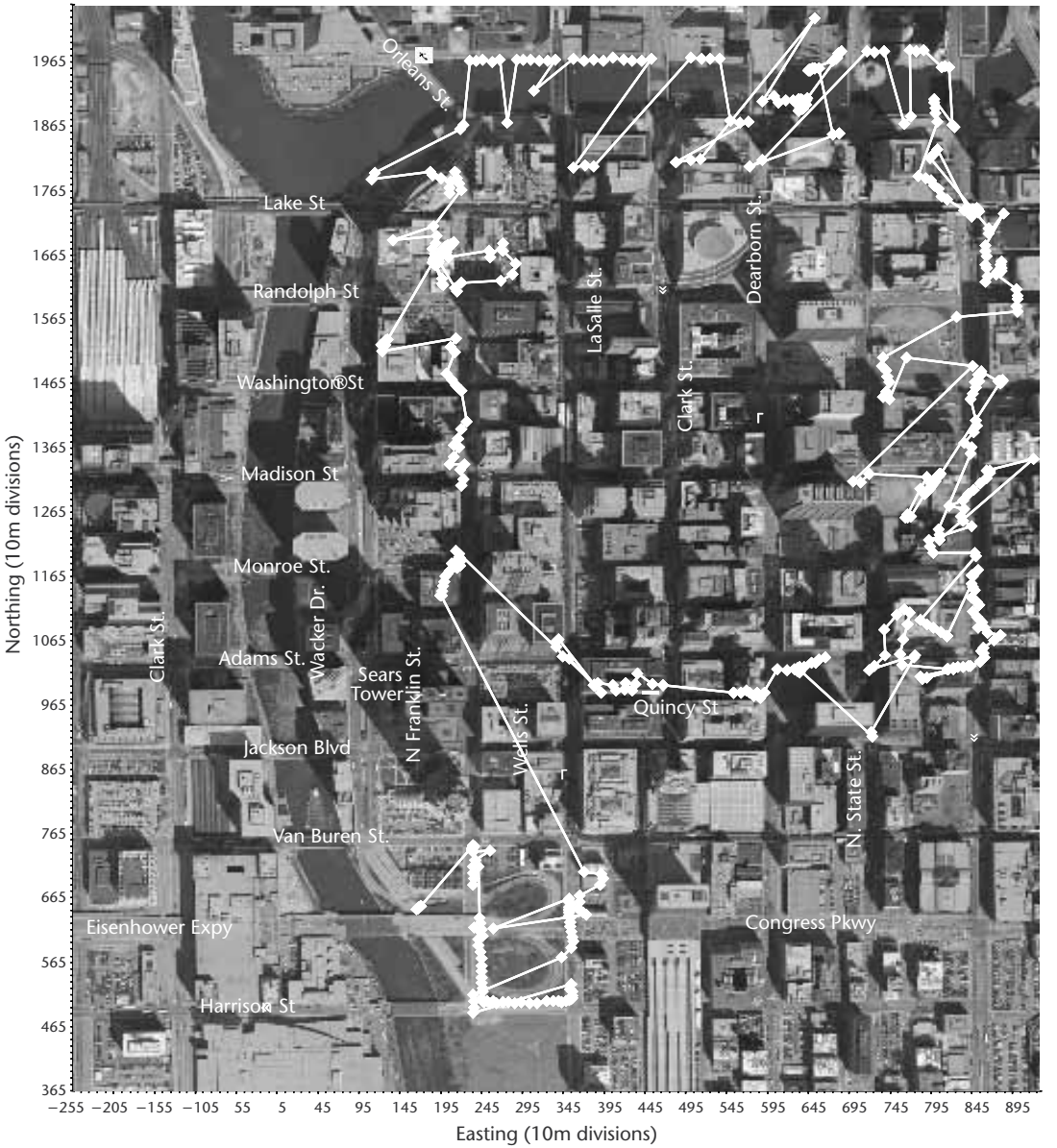


Figure 13.20 GPS performance in severe urban canyon (Chicago).

lower cost) sensors and still achieve complete coverage and improved accuracy over the performance provided by GNSS alone.

13.3.2 Land Vehicle Augmentation Sensors

13.3.2.1 Inertial Systems and Sensors

The use of inertial and various automotive sensors to augment GNSS performance in automotive applications is often termed *dead-reckoning* (DR). Since this term may appear strange to the reader, and since there is some controversy associated

with its origin, some explanation is in order. The term is much broader than automotive in its application and in fact originated long before automobiles were invented. A popularly held belief is that it derives from deduced reckoning, and it is often abbreviated as "ded. reckoning," consistent with this interpretation. Certainly, this view is consistent with its meaning, that is, to deduce one's current position by applying course and distance traveled to a previously determined position. However, according to the Oxford English Dictionary, the phrase "dead reckoning" dates from Elizabethan times, in 1605 to 1615. At that time, it applied to navigation in ships in the absence of stellar observations. With stellar observations, navigation was viewed as navigating "live," working with the stars and the motion of the Earth; however, navigating without sky visibility, by using logs (the process of determining speed by timing the transit of a log dropped in the water from bow to stern), compasses, and clocks, was viewed as navigating "dead," and hence the term dead reckoning. So either expression is valid, both consistent with the modern day application, and both consistent with the abbreviation DR.

Inertial sensors measure change in direction, speed, or orientation directly by means of physically measuring magnetic heading, acceleration, or rotation, respectively. Sensors on vehicles can be used to measure speed by monitoring the drive train or heading by monitoring two different wheels as will be discussed. The use of inertial sensors to augment GNSS in automotive applications offers several advantages over approaches based upon measuring wheel rotation. The quality of inertial sensor information does not vary with tire wear or road conditions, whereas measures of distance traveled using wheel rotation certainly do, as their performance will vary with tire wear, tire slipping, and skidding due to nonideal road conditions. However, very low-cost inertial sensors require nearly continuous calibration: large bias and scale factor errors are typical, as are high sensitivities to temperature variations.

In terms of their usage in automotive and other land vehicle applications, the following inertial system options have emerged as attractive alternatives, with varying limits of practicality:

- Three orthogonal gyros, three orthogonal accelerometers, and three orthogonal megnetometers;
- Three orthogonal gyros and three orthogonal accelerometers;
- Three orthogonal gyros and two level axis orthogonal accelerometers;
- Three orthogonal dual accelerometers;
- Two-level axis orthogonal accelerometers;
- Single longitudinal axis accelerometer and a vertical gyro;
- Single, lateral axis accelerometer with an interface to the vehicle's odometer;
- A single vertical gyro with an interface to the vehicle's odometer.

Obviously, since the last two options above make use of an interface to the vehicle's odometer, they do not take full advantage of purely inertial instrumentation, and so are sensitive to both tire wear and road conditions. To better understand the relative strengths and weaknesses of the various options, it is helpful to first review

the basics of inertial sensing. An in-depth treatment of inertial sensors and systems is beyond the scope of this text and can be found in [6].

A common misconception is that an accelerometer directly measures a component of acceleration: in fact, the accelerometer senses what is often referred to as "specific force" [6], the difference between the component of acceleration along its input (sensitive) axis and the component of gravity along the same axis. Figure 13.21 illustrates the specific force measurement for an accelerometer mounted along the lateral axis of an automotive vehicle. Note that it is implicitly assumed that the input axis of the accelerometer is perfectly aligned with the vehicle's lateral dimension in the figure, which is not realistic. More generally, the misalignment between the accelerometer's sensitive axis and the vehicle's lateral axis is a source of error which must be considered in the design of the navigation system. Neglecting this misalignment in Figure 13.21, the angle φ (in radians) represents the roll of the automobile, or the rotation of the vehicle's vertical axis about its longitudinal axis with respect to the local vertical, b the inherent bias of the accelerometer (in m/s^2), and a_L the lateral acceleration component (also in m/s^2). Accounting also for a dimensionless scale factor error s_L, the output of the accelerometer can be modeled (in m/s^2) as:

$$a_L^m = (1+s_L)a_L + b_L - g\sin\varphi \approx (1+s_L)a_L + b_L - g\varphi \tag{13.38}$$

where the indicated approximation is valid for small roll angles and the m superscript denotes measured value. A similar equation exists for an accelerometer mounted along the longitudinal axis of the vehicle, with independent bias and scale factor errors and with the roll angle replaced by the pitch angle of the vehicle. Equation (13.38) and Figure 13.21 illustrate the difficulty in directly measuring acceleration.

A similar misconception exists relative to the gyro (i.e., that it simply measures the rate of rotation of the vehicle in which it is mounted along its sensitive axis). While this is true to excellent approximation even for low cost gyros, the gyro, in theory, senses inertial angular velocity along its sensitive axis, which will include a component of the earth's rotation rate. It is this property that has been exploited in initializing the heading of inertial systems, using a process generally referred to as *gyrocompassing* [6]. Because the sources of error associated with low cost gyros are orders of magnitude greater than earth rate (e.g., drift rates approaching 1°/s, as contrasted with 15°/h), an alternate means of initializing heading is necessary until low cost gyro technology dramatically improves.

Let us return now to the issue of gyro and accelerometer initial alignment. Any misalignment of either sensor, due either to imperfect mounting of the sensitive

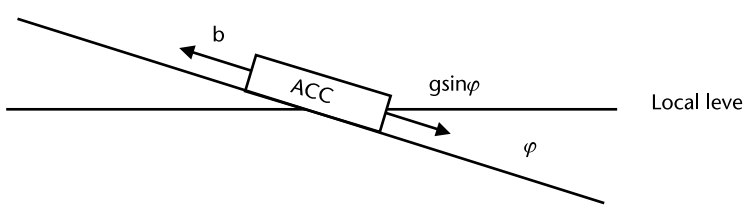


Figure 13.21 Error effects upon lateral accelerometer.

element(s) within the sensor's housing, or imperfect alignment of the sensor housing within the vehicle upon installation, will lead to a cross-axis sensitivity which can be significant. For the lateral accelerometer equation above, a misalignment about the vertical axis of the host will cause the accelerometer to sense a component of longitudinal acceleration, and a misalignment about the roll axis will cause the accelerometer to sense gravitational acceleration, even when the vehicle is level (i.e., at zero roll angle). In each case, the magnitude of the error, for small misalignment angles, is the product of the angle (in radians) and the off-axis acceleration. For example, a 5° misalignment about the vehicle's roll axis will produce an error in the lateral accelerometer of roughly 0.1*g*, or about 1 m/s^2. A gyro mounted with its sensitive axis in the vertical direction, intended to sense the turns of the vehicle, will produce an output that may be modeled (in units of rad/s) as:

$$\omega_H^m = (1 + s_H)\omega_H + b_H + m_\varphi \omega_\theta + m_\theta \omega_\varphi \tag{13.39}$$

where s_H is the gyro's scale factor error, b_H is the gyro bias, m_φ and m_θ are the small angle misalignments (in radians) of the gyro sensitive axis about the roll and pitch axes, respectively, and ω_θ and ω_φ are pitch and roll rate (in rad/s), respectively. In addition, any misalignment of the gyro with respect to the local vertical will appear as a component of gyro scale factor error, since it will contribute an error which is proportional to the angular rate about its sensitive axis. The scale factor error term is expressed in (13.40), where α (in radians) is the misalignment value:

$$\delta s_H = \cos\alpha - 1 = -\alpha^2/2 \tag{13.40}$$

So, for a gyro that is misaligned by 5° relative to the vertical axis of the car, the effective scale factor error is changed by 0.5%, which is generally not significant for low-cost gyros (the nominal scale factor error can be 10 times this level).

Now, given this very basic review of inertial sensing technology, we can return to the issues associated with the options for inertial sensor augmentation of GNSS in automotive vehicles. The first option includes magnetometers to measure compass heading directly once calibrated and magnetic deviation is accounted for. This INS configuration is the most robust; however, it is also the most expensive. Recent advances in low-cost sensors have made this configuration more practical for many applications.

The second and third options do not have magnetometers so monitor heading must change by initial calibration and using gyro/accelerometer solution to deduce heading change. These two options differ only in that the third abandons the vertical accelerometer, based upon the fact that the vertical motion of an automobile is not expected to be significant, and GNSS aided by an altitude constraint may suffice. Referring to (13.38) and Figure 13.21, initialization of the pitch and roll angles for both systems begins (upon turn-on of the system) by assuming that the car is stationary and level, which implies that the accelerometers (after gravity compensation for the vertical axis for the first option) should read as 0. Under zero acceleration, the accelerometers will read the bias error level associated with its current operation. Under the assumptions of the initial-level operation, which is probably a good assumption for an aircraft on a runway, or in a hangar, is generally not a good assumption for an automobile. Even if the road which the car is

parked on is level, the road crown will induce a nonzero roll angle. In general, both the car's pitch and roll angles will be nonzero at IMU turn-on. From (13.38), this implies that each level accelerometer will sense a component of gravity. The sum of the sensed gravity component will be nulled by assumed roll and pitch angles as part of the process which initializes the vehicle's attitude: this so determined pitch and roll will not, in general, match the actual pitch or roll of the vehicle. These initial attitude errors, through the actions of the inertial system, will induce a Schuler oscillation [6] in attitude and position and velocity error in the level axes. The Schuler oscillation period is 84 minutes. This error oscillation, if not disturbed by other error inducing effects (e.g., maneuvers), will persist until the Kalman or integration filter has had time to estimate the sensor errors. Typical Kalman filter designs will be addressed in Section 13.3.3.

Unlike the initialization of pitch and roll, however, because low-cost gyros have bias errors that are very large relative to earth rate, the heading of the vehicle must be initialized by an auxiliary sensor (e.g., a magnetic compass), and/or use of a GNSS determined heading, and/or use of the vehicle heading as last computed by the navigation system. In the case of a GNSS heading, care should be taken that a minimum speed has been attained, and that at least four GNSS satellites are tracked to ensure adequate accuracy.

Returning to the two-accelerometer INS, use of a vertical accelerometer in an INS brings a potential stability problem. As is well-known [6], an INS vertical channel is inherently unstable due to the dependence of gravitational acceleration upon altitude (in general a gravity model is needed to remove gravitational acceleration from the accelerometer outputs to enable sensing of inertial acceleration). The fact that modeled gravitational acceleration may decrease with altitude increase leads to an effective positive feedback loop in the error equations for the vertical channel [6], which produces an exponential error growth. This error growth will produce more than a doubling of altitude error roughly every 10 minutes if not corrected. Thus, an independent source of altitude information is needed, which could be provided by an additional sensor (e.g., a barometric altimeter) or an altitude constraint (e.g., the assumption that the vehicle is at mean sea level or at the known altitude for a certain road).

Because gyro design and development are generally more complex and less reliable than accelerometer design and development [31], it is attractive to consider an accelerometer-only INS, which develops angular acceleration estimates by placing dual accelerometers at known displacements (referred to as *lever arms*) from the vehicle's center of gravity. For example, as illustrated in Figure 13.22, the two accelerometers illustrated could be used to sense both linear and angular acceleration. Before discussing a reference [32] where such a prototype system is constructed, some high-level comments are worth making. First, since we have replaced the gyro, an angular rate sensor, with an angular acceleration sensor, accelerometer errors will have a different effect upon the INS position and velocity error. Any biases in the accelerometers will produce a time-varying rate error in angular velocity: the accelerometer biases add, while error effects due to sensing of gravitational acceleration from pitch or roll error is largely cancelled. The quality of the angular acceleration sensing improves as the separation between the accelerometers increases. To understand this, consider the treatment in (13.41), valid for two accelerometers placed along the longitudinal axis of the vehicle:

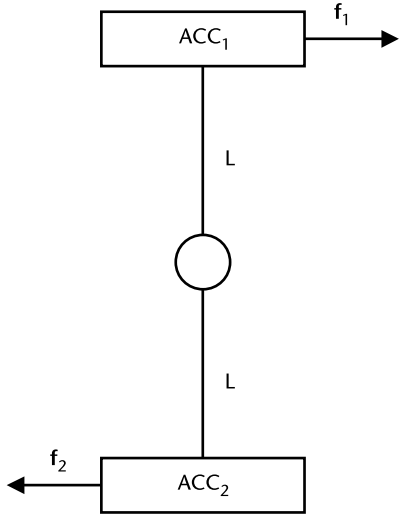


Figure 13.22 Dual accelerometer approach to linear and angular acceleration sensing.

$$a^m = (f_1 - f_2)/2 \tag{13.41a}$$

$$d\omega/dt^m = (f_1 + f_2)/2L \tag{13.41b}$$

$$f_1 = (1 + s_1)a + Ld\omega/dt + b_1 - g\sin\varphi \tag{13.41c}$$

$$f_2 = -(1 + s_2)a + Ld\omega/dt + b_2 + g\sin\varphi \tag{13.41d}$$

$$\delta(d\omega/dt) = \left[(s_1 - s_2)a + b_1 + b_2\right]/2L \tag{13.41e}$$

In (13.41), (13.41a) and (13.41b) represent the equations that would be used to measure linear and angular acceleration, labeled a^m and $d\omega/dt^m$, respectively; (13.41c) and (13.41d) represent the error equations associated with the measured quantities in (13.41a) and (13.41b): thus, a represents the true acceleration of the vehicle along the sensitive (lateral) axis, and b_1 and b_2 are the accelerometer biases, all preferably represented in units of m/s^2. As used previously, φ is the roll angle of the vehicle in radians, and g represents gravitational acceleration in m/s^2. The accelerometer scale factor errors (unit-less quantities) are denoted as s_1 and s_2, respectively. The lever arm is represented by the variable L, expressed in meters to maintain consistent units. Finally, note that (13.41c) is an equation for the rate of change of the error in sensing angular rate (i.e., yaw rate, which is roughly the heading rate), which would typically be modeled in a Kalman filter which attempted to reduce this error by processing GNSS measurement data.

Thus, the error contributors to angular acceleration, the individual accelerometer bias and scale factor errors, b_1 and b_2 and s_1 and s_2, are reduced by increasing the lever arm, L, between each sensor and the center of gravity of the vehicle. In

the specific case illustrated by Figure 13.22, best performance would be achieved by placing one accelerometer near the front of the car, and the second near the rear of the car. The lever arm does not affect the quality of the determined linear acceleration. Since the accelerometer bias contributes to an angular rate bias in this formulation, it produces different position and velocity error behavior than its gyro bias counterpart. As is well known [6], level axis gyro bias errors produce biased velocity errors superimposed on a Schuler oscillation in the level axes. The bias component of the velocity error can dominate the INS drift for periods which are less than the Schuler period, leading to the familiar "nm/h" rating often associated with inertial systems [3]. A bias angular acceleration error can therefore be expected to produce a ramping velocity error over a similar time period.

The concept of using accelerometers to sense angular acceleration is not new [33]. In the 1990s, this concept received new attention, driven largely by the presence of very low-cost microelectromechanical sensors (MEMS) technology that could be used to produce suitable accelerometers for vehicles for a fraction of the cost of gyros [34, 35]. One study [36] focused on the placement of accelerometers within the vehicle for best performance. Another treatment [32] attempts to make use of existing accelerometers [e.g., as could be associated with air bag deployment or the vehicle's antilock brake system (ABS)] distributed throughout the car to support an inertial navigation capability. Tests of a prototype system have demonstrated that the accuracy of measured angular accelerations using accelerometers is nearly equivalent to that provided by low-cost gyro sensors.

Use of single accelerometers aligned with the lateral and/or longitudinal axis of the vehicle is an option worthy of consideration. The longitudinal accelerometer measures vehicle accelerations and decelerations, which, once integrated, could potentially replace use of the vehicle's odometer. The lateral accelerometer could potentially replace a heading or heading rate sensor, since a lateral acceleration is generally indicative of a turn: the product of the vehicle's speed and the turn rate is the lateral acceleration of the vehicle. However, use of single accelerometers has its drawbacks. As previously discussed, both accelerometers will generally sense a component of gravity, due either to initial misalignment of the sensor as installed in the vehicle, or the pitch (affecting the longitudinal accelerometer) and roll (affecting the lateral accelerometer) of the vehicle. Although the pitch and roll of a vehicle during normal operation are expected to be small, the error effect, if uncompensated, can be significant. Relatively high-frequency pitch-and-roll variation, as could be induced by road or speed bumps, is not as troublesome as a steady offset. A 5°, steady roll angle induced by the crown of the road induces an effective acceleration error of 0.1*g*, or roughly 1 m/s^2. Without compensation, this will integrate to a velocity and position error, even when the vehicle is stationary; for example, in 10 seconds, roughly 50m of cross-track error will develop. In addition, since the lateral accelerometer measures the product of heading rate with the vehicle's speed, heading changes may be very difficult to detect at low speed. Similarly, a steady climb or descent on a road will be incorrectly interpreted as an acceleration or deceleration of the vehicle by the longitudinal axis accelerometer, which, without compensation, will be integrated into significant along track velocity and position error.

Finally, the use of a low-cost gyro to track the heading changes of the vehicle is an attractive option used in several of the current navigation systems. The vehicle's

pitch and roll have a second-order effect upon the gyro scale factor, as indicated in (13.14), but this should be small relative to its nominal scale factor error.

Given the preceding discussion on inertial system options, the error characteristics of gyros and accelerometers can now be addressed. For the low-cost sensors considered for automotive applications, the bias and scale factor errors can be very large relative to those of gyros and accelerometers associated with commercial grade systems; for example, for the gyro, a bias of several degrees per second is expected, and a scale factor error as large as 5% is possible. A summary of gyro and accelerometer bias and scale factor errors for different applications may be found in [33]. These errors can be calibrated using GNSS and other means. For instance, an estimate of the gyro bias can be obtained each time the vehicle is stationary in a calibration procedure referred to as a *zero velocity update* or ZUPT. However, the errors can also be quite unstable and have high-temperature sensitivities.

Figure 13.23 illustrates the laboratory measured gyro bias temperature sensitivities for two samples of a low-cost, vibrational gyro. The term vibrational indicates that the gyro has a vibrating element which senses angular rate through the Coriolis force, which is exerted on the vibrating element. This force is directly proportional to the angular rate of rotation, and is measured through the actions of the gyro electronics. Figure 13.24, abstracted from [37], illustrates the driving and detection and control mechanisms for the Murata Gyrostar gyro, as an example of vibrational gyro technology. The illustrated bar has a triangular cross-section,

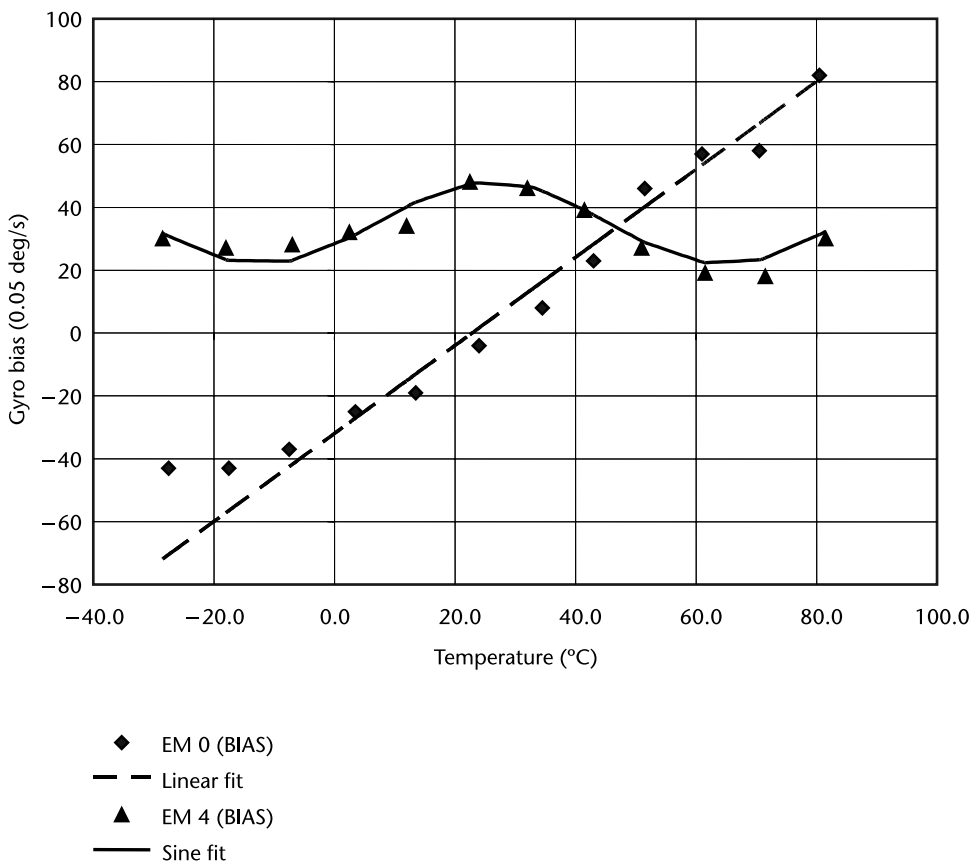


Figure 13.23 Bias versus temperature for two low-cost vibrational gyro samples.

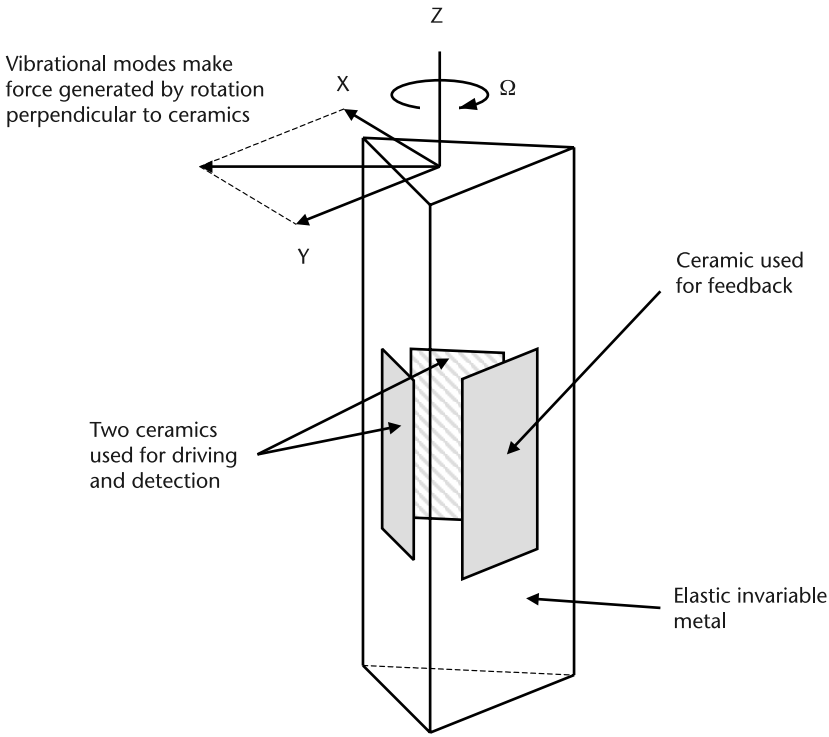


Figure 13.24 Gyrostar free-free bar and ceramics [37, Figure 3.3].

with the bar faces forming an equilateral triangle. Two sides are used for driving the beam at a resonant frequency and detecting the Coriolis force; the third face is used to close the vibration control loop.

Returning to Figure 13.23, several conclusions relative to temperature sensitivities can be drawn from just these two gyro samples. First, the temperature sensitivity can be very large. For the sample denoted EM 0, the sensitivity is roughly linear over the temperature range, and its magnitude is 0.07°/s/°C. If the sensitivity is ignored, and the gyro is in an agile temperature environment (e.g., a car left outside overnight in the winter in Boston heating up), the gyro will require frequent calibration. At constant speed, an uncompensated gyro bias error will produce a quadratic growth in cross-track position error proportional to the product of the bias and the speed of the vehicle. Second, the temperature sensitivity is individualized to each gyro (i.e., if compensation is desired, every gyro must be tested prior to installation in the vehicle, unless this requirement is levied upon the manufacturer). Such requirements inevitably increase the cost of the gyro and implementation. The sample denoted as EM 4 has a nearly sinusoidal variation that is relatively minor over the temperature range tested. Given a temperature curve for a gyro or accelerometer bias or scale factor, it is tempting to use a curve fit or other means to compensate its output in real time. There are several issues here, in addition to the expense associated with the curve-fit generation for each gyro sample. First, a temperature sensor will be needed to perform the compensation, and the sensor must certainly be installed near the gyro or accelerometer sensitive element. Although some sensor assemblies may provide temperature information, all do not. The second issue that must be considered is the stability of the underlying sensitivity itself.

Can the temperature compensation curve, without adjustment, be used for several months or even several years? The answer to such a question may not be known by the manufacturer, so it is therefore advisable to at least monitor the curves for stability. This subject is addressed further in Section 13.3.3.

The use of gyros and accelerometers based upon MEMS technology first received much attention for military systems [38] as a result of expected cost, weight, size, and power savings. Since their introduction, MEMS sensors have been widely adopted for use in vehicles and consumer electronics. Due largely to their use as sensors for air bag deployment in cars, accelerometer development was initially more mature than that for gyros. The use of gyros in image stabilizers for cameras drove further advances in gyro stability and lower cost. Cost versus performance is a constant trade-off. In [39–43], accelerometer developments are described that achieve navigation grade accuracies, that is, with bias errors as low as 20 μg, and scale factor errors approaching 50 ppm. Key developments in gyro technology are summarized in [44–46]. Although 1°/h gyro bias performance is predicted, reported performance levels are limited at 10°/h, with scale factor errors approaching 500 ppm. MEMS applications in the commercial were very promising early on [47] and have since proven a key component in billions of devices. Characterization of MEMS-based sensors for land vehicle applications is treated in [48]. Like the existing sensors (e.g., the vibrational gyros), MEMS gyros, as well as accelerometers, are expected to have significant temperature sensitivities which must be compensated to realize their full performance potential. MEMS sensors are lower cost and lower performance than traditional inertial sensors; therefore, system design and performance requirements drive the selection of sensors used.

13.3.2.2 Map Databases

As mentioned in Section 13.3.1, the emergence of high-quality, affordable, digital maps was a significant factor in the wider acceptance of automotive navigation. Digital road maps are not only an essential component for selecting destinations, pathfinding, and route guidance in navigation systems, but also a high-value addition to the positioning subsystem.

There are several companies that publish digital road map data for navigation on a global basis. Navteq, one of the pioneers in digital maps for navigation, was acquired by Nokia in 2007 and by 2012 was part of the Here business unit within Nokia. In 2015, Nokia sold the Here mapping unit to a consortium of German automakers, including Audi, BMW Group, and Daimler, and since then Here has operated as an independent company distributing digital map data and mapping tools for vehicle navigation systems and mobile devices. The Here maps have the highest global market share in installed vehicle navigation systems. TeleAtlas group was established by Robert Bosch GmbH and Janivo BV in 1995 to speed up the collection of digital road map data in Western Europe and publish it in a uniform format [49]. TeleAtlas acquired Etak in 2000 and acquired Geographic Data Technology (GDT) in 2004 to expand coverage in North America and globally. TeleAtlas was acquired by TomTom in 2008 and now operates as a wholly owned subsidiary of TomTom. TeleAtlas maps are used in TomTom and other navigation systems and also in various Internet map portals. Google is the most recent major entrant into the navigation mapping market, launching a Web-based mapping service in 2005.

Google first licensed map data from companies including Navteq and TeleAtlas and then gradually built their own mapping capability outfitting cars with cameras and other sensors. In 2007, Google launched route planning and driving directions, and then in 2009, Google introduced free turn-by-turn navigation on a mobile phone application. Here, TeleAtlas, and Google have extensive digital road network databases attributed for navigation covering the United States, Canada, Europe, Asia, and other emerging markets worldwide. The accuracy of these databases, as determined by comparing road centerline vectors to ground truth, ranges from under 5m in urban areas to 20m or more in rural areas. New initiatives are under way to map road center lines to better than 1-m accuracies and to include vertical information for use in advanced driving systems using aerial imagery, dedicated mapping vehicles, and recorded traces of GNSS data from vehicles [50, 51]. Over time, both the positional and topological accuracies are being improved through GNSS surveying, photogrammetry, and other data acquisition methods [52].

Even before GPS became a viable positioning system for use in commercial products, digital road maps were used as a component in the positioning subsystem of navigation systems. The Etak Navigator, introduced in 1984, consisted of a cassette tape player, an 8086-based computer, dual odometers, a compass, and a small cathode ray tube (CRT) display. A digital road map was stored on the cassette tape. The system used the compass, differential odometry and map matching to position the vehicle [53–55]. Map matching is the process of correlating the vehicle path with a drivable path in the digital road map [56]. The map and the vehicle position were displayed, and as the vehicle moved, the map would move, keeping the vehicle symbol in the center of the screen. With map matching, a basic assumption made is that the vehicle is on the road network so that the calculated vehicle position is constrained to one of the road segments in the map. As the vehicle travels, the dead reckoning sensors provide a path of the vehicle, which is matched up with road segments in the map database that have the same approximate shape and orientation in order to determine the position of the vehicle.

One major challenge with map matching before GPS was available was the initialization of the system when the starting position was not known. In early navigation systems, the user sometimes had to be prompted to enter the current position. This was difficult if the user did not know where he or she was. With GNSS, the absolute position is readily determined, and in time, GNSS receivers were added to navigation systems. Initially, GPS was only used to get the DR/map matching system started or to detect large errors. Then systems emerged where the GPS/DR trace was compared with the digital road map in order to find the most probable location of the vehicle [57]. Modern navigation systems rely primarily on GNSS and use DR and map matching to correct GNSS errors and bridge the coverage gaps.

A robust map matching implementation uses confidence measures to determine all possible road segments in the map that the vehicle could be traveling on as illustrated in Figure 13.25 [58]. As the vehicle travels, distance traveled and changes in direction are used to continuously determine the shape of the route traveled; this shape is used to match the road network in the map through shape correlation. When an accurate heading is known, the list of roads is reduced to those that have a bearing within a tolerance of the vehicle heading. When the vehicle makes a turn, the list of candidate segments is further reduced based on examining the topology

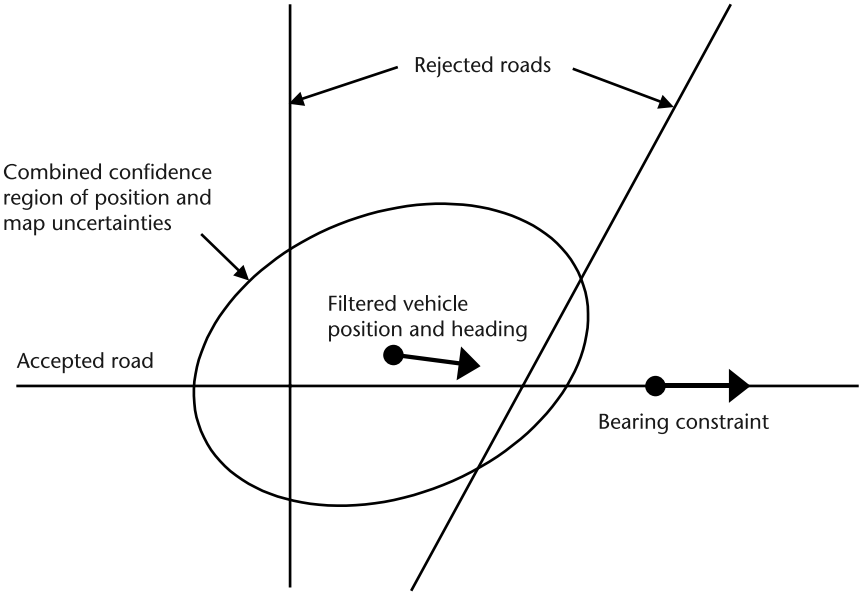


Figure 13.25 Road selection and map aiding [58, Figure 7.3].

of the road network to find candidate segments that have a turn in the direction the vehicle turned. Through this process, the list of possible vehicle positions is eventually reduced to a unique segment and the confidence in the positioning solution increases accordingly. When there is only one possible vehicle position, the map matched position solution will have a small confidence region and therefore can be considered highly reliable. If a position jump or a turn occurs that introduces additional potential positions in the road network, then the confidence region should grow reflecting a lower confidence in the map-matched solution.

In order to support map matching, the map data should have high position accuracy, ideally better than 5m, to minimize incorrect road selections. The map data should also be topologically correct, reflecting the real-world road network, so that the algorithm does not get confused if the user drives on a road that is not in the database. The expected accuracy of the road center line data should be used in the map-matching process to determine the overall confidence region of the map-matched position solution.

Once a match is determined, the vehicle position is then displayed on the matched road segment and used for the route guidance instructions. The map can also be used as a sensor itself to provide useful information to the positioning subsystem and/or to calibrate inertial and other dead reckoning sensors. These capabilities have been broadly referred to as map aiding [58] and map calibration.

Map aiding is most useful when map matching has determined that the vehicle has just turned a corner, in which case the vehicle position is in close proximity to the intersection of two streets, which has a known location in the map database. This reference position may be treated as a single position fix by the integration filter (see Section 13.3.3 for further discussion), which serves to correct or improve the accuracy of the absolute position determined by GNSS. Further, if map matching has determined, with high probability, that the vehicle is traveling on a specific road in the database, and that road is straight, then a heading fix may be generated

for the integration filter (see Section 13.3.3) based on the bearing of that road segment according to the map database as shown in Figure 13.25. Another way to utilize the heading information after a turn is to impose a constraint on the model to force the heading of the vehicle to match the bearing of the road. Map feedback can be used instead of dead reckoning sensors to improve the performance of GNSS in low-cost navigation systems [58].

In addition to the horizontal position components of road vectors, ground elevation data may be used to augment the performance of GNSS. A digital terrain model (DTM) is a representation of the Earth's surface that can be used to extract elevation data. A digital elevation model (DEM) is a type of DTM with a regularly spaced grid of elevations corresponding to the elevation of the Earth's terrain at that point. Modern DTMs are derived from airborne or satellite-based remote sensors, are georeferenced using GNSS coordinates, and have vertical accuracies better than 10m, in some cases better than 5m.

Terrain elevation can be used to improve the accuracy associated with GNSS fixes for land applications. As is well known, and addressed previously in the text, the vertical axis is the weakest part of the GNSS solution. Terrain elevation data, if sufficiently accurate, can be added as a constraint to an LS or WLS GNSS fix, or added as a measurement to a real-time Kalman filter. To apply a height constraint, an approximate or previous position can be used to extract the corresponding elevation from a DTM, DEM or other source of elevation data. If the terrain elevation varies greatly in the vicinity of the position, iteration may be necessary to converge on the solution. Using a DEM for this purpose may be easier from a computational perspective since it would involve a simple value lookup and interpolation based on the coordinates; however, a large amount of storage would be required for the DEM. A DTM that has the elevation data organized into vectors would use less storage, but would require more complicated computations to determine the elevation at a specific point. Elevation data can also be integrated into digital road maps as attribute data, which would simplify elevation lookup and keep the storage requirements lower. Terrain elevation data is now being used to augment GNSS in driver safety applications that monitor speed and slope.

Map calibration is very similar to the process of using GNSS data to calibrate inertial and other dead reckoning sensors. For example, with the same set of conditions that support the heading fix generation, the constant road heading may be used to calibrate a low-cost gyro or magnetic compass: since the road heading is constant, the gyro reading is then a direct measure of its bias. Another example is when the vehicle makes a turn at an intersection, the change in heading between the inbound segment and the outbound segment can be used to calibrate a heading sensor. With the current performance of GNSS, map calibration of sensors is less common than it once was.

As discussed, digital road map data is a valuable component of the positioning subsystem in vehicles. However, the usefulness is limited by the accuracy of the data. As roads are constructed and rebuilt over time, the geometry and connectivity of the digital road map database segments changes. An incorrect road segment in the database will have a negative effect on the position computation when used for map matching or map aiding. The internal weighting must accommodate for this possibility and allow the system to correct itself back to another segment should the probabilities dictate. Storing and updating map data on mass storage media

within the vehicle are especially susceptible to this condition since consumers and commercial operators do not always update their map data right away and even if they do, there may be temporary changes due to construction or weather that are not accounted for. The emergence of better connectivity in vehicles enables systems to utilize map data updates from a server so that the system can download and use the most recent map data and even the effects of traffic, weather, and construction in near real time such as the case with Google Maps.

13.3.2.3 GNSS

As mentioned in Section 13.3.1, the discontinuance of SA enabled commercial use of GPS at close to full accuracy with low-cost stand-alone receivers, excepting for the inability to remove the majority of the ionospheric delay. Now that a secondary civilian signal has been introduced, it is possible to remove the effects of the ionosphere with a dual-frequency receiver. Previous sections in the text have identified and discussed the major sources of GNSS errors, both in the measured pseudoranges and delta ranges or Doppler measurements, and the determined positions and velocities. Of interest here are the sources of error in the GNSS-determined speed and heading, and sources of error which may be unique to the automotive environment. GNSS-determined speed and heading are useful in calibrating automotive sensors that are then used as sources of speed and heading information when GNSS is not available. This direct comparison enables rapid calibration of sensor errors when GNSS is accurate. For errors that are small relative to the vehicle speed, the error in the GNSS-determined speed and heading can be expressed as:

$$\delta v = \left(v_n \delta v_n + v_e \delta v_e\right)/v \tag{13.42}$$

$$\delta H = v_n \left(v_n \delta v_e - v_e \delta v_n\right)/v^2 \tag{13.43}$$

where δv_n and δv_e are the north and east velocity error components; v_n and v_e are the north and east velocity components; δH and δv are the heading and speed errors, respectively; and v is the vehicle speed in a horizontal plane.

All velocity components (both whole value and error quantities) in (13.42) and (13.43) should be expressed in consistent units (e.g., m/s for velocity, and radians for the heading error). Equations (13.42) and (13.43) can be derived by simply perturbing the equations for speed and heading expressed in terms of the velocity components.

An additional source of error in the GNSS-determined heading is worthy of mention and can be a significant error, depending upon the antenna placement in the vehicle. The GNSS antenna will generally not be installed close to the center of turn rotation of the car. As illustrated in Figure 13.26, where the antenna is installed a distance L from the center of rotation of the vehicle, the GNSS receiver will detect the heading rate multiplied by the distance L as a velocity component orthogonal to the true velocity of the vehicle. Since GNSS (in a nonmulti-antenna configuration) can only derive heading from the determined velocity components, a heading error given by (13.44) results:

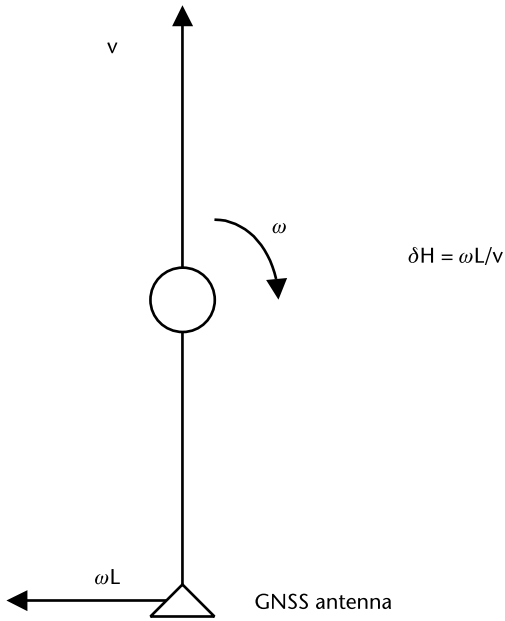


Figure 13.26 Effect of antenna placement on GNSS heading.

$$\delta H = \omega L / v \tag{13.44}$$

where ω is the heading rate of the vehicle, typically represented in radians per seconds; L is the distance from the center of rotation in meters; v is the vehicle's speed in meters per second; and δH is the resultant heading error in radians.

To assess the magnitude of this error source, assume that the GNSS antenna is 1m from the center of rotation, the heading change rate is 20°/s, and the vehicle speed is 5 km/h. A heading error of more than 14° results, which is generally unacceptable for navigation purposes, producing cross-track error which is more than 20% of distance traveled. The error persists only as long the vehicle is turning. If the lever arm L can be measured, the error effect can be compensated. However, at a minimum, the real-time navigation filter should recognize this error effect in its weighting of GNSS headings in turns. If the system employs a digital road map database, map-aiding can be used to calibrate or even determine the lever arm distance by observing the heading rate of change and speed directly, then determining the change in heading between the heading of the road prior to a turn and that of the road after a turn to determine the heading error and then solve for the lever arm distance in (13.44).

As efforts continue to lower acquisition and tracking thresholds for GNSS receivers, additional sources of error must be considered, including false signal acquisitions and tracking of reflected signals (commonly referred to as multipath). As discussed in Section 8.5, acquisition of signals below normal thresholds requires longer coherent and noncoherent integration times. As signal-to-noise ratio thresholds for acquisition and tracking are lowered by more than 20 dB, the potential for cross-correlation (i.e., declaring detection for a higher power signal with an incorrect PRN code) increases. In addition, the conservatism associated with normal detection thresholds (i.e., the threshold placed upon the peak to noise floor ratio) may

be relaxed in order to increase coverage. Alternate tests may also be employed [e.g., use of a neighbor test, where a detection may be declared if the peak magnitude and the next largest peak magnitude are in neighboring code phase positions (i.e., separated by one-half chip)]. Such relaxations of conservatism in detection inevitably bring a higher probability of false signal acquisition (i.e., interpreting integrated noise as a signal). Both false signal acquisition and cross-correlation will produce pseudoranges that are grossly in error; generally, these errors do not persist as the transition is made to tracking the direct signal. If this should happen for a short period of time, however, statistical rejection tests employed by the navigation filter should remove them.

Reflected signal tracking is a serious problem that can arise in urban canyons and can occur when the direct signal path is obscured by a high-rise building, yet a reflected signal path is visible to the GNSS receiver. Note that this condition is not truly multipath, as the direct path cannot be seen, and only the reflected version is tracked; however, the nature of the error introduced is similar. The reflected signal will be attenuated relative to the direct path, and the geometry of the reflection cannot persist indefinitely. Pseudoranges presented to the navigation filter will have additional, unexpected error due to the additional range delay associated with the reflected path, and Doppler measurements derived from the reflected signal can be significantly in error. The measured Doppler component due to receiver motion may be opposite in sign to the actual Doppler component induced by the receiver motion. It is generally a function of the velocity of the vehicle relative to the surface which is causing the reflection, as illustrated in Figure 13.27. As was the case with the false acquisitions, we must rely upon the integration filter's statistical rejections to preserve acceptable navigation performance.

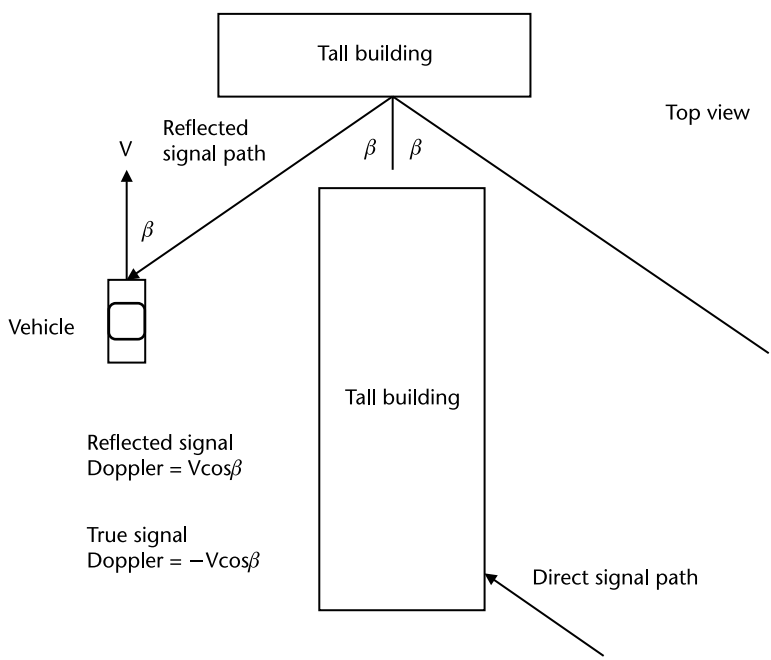


Figure 13.27 Illustration of reflected signal tracking geometry.

13.3.2.4 Transmission and Wheel Sensors

The use of elapsed distance traveled information available in the vehicle is generally a low-cost, high-value augmentation of GNSS. Vehicle transmission and wheel sensors can be used to determine the speed and heading changes of the vehicle. Depending upon the type of sensor utilized, the distance determination can become unreliable at low speed; if variable reluctance sensing [59] is used, the sensor output becomes zero as the magnetic flux change becomes small as illustrated in Figure 13.28. When the motion of the protruding tab through the magnetic field becomes too slow, the signal processor will not be able to detect a pulse, corresponding to a certain distance moved by the wheel. Depending on the specific sensor utilized and the signal processing circuitry, speeds of 0.5 m/s to several meters per second may be undetectable. However, Hall-effect sensors [59], whose output is position rather than rate sensitive, can detect vehicle speed reliably down to stationary conditions. For this reason, Hall-effect sensors are preferred, but are generally more expensive to install.

Independent of the type of sensor utilized, transmission odometer-based speed determination can be unreliable under three distinct conditions: wheel slipping, wheel skidding, and vehicle motion when the tires are stationary.

The first problem can be reduced by installing sensors so they detect the motion of the nondriven wheels (e.g., the nondriven wheels of a front wheel drive vehicle are the rear wheels and vice versa). Otherwise, tire slippage can lead to gross positioning errors in the dead reckoning (DR) system, since the sensed speed will greatly exceed the actual speed of the vehicle. Some slippage will occur, even with

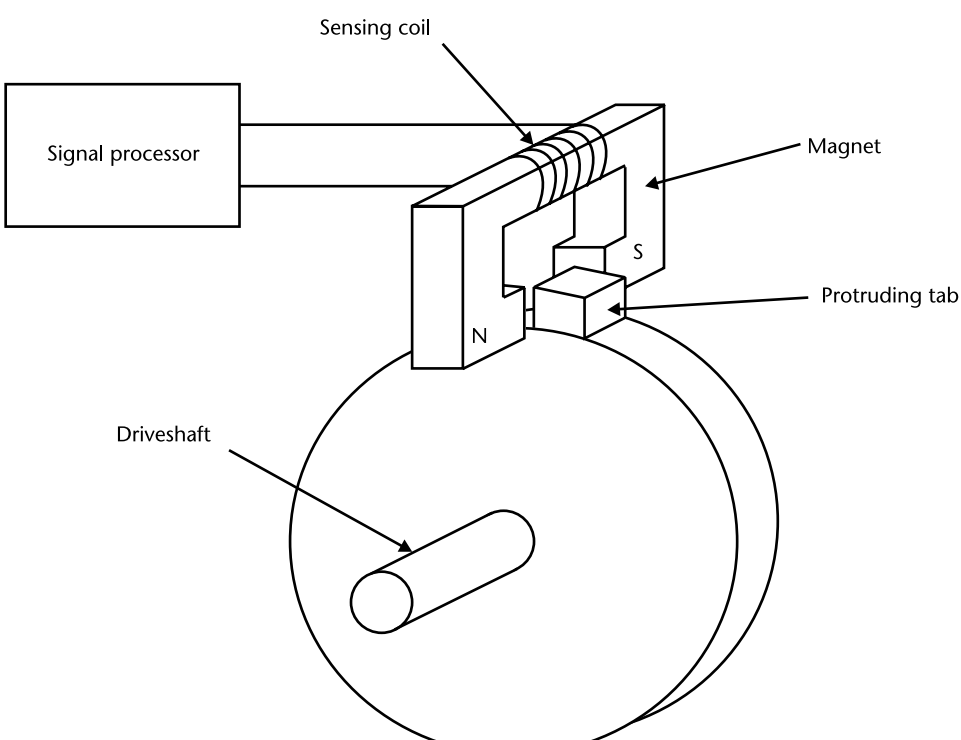


Figure 13.28 Variable reluctance rotation sensor [37].

nondriven wheel installation, but generally only during braking and cornering. The effects of wheel skidding is much more difficult to eliminate; however, the potential for it can be reduced significantly by the use of an ABS. Detection of and recovery from skidding conditions should be important considerations in the design of the sensor integration algorithm (this concern is discussed further in the next section). Finally, motion of the vehicle when the tires are stationary (e.g., as could occur when the vehicle is transported with a tow-truck or onboard a ferry) can also lead to excessive positioning error and necessitates the need for a second recovery mode in the sensor integration algorithm.

Excepting these anomalies, speed determination is affected by the ability to measure distance traveled using the circumference of the wheel. Typically, 24 to 48 pulses are generated for each wheel revolution. The scale factor that converts pulse counts to distance traveled can be accurately calibrated at installation by driving a known distance. However, slow variations in tire pressure can degrade the initial calibration and, over time, affect the accuracy of the scale factor. Wheel sensors suffer from the same problems described for the transmission sensors but with potentially more serious error conditions. In addition to speed, individual wheel sensors can be used to determine heading changes of the vehicle. This is done by measuring the difference in the distance traveled by each nondriven wheel, a technique known as *differential odometry*. If the vehicle is making a right turn, the left wheel has to travel farther than the right wheel to complete the turn and vice versa. Assuming that the sensors are installed on nondriven rear wheels, the following equation can be used to compute heading change, ΔH, and is illustrated in Figure 13.29:

$$\Delta H = (d_R - d_L)/T \tag{13.45}$$

where d_R and d_L are the distances traveled by the right and left wheels, respectively, and T is the wheel track (the distance between the tires).

In (13.45), the right and left wheel distances are represented in meters, as is the wheel track, resulting in the computed heading change ΔH in radians. Note that (13.45) is valid only when the sensors are installed on the rear wheels. When the front wheels are used, the geometry changes, since the front wheels develop wheel angles, denoted by γ in Figure 13.30.

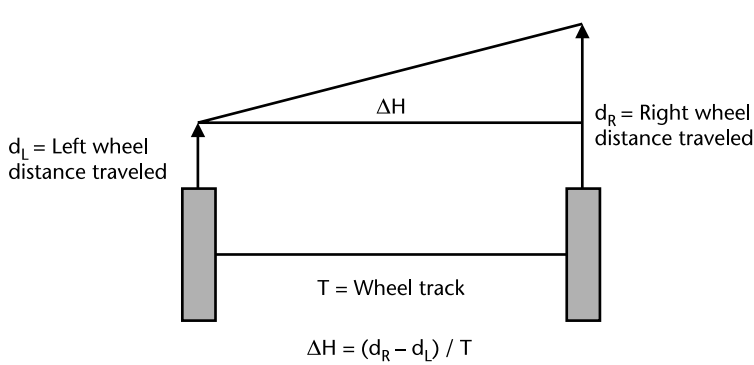


Figure 13.29 Heading change determination using rear wheels.

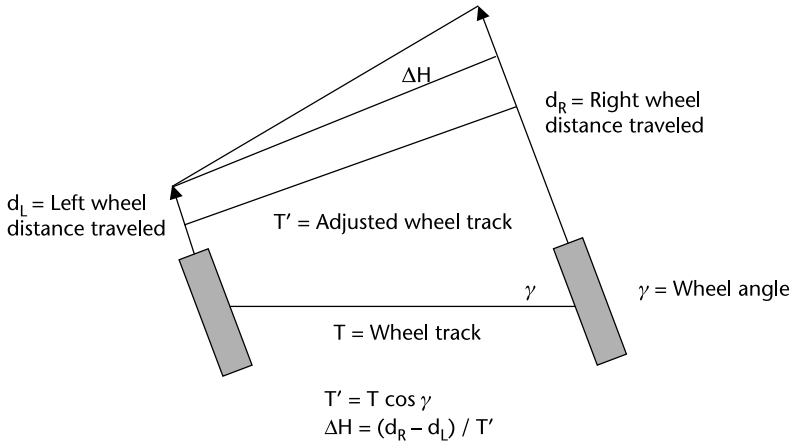


Figure 13.30 Heading change determination using front wheels.

Because of the wheel angles, the original wheel track, T, is no longer perpendicular to the tires, and the effective wheel track, denoted as T' in Figure 13.30, is reduced by $\cos \gamma$. This shortening of the wheel track becomes more significant as the wheel angle increases (i.e., the heading rate is more rapid). Ignoring this effect for front wheel installations can induce significant heading errors in turns. Since the wheel angles are generally not known to the dead reckoning system, an approximate method for computing the effective track width is required. Several such methods are described in [60]. The track width variation with vehicle speed is also referenced in [60].

Heading determination via differential odometry is susceptible to gross errors when the pulse count difference is induced by either tire slipping or skidding, as discussed for the transmission odometer, but also due to significant differential tire pressure. A relatively small difference in tire pressure, if not calibrated, can lead to significant error growth: a difference in tire size of 1% produces a heading rate error of over 2°/s at 20 km/h, assuming a 2-m wheel track. Calibration of this error source by the integration filter is therefore essential and will be discussed in Section 13.3.3. The effects of wheel sensor pulse count quantization are not a significant error contributor to accumulated heading error, as demonstrated [61]. The reason for this is that a quantization error, which can induce a heading error of one distance quantum divided by the wheel track in one sampling interval, will tend to correct itself in the following sampling interval. For example, if the left tire had just missed registering a pulse in the current sum of pulse counts, it will certainly register that pulse in the next pulse count and so catch up in its measure of accumulated heading change. In statistical parlance, successive quantization induced heading errors are strongly negatively correlated; hence, their summation approaches zero. However, ABS will sometimes exhibit a random heading error, whose one-sigma level is roughly the size of the pulse quantization, and so behaves like an uncorrelated, quantization error. The error can be attributed to noise in the sensor that generates the pulses, which seems to be accurate, by design, to the pulse quantum level. Thus, (13.46), although not representative of the effects of true pulse quantization, may still be representative of the actual heading error growth and so is

generally recommended for consideration in the design of any real-time Kalman filter algorithm for conservatism.

$$\sigma_H = \sigma_q \sqrt{t} \tag{13.46}$$

where σ_H = one-sigma heading error (rad); and σ_q = quantization level (rad).

Table 13.2, abstracted from [37], assesses the magnitudes of various factors affecting differential tire size.

13.3.2.5 Barometric Altimeter

As mentioned in Section 13.3.2.3, a barometric altimeter can be used to stabilize an inertial indication of altitude derived by a vertical accelerometer and a gravity model. In addition, it can be used to augment a GNSS-based altitude, as with a gyro/odometer or ABS-based augmentation of GNSS. For land vehicles, all visible satellites will be above the horizon and the resulting geometry yields more uncertainty in the vertical direction. Therefore, GNSS-based altitude estimates are less accurate than horizontal position estimates, so an augmentation for improved altitude estimation is attractive. Relatively low-cost barometric altimeters are available [62], which can sense changes as low as 1m in altitude and, with proper calibration, can provide absolute altitude measurements better than 10m. Both the absolute and relative altitude data are valuable in aiding GNSS, whether inertially augmented or not.

Since any barometric altimeter determines altitude through sensing air pressure, calibration is necessary. The calibration, which associates pressure readings and altitudes, will degrade relatively slowly with time, as local weather conditions change. The calibration accuracy will also degrade as the physical separation between the vehicle and the reference location for the calibration increases. Thus, operation of a vehicle navigation system that makes use of a barometric altimeter as a source of absolute altitude information will require that calibration data from a reference station be supplied to it, or that similar calibration in formation be supplied by GNSS. If the navigation system is integrated with a cellular phone or other communication means within the car, then the communications network can provide the calibration information applicable to the location of the vehicle. Air pressure sensors can be implemented on a small silicon chip at low cost [62] and cellular phone base stations can provide calibration information [63]. Calibration using GNSS is preferably done by inclusion of a barometric altimeter bias state in the

Table 13.2 Factors Affecting Wheel Scale Factor

Error Factor	*Possible Error in Radius*
Pressure	1 mm/lbf/in^2
Temperature	1 mm/5°C
Wear	5 mm
Speed	1 mm
Weight	1 mm/100 kg

Source: [37].

Kalman filter, which compares GNSS altitudes with barometric altimeter readings. An appropriate level of process noise associated with the barometric altimeter bias will ensure that the calibration is not static. In addition, if the barometric altimeter derived pressure changes are used as a source of vertical velocity information by the integration filter, a scale factor error state may be necessary, which calibrates altitude change derived by pressure change using GNSS determined vertical velocity.

13.3.2.6 Magnetic Compass

Magnetic compasses provide an inexpensive means of determining vehicle heading and have been used to augment DR systems [64]. The major problem associated with the use of a magnetic compass as a primary or sole heading reference is its sensitivity to magnetic anomalies such as large metal structures. Although compass designs can be self-calibrating, this calibration serves only to remove the static disturbance of the Earth's magnetic field (e.g., as could be induced by the vehicle itself). The error induced by the tilt of the sensor can also be compensated [37]. Dynamic sources of disturbance, which could be generated by other passing cars or the steel trusses of a bridge, can induce very significant errors in the compass' heading indication. Thus, the compass is usually relegated to a backup role, or as a complement to another system. If integrated with a source of heading rate information (e.g., as could be supplied by a low-cost gyro or through differential odometry), the integration filter residual test can usually be used to screen gross errors induced by magnetic disturbances. Such a test compares the current magnetic compass reading with the current best estimate of heading propagated forward in time using the heading rate information.

13.3.3 Land Vehicle Sensor Integration

13.3.3.1 Position Versus Measurement Domain Integration

Integration of GNSS with any of the systems and sensors discussed in the previous section can generally be done in either the position domain or the measurement domain. Position domain integration means that GNSS positions and velocities are processed by the navigation filter along with data from additional sensors. Measurement domain integration means that individual GNSS satellite pseudorange and Doppler measurements are processed by the navigation filter along with data from additional sensors. Generally speaking, measurement domain processing is preferred, but it is not necessarily required for acceptable performance (see [65] for a description of a gyro-based DR system which uses position domain integration). An advantage of measurement domain integration is that the pseudorange and Doppler measurements can be individually weighted based on measurement uncertainty with the data from additional sensors allowing weak or noisy signals to be further deweighted when other sensors have low uncertainties. The measurement domain approach also enables partial updates of the DR system using less than the number of satellites required for a position fix (i.e., three for a two-dimensional fix or four for a three-dimensional fix), thus performance should be improved. However, this improvement comes at a cost. The cost is the requirement for the integration filter to compute the satellite positions and velocities from the ephemeris data decoded

by the GNSS receiver or to request this from the GNSS receiver. In cases where the integration filter runs in the same processor as the GNSS receiver, the cost is zero.

A common misconception in sensor integration is that a measurement domain integration is needed for sensor calibration. Both integration approaches enable calibration of the various sensors. GNSS heading and speed information (derived from the GNSS determined velocity), as well as individual satellite Doppler measurements can be used to calibrate the gyros, accelerometers, and wheel sensors of the DR system.

13.3.3.2 The Ubiquitous Kalman Filter

The Kalman filter remains the most widely used tool in integrated navigation systems. In this section, the key aspects of Kalman filter designs for three of the integrated systems identified in the previous section will be provided. It is assumed that the reader is familiar with Kalman filters, or can consult one of the many excellent textbooks on the subject [11, 14] as well in Section 13.2.3. The three systems that will be examined in detail include an INS with GNSS, three gyros, and two accelerometers; a system with GNSS, a single gyro and an odometer; and a system with GNSS and differential odometers using an ABS.

Kalman Filter Model for Two-Accelerometer INS

The error equations for an INS are well known and will not be repeated here [8]. Suitable error models for automotive quality sensors should include the basic nine error states associated with the unforced error dynamics of any INS, excepting the two states specific to the vertical axis (i.e., two INS position errors, two INS velocity errors, non-INS altitude and vertical velocity errors, and three attitude errors). The fundamental (**F**) matrix associated with the INS error dynamics has two distinct frequencies of oscillation when the INS is at rest: the Schuler frequency, with an 84-minure period, and Earth rate, with a 24-hour period. Because the longest GNSS outages in the automotive environment are expected to be no more than several minutes long, the Earth rate dynamics can be ignored, and the Schuler dynamics are well approximated by much simpler equations. Now returning to the state vector selection, the basic nine error states (i.e., 3 position errors, 3 velocity errors, and 3 attitude errors) will be augmented by three gyro bias states, two accelerometer biases, three gyro scale factor errors, and two accelerometer scale factors, resulting in a total of 19 states. The resulting state vector is summarized here:

$$\mathbf{x}^T = \left[\delta\mathbf{p}^{\mathrm{T}}\delta\mathbf{v}^{\mathrm{T}}\delta\boldsymbol{\theta}^{\mathrm{T}}\mathbf{b}_{\theta}^{\mathrm{T}}\mathbf{s}_{\theta}^{\mathrm{T}}\mathbf{b}_{\mathrm{a}}^{T}\mathbf{s}_{\mathrm{a}}^{\mathrm{T}}\right] \tag{13.47}$$

The 19 states must be augmented by GNSS clock phase and frequency errors if a measurement-domain integration approach is chosen, resulting in a total of 21 states. Preferably, the modeled position errors ($\delta\boldsymbol{p}$) in (13.47) are represented in meters, velocity errors ($\delta\boldsymbol{v}$) in meters per second, attitude errors ($\delta\boldsymbol{\theta}$) in radians, gyro biases ($\boldsymbol{b}_{\theta}$) in radiand per second, and accelerometer biases ($\boldsymbol{b}_{a}$) in m/s^2. Note that scale factor errors for both the gyro ($\boldsymbol{s}_{\theta}$) and accelerometer ($\boldsymbol{s}_{a}$) are unitless.

Given that most GNSS outages due to signal blockage are less than a few minutes in duration, the sine or cosine of the Schuler angle, which appear in various terms in the INS error dynamics equations, can be well approximated by:

$$\sin(\omega_s t) = \omega_s t \tag{13.48}$$

$$\cos(\omega_s t) = 1 - \omega_s^2 t^2/2 \tag{13.49}$$

Given these substitutions, the INS error dynamics simplify significantly and become more intuitive:

$$d\delta\mathbf{p}/dt = \delta\mathbf{v} \tag{13.50}$$

$$d\delta\mathbf{v}/dt = \mathbf{b}_a + g\delta\boldsymbol{\theta} + \mathbf{S}_a\mathbf{a} \tag{13.51}$$

$$d\delta\boldsymbol{\theta}/dt = \mathbf{b}_\theta + \mathbf{S}_\omega\boldsymbol{\omega} \tag{13.52}$$

where $\mathbf{S}_a$ and $\mathbf{S}_\omega$ are matrices with the scale factor elements on the diagonal, and instrument input axis misalignments as off-diagonal terms, with g representing gravitational acceleration in m/s^2. Our Kalman filter state vector per (13.47) only estimates the accelerometer and gyro scale factor errors (i.e., the misalignments are set to 0 in these equations). A real-time Kalman filter would generally have a very difficult time observing these misalignments, as controlled maneuvers are generally required for observability, so they are generally assumed to be calibrated to negligible levels prior to the filter's operation. The altitude and vertical velocity error behavior is noninertial, yet must be modeled by the filter, since errors in these states drive the inertial errors. A simplified model providing acceptable performance for many applications is:

$$d\delta p_3/dt = \delta v_3 \tag{13.53}$$

$$d\delta v_3/dt = -\beta\delta v_3 + w \tag{13.54}$$

In (13.53) and (13.54), units are consistent with those already referenced, with position errors in meters, and velocity errors in meters per second.

In (13.54), the velocity error is modeled as a Markov process [8], which, through the appropriate choice of the variance associated with the white noise, w, reaches a steady state error variance in the absence of updates. This error variance represents the expected variation in the vertical velocity of the car. Altitude and vertical velocity can be maintained through GNSS measurement processing and can also be augmented with barometric altimeter measurements. In this case, as discussed in Section 13.3.2.5, a barometric altimeter bias state should be added to the state vector, resulting in a total of 22 states.

Implementing a Kalman filter with 21 or 22 states may pose some problems from a computational burden standpoint, depending upon the processing bandwidth available to the filter. Some of the states can perhaps be removed. Leading candidates for removal are the scale factor errors associated with the pitch-and-roll gyros, since pitch-and-roll rates are not expected to be large for car maneuvers, except for relatively high-frequency effects, as could be induced by speed bumps, but which do not integrate to significant attitude error. It may also be worthwhile to consider removing the accelerometer bias states, since the initial determination of vehicle pitch and roll will remove their effect. Their inclusion is therefore largely a function of the bias instability and the expected pitch-and-roll agility of the vehicle.

Because of the potentially significant temperature sensitivities associated with the gyro and accelerometer bias and scale factor errors, it is highly desirable that temperature information be supplied with their high-rate outputs (i.e., the gyro measured delta-angles, and the accelerometer measured delta velocities). The temperature sensitivities can be measured in a laboratory environment (as previously discussed, this must be done for each gyro and accelerometer), and so the resulting bias and scale factor error estimates will be comprised of a precomputed temperature-dependent component, preferably represented as a curve fit, and a correction to that generated by the Kalman filter from processing GNSS. A consistent and statistically significant trend in the correction component away from the sensitivity curve may result in a modification of the temperature sensitivity curve, as could be determined using the statistic below for the gyro bias:

$$S_t = \sum \left(\delta b_\theta / \sigma_{bg} \right) \Big/ n \tag{13.55}$$

In (13.55), δb_θ represents the corrections to a component of the gyro bias vector $\mathbf{b}_{\theta,}$ preferably represented in rad/s, over the most recent set of n Kalman filter updates. The value in the denominator of the summation, σ_{bg} in (13.55) represents the a priori uncertainty associated with each gyro bias component correction, representing the designer's best knowledge about its temporal stability. If the process noise associated with the gyro bias state considered in (13.55) assumes that the factory generated temperature compensation curve is effective in removing the gyro bias sensitivity, then the value of the normalized statistic in (13.55) can be used to detect a departure from those conditions. Such a detection must be gated by two conditions: a significant temperature change occurring over the set of n updates used in (13.55), and the establishment of an upper limit, or threshold for the statistic. Such a threshold selection will typically be chosen to represent a three-sigma condition, dictating use of a value of nine for testing S_t. However, simulation study and test experience will generally be required to achieve the desired response characteristics from the test. When the threshold is exceeded, the precomputed temperature curve for this error source can be revised. Such revisions are generally done cautiously; incorrectly revising the temperature sensitivity curve can adversely affect performance for a long time until the erroneous adjustment is detected and removed. Similar statistics and tests can be generated for each error source for which a predetermined temperature compensation exists.

Low-cost sensors may also exhibit significant scale factor asymmetry (i.e., it may be advisable to separately model gyro and accelerometer scale factors for positive and negative rotations and accelerations, respectively). Usually, however, the component of the scale factor which is common for both directions is dominant, and the asymmetry can barely be observed in the normal operation of the vehicle.

Given the state vector definition in (13.47), the process noise selection should consider all sources of error that have been excluded [i.e., scale factor asymmetry, sensor misalignments, and gyro g-sensitivity (if significant)]. In addition, the expected noise floor of each sensor is also included. Since most of the unmodeled effects behave more like biases than noise, caution must be exercised to select appropriate levels. As is well known, bias errors do not behave like white noise; for example, a bias acceleration error produces a velocity error that grows linearly or an error variance that grows quadratically. However, representing a bias acceleration error as white noise (implied through a process noise representation) produces a velocity error with a variance that grows linearly.

Consider the misalignment of the roll gyro about the lateral axis of the vehicle as an illustrative example. This error source is generally expected to be constant, assuming that the gyro case is rigidly attached to the vehicle, and does not experience significant shock (which could change the sensitive element's alignment within the case). During a heading maneuver, for example, this error source produces an angular velocity error in the roll gyro's output:

$$\delta\varphi = \Delta H \; m_\theta \tag{13.56}$$

where m_θ is the misalignment of the roll gyro about the pitch (or lateral) axis of the vehicle, measured here in degrees, ΔH represents the magnitude of the heading maneuver in radians, and $\delta\varphi$ is the resultant roll error in degrees. If the vehicle makes a U-turn at a stoplight, the heading change will be π radians, and let us assume that the maneuver is completed in 5 seconds. The actual roll error that is induced, assuming a 1° misalignment, will be slightly more than 3 (π) degrees. If we select a process noise variance as in

$$q_\varphi = \Delta H^2 \sigma_m^2 \tag{13.57}$$

where σ_m^2 is the error variance assigned to the misalignment, and ΔH is the sensed heading change of the gyro in each assumed 1-second propagation step, use of the (13.57) representation will increase the roll error variance by less than 2.0 degrees^2 at the end of the turn, or roughly 1.4°, one sigma, compared to the actual error, which is more than double this predicted one-sigma value. The reason for this optimistic prediction is that the filter assumes a white noise model, such that the error accumulation root-sum-squares from second to second: but the actual error is a bias, which adds each second. A way to force the filter to be more conservative, and so more realistic, is to assume a maneuver duration associated with the heading change, and scale the process noise variance by this amount. If a 3-second average maneuver change is assumed, the resulting prediction will be 2.4°, one sigma, closer to the actual induced roll error.

Kalman Filter Model for Gyro/Odometer

Integration of a vertical gyro (to sense heading changes) with an interface to the vehicle's odometer is one of the first GNSS augmentations considered [65], and was one of the most popular options for its relative simplicity and lower cost. A commonly selected state vector for the Kalman filter is given as (13.58) in row vector form:

$$x^T = \left[\delta \mathbf{p}^T \;\; \delta \mathbf{v}_o \;\; \delta v_z \;\; \delta H \;\; b_H \;\; s_H\right] \tag{13.58}$$

where $\delta\mathbf{p}$ is the three-dimensional position error vector, δv_o is a scale factor error associated with the odometer, $\delta \upsilon_z$ is vertical velocity error, δH is heading error, and b_H and s_H are the gyro bias and scale factor error, respectively. In (13.58), position errors are represented in meters, velocity errors are represented in meters per second, heading error in radians, and gyro bias error in radians per second. In general, temperature error curves can be derived and applied for both the gyro bias and scale factor error, if temperature information in the vicinity of the gyro is available. The state vector definition in (13.58) implies a centralized filter approach, that is, where a single filter (8 states) is used; however, adequate performance can be obtained using a decentralized approach [65]. In this system, individual, mostly single-state filters are used.

Appropriate levels of process noise are required to force the filter to track variations in average tire pressure due to changes in temperature and driving conditions that affect the scale factor error associated with the odometer. In addition, if the odometer cannot accurately track very low velocities due to sensor limitations (see discussion in Section 13.3.2.4), additional process noise can be injected into the horizontal position error states directly (the velocity error is therefore represented as the sum of the odometer scale factor induced error plus other, unmodeled effects which are represented as white noise). Any filter designed to operate with sensors that derive velocity information from the vehicle's wheels must deal with the anomalous sensor performance induced by wheel skidding and slipping. As mentioned in Section 13.3.2.4, the preferred solution to tire slipping is to derive information from the nondriven wheels; however, this may not always be possible. For tire skidding, there may be an indication of ABS activity (if the car has an ABS) which can be made available to the filter. This serves as an alert, and conservatism would dictate that when this occurs, additional process noise should be injected to keep the filter aware of potential error in its propagation. The amount should be derived from test experience.

For either skidding or slipping, then, the Kalman filter may have to adjust to a potentially significant and unmodeled source of error. Since its a priori levels of process noise do not reflect the presence of either condition, they must be treated as failure conditions by the filter. Generally, gross discrepancies between the GNSS measurements (in this case, a Doppler or velocity component) and the reference speed and heading may indicate such a failure; however, distinguishing between slipping or skidding and a large Doppler error (as could be induced by tracking a reflected signal or tracking beyond the limits of the lock detector) is not straightforward. Failures are generally detected by a Kalman filter through a statistical test applied to the measurement residuals, as explained in Section 13.2:

$$\text{if } \left(D_{res}^2 > r_{scale} r_{\text{var}}\right) \text{ bypass this Doppler measurement} \tag{13.59}$$

where D_{res} is the Doppler residual for the current satellite represented in meters per second, and r_{var} is the Kalman filter computed residual variance (in $(\text{m/s})^2$). The parameter r_{scale} is typically set to 9, implying that the probability of a residual failing the test (under the assumed unfailed error conditions of a Gaussian process) is roughly 0.01. If the failure condition is the reference trajectory (as would be the case if significant tire slipping or skidding was occurring), then several or perhaps all Doppler measurement residuals should fail. This is therefore a way to distinguish skidding and slipping from Doppler failure, since it is unlikely that several or all Doppler measurements would fail at the same time. In this case, two approaches can limit the errors induced in the integrated trajectory: reinitialization to a GNSS position and velocity (if that is possible, given the GNSS coverage at the time of the failure), or addition of sufficient process noise such that measurement rejections no longer occur. The appropriate level can be determined through experiments conducted with test data, or it may be possible (depending upon the number of Doppler measurements available during the failure condition, to solve for the needed amount of process noise:

$$\mathbf{h}^T \mathbf{\Delta Q h} = r_{\text{var}} - D_{res}^2 \tag{13.60}$$

The vector $\mathbf{h}$ in (13.60) represents the measurement gradient for each measurement which produces a detected failure using the test of (13.59). Since (13.60) is a single equation, each residual that produces a failure detection through (13.59) should be included to enable a possible solution for the process noise increment $\mathbf{\Delta Q}$, which will generally have more than a single nonzero component. An overdetermined set of equations for $\mathbf{\Delta Q}$ may be ensured if we limit the increment to the horizontal velocity components, or further limit the increment to a speed adjustment or a scale factor adjustment to the a priori process noise levels. Once determined, the covariance propagation can be repeated and the Doppler measurements re-processed, if sufficient processor throughput exists.

Kalman Filter Model for ABS

Integration of the sensed wheel speeds, or distances traveled from an ABS in a vehicle is perhaps the most cost effective augmentation of GNSS, since the ABS sensors are already present and no other sensors are procured. A commonly selected state vector for the Kalman filter is given as (13.61) in row vector form:

$$\mathbf{x}^T = \left[\delta \mathbf{p}^T \delta v_L \delta v_R \delta v_z\right] \tag{13.61}$$

where $\delta \mathbf{p}$ is the three-dimensional position error vector, represented in meters, δv_L is a scale factor error associated with the left wheel, δv_R is a scale factor error associated with the right wheel, and δv_z is vertical velocity error in meters per second. It is possible to include information from more than two wheels; however, inclusion of separate scale factors for each wheel can then lead to observability problems for the integration filter. Essentially, the average of the left and right scale factors is es-

timated by comparison with GNSS derived speed, while the difference is determined using GNSS derived heading.

ABS determined speed and heading is also subject to failures induced by slipping and skidding, but in a potentially more damaging way than for the gyro/odometer system. Since heading is also determined from the wheels, the potential exists for very large heading errors to develop, for example, one wheel slipping over ice while the other is stationary produces a heading error rate equal to the wheel speed divided by the track. A slipping rate of 30 km/h corresponds to a heading error rate of almost 300°/h. In general, heading errors are more of a concern in the use of DR systems than speed errors, owing to the potential for excessive error growth as heading errors become large.

Another issue worthy of mention is the possible adjustment of the covariance equations as heading errors become large. Due to the additional failure mechanisms just discussed, heading errors exceeding the expected linear range (e.g., 10°) can and will occur. In these cases, filter conservatism can be lost with a linear model. In developing a linear model involving the sine and cosine of heading, the usual (linear) approximations are:

$$\sin(\delta H) = \delta H \tag{13.62a}$$

$$\cos(\delta H) = 1 \tag{13.62b}$$

In (13.49), the heading error is represented in radians. As heading error becomes large, the cosine function can be better approximated as $1 - \delta H^2/2$. The error variance propagation equations have become nonlinear, since expressions involving error variances associated with the sine and cosine of heading error can no longer be linearized. These expressions can be approximated by including additional terms involving the variance of $\delta H^2/2$. Its variance can be approximated using a Gaussian assumption, and noting that:

$$\text{var}\left(\delta H^2\right) = 3\sigma_{\delta H}^4 \tag{13.63}$$

Thus, the traditionally linear variance propagation equations can be replaced by equations that approximate the nonlinear distortion of the statistics.

Gyro/ABS Performance Comparisons

A comparison of urban canyon performance for experimental gyro and ABS-based dead reckoning systems is performed in [37]. Both are integrated with two types of GNSS receivers: wide and narrow correlator spacing. As discussed in Section 9.5, the receiver with narrow correlator spacing is expected to reduce the effects of multipath on each pseudorange measurement. Many sets of comparison data are generated and discussed in [37]; however, only a small subset of the performance data is summarized here. The reported tests of primary interest are those tests performed in downtown areas, as these are expected to be most limiting for GNSS coverage. Figure 13.31 is a sample result from the gyro/odometry integration, where the map truth and unaided GNSS trajectories are also shown. Corresponding results for the

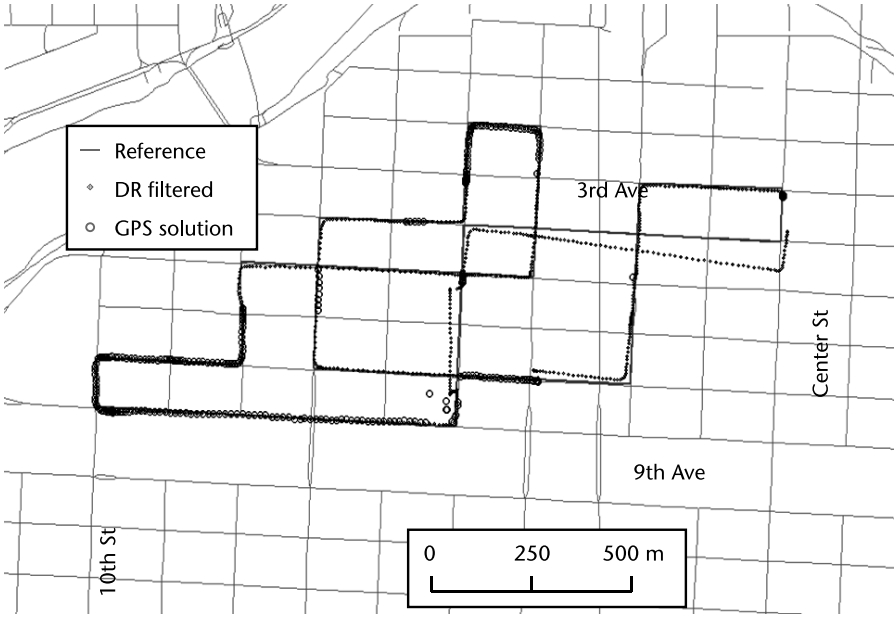


Figure 13.31 GNSS and gyro/odometry integration filter results.

ABS integration are shown in Figure 13.32. Both tests were performed with narrow correlator receivers.

Since it is difficult to make quantitative comparisons from the plots, the following summary table is also abstracted from [37], and provides a rough characterization of the relative performance of the integrated systems. The results represent a summary of roughly a dozen tests, and indicate that the gyro based system has performance advantages, particularly in reducing the maximum excursions from the road.

Both of these DR systems can provide complete solution availability under nominal sensor performance conditions; however, both systems are subject to conditions that can lead to excessive error growth, which inevitably forces a reset to a GNSS solution in order to recover. For the ABS, road conditions can induce such error behavior due to skidding or slippage, while for the gyro-based system, a gyro failure or abrupt and unknown temperature change can induce this behavior. Generally speaking, this is expected to occur more frequently for the ABS. The choice then for the systems designer is whether or not the cost of the gyro is worth the expected reduction in excessive drift conditions.

13.4 A-GNSS: Network Based Acquisition and Location Assistance

Since its first appearance in consumer handsets in the early 2000s and with the explosion of smart-phone applications using location based information, the number of deployed assisted GNSS (A-GNSS) receivers has surpassed all other applications of satellite navigation combined. This section summarizes A-GNSS methods that are enabled by network assistance messaging in modern cellular applications. For readers interested in a more in-depth treatment of the technology, see [66].

Table 13.3 Summary Comparison, Gyro Odometry, and ABS Integrations

Dead Reckoning	*Maximum Error*	*RMS Error*
ABS	115m	17m
Gyro/odometry	69m	13m

Source: [37].

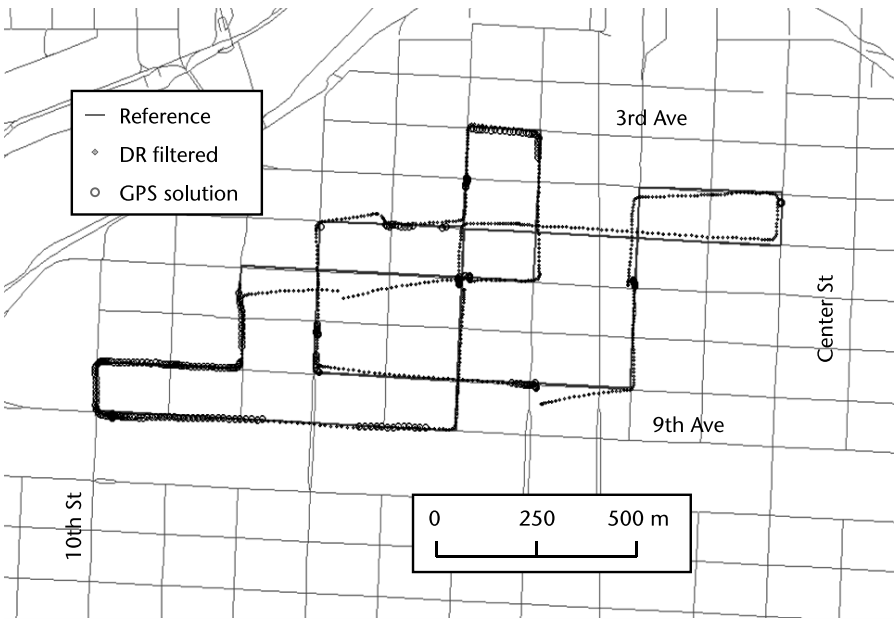


Figure 13.32 GNSS and ABS integration filter results.

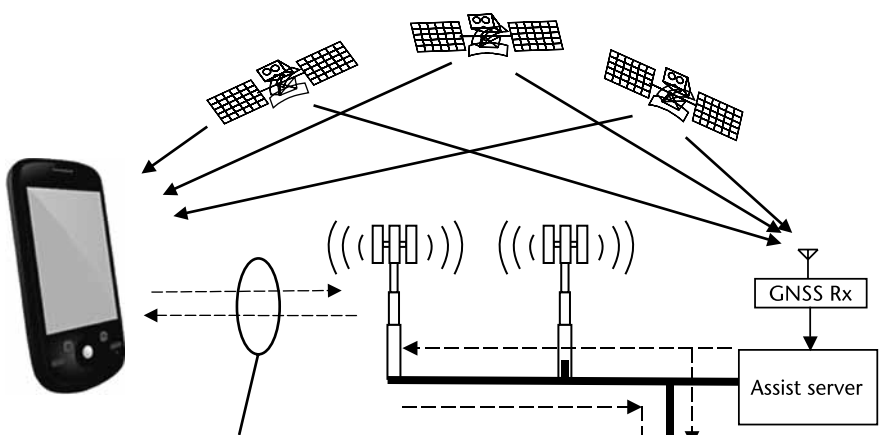
Network Assisted GNSS (or A-GNSS) grew from a need to overcome some of the system shortcomings of all GNSS systems when the technology is used in battery-powered mobile wireless devices to enable applications such as emergency location reporting (E911, E112) or location-based services. Mobile wireless units cannot afford to leave the GNSS receiver on all the time as battery life would suffer. Because wireless mobile devices are used more frequently indoors than outside, the received signals are weak, blocked or attenuated by buildings. In addition, GNSS receive antennas integrated in small handsets are, by nature, small, and small antennas are synonymous with lossy antennas. A-GNSS provides methods of reducing receiver on-time while enabling signal processing gain (i.e., increased sensitivity) to overcome poor antennas and indoor environments.

One drawback of stand-alone GNSS is the long time needed to demodulate the satellite orbit (ephemeris) and satellite clock correction parameters that are essential to computing user position. If a receiver could acquire the satellites instantly, 30 seconds of additional time (example, GPS) is required to demodulate the 50-bits-per-second (bps) navigation data message (NAV) before a position solution can be computed. In applications in which the GNSS receiver is part of an emergency response system, waiting for data demodulation to occur can seem like an eternity.

All existing and future GNSS systems (GPS, GLONASS, Galileo, BieDou) have this same drawback. As such, methods to eliminate the need to demodulate the NAV message and to decrease the signal acquisition time in weak signal environments are the two primary drivers for A-GNSS. A-GNSS takes advantage of the communications link enabled by the embedded wireless modem to exchange data with the network. The data exchange allows the unit to overcome the NAV data decoding time and weak signal obstacles.

There are two basic methods of A-GNSS employed in cellular handsets (Figure 13.33): *MS-assisted* and *MS-based*. In cellular telephone terminology, MS refers to the mobile station (MS) or cellular phone. The two methods are quite different, but both require a complete or nearly complete GNSS receiver to be integrated into the MS and a data exchange with the network enabled by a predefined over-the-air protocol.

The position solution is computed in the network when using the MS-assisted method. The MS-assisted handset shifts some of the functions of the traditional GNSS receiver to a network-based processor or server. This method requires most of the hardware elements of a stand-alone GNSS receiver (an antenna, RF section, and digital processor), but generally can get by with less embedded RAM and read-only memory (ROM) as the firmware required to compute the position solution exists elsewhere in the network. The network transmits a very short assistance message to the MS, consisting of time, visible satellite list, predicted satellite Doppler, and code phases, among other things. This visible satellite list tells the GNSS which



-	MS-based	MS-assisted
Handset inputs	Ephemeris, approx position and time, almanac, DGPS real-time integrity, sensitivity assist	SVIDs, Dopplers, code phases, AZ/EL, sensitivity assist
Handset outputs	Position, velocity, time, quality, error ellipse	Measured code phases, measured Dopplers, C/N_0s
Position computed	In handset	In network

Figure 13.33 Assisted GNSS positioning methods: MS-based and MS-assisted.

satellites to acquire, and the Doppler/code phase data indicates where to look. Acquisition time is reduced because the Doppler and code phase search space is much smaller than in autonomous GNSS processing because the search space is small as predicted by the network. This allows for rapid search and the use of narrower signal search bandwidth that enable enhanced sensitivity by allowing the receiver to dwell longer in each of the reduced Doppler/code phase search bins.

The how and why are explained in detail in Section 13.4.3. The MS-assisted handset acquires the signals and returns the measured pseudorange data for all detected satellites to the network. There, a position determining entity (PDE) such as a server does the work of computing the position solution. MS-assisted solutions are inherently differential in nature, since the PDE has access to DGPS corrections, either from a local receiver or via the Internet.

In the MS-based method, the position solution is computed in the handset. The MS-based solution maintains a fully functional GNSS receiver in the handset. This requires the same functionality as described in MS-assisted handset with the additional means for computing the position of the mobile station. Computing position locally to the handset generally adds to the handset's total memory (RAM, ROM) requirements in addition to increasing the loading on the host processor [e.g., as might be measured in millions of instructions per second (MIPS)]. The MS-based handset may work in an autonomous mode as well, providing position solutions to the user or embedded applications without cellular network provided aiding data.

MS-based methods are better for applications requiring the position solution in the handset, an example of which is personal navigation that can provide the user with turn-by-turn real-time directions (think Google Maps on a smart phone). Turn-by-turn navigation is awkward in the MS-assisted mode because each update of the position solution in the network must be communicated back to the mobile. In the MS-based case, significantly more data needs to be delivered to the handset in the form of the precise satellite orbital elements (ephemeris), but once it is transferred to the handset, little or no additional data is needed to perform periodic fixes as long as the ephemeris remains valid (several hours). MS-based solutions can be differentially corrected by sending corrections to the handset.

The reduction in search space allows the receiver to spend its search time focusing on where the signal is expected to be, which, in turn, allows it to search at a much narrower bandwidth, increasing signal detection sensitivity.

As referenced in this section, network assistance can refer to any one of three forms:

- Acquisition assistance, intended to reduce the receiver's time to generate a fix [time to first fix (TTFF)];
- Sensitivity assistance, intended to help the receiver to lower its acquisition thresholds;
- Navigation assistance, intended to improve the accuracy or integrity of the position solution generated by the receiver.

Certain types of information can qualify as more than a single type of assistance: for example, supplying the GNSS receiver with an initial, coarse position estimate can assist both acquisition and navigation.

13.4.1 History of Assisted GNSS

Many believe that the 2008 U.S. Federal Communications Commission (FCC) mandate to require locating cellular-telephone-based 911 calls [67] triggered the creation of A-GPS technology. The mandate did not create A-GPS, but it certainly made it mainstream. Indeed, most mobile wireless handsets today have embedded A-GNSS technology that is used not only for E911, but also to enable a huge number of innovative location-based applications (apps) that consumers use every day on their smart phones.

Examples of the earliest uses of network assistance predate the introduction of cellular telephones. Perhaps the earliest formal reference to the use of assistance information is disclosed in [68]. NASA inventors realized the potential benefit of transmitting an initial almanac or ephemeris to a mobile GPS receiver to enable prediction of satellite visibility and Doppler and eliminate the long data demodulation time inherent in collecting the required data bits from the signals directly.

The first standard for sending ephemeris data over a wireless link was included as part of the RTCM DGPS Standard [69]: message type 17 includes the ephemeris data for all satellites visible to the DGPS reference station receiver. One of the earliest references to sending measured pseudorange data over a wireless link to support location determination external to the GPS receiver is described in [70]. It describes a vehicle tracking system in which pseudoranges can be sent over a wireless link to a workstation that calculates the position of the host vehicle. In determining the vehicle's position, altitude information derived from a terrain map of the local area can be used to improve the accuracy and reliability of the solution.

An early example of ephemeris-aiding was used in 1985 by the Motorola Eagle receiver. It was one of the first commercially available GPS receivers to offer a form of ephemeris aiding [71] in one of its operational modes. Inherent in its design, and when two receivers were used in a differential master/slave configuration, the master station sent DGPS range and range rate data for all satellites tracked by the master station. In addition to this information, the master station transmitted the ephemeris data for all satellites tracked using a commutated message structure. A few parameters of each ephemeris were sent with each DGPS correction message, allowing eventual broadcast of all ephemeris data for all visible satellites to the slave receiver. The master-station ephemeris information was used by the DGPS slave receiver to:

- Enable the best DGPS position performance by ensuring both master and slave units were using the same ephemeris set for each satellite (this was prior to the development of the RTCM-104 DGPS messaging standard discussed in Chapter 12).
- Ensure that the slave unit acquired the ephemeris data for all satellites visible by the master, maximizing the availability of DGPS solutions.

The latter was especially useful when the slave unit was partially blocked from acquiring the data directly from the satellite because of blocked or reduced signal power to one or more satellites, which occurs near tall mountains, in canyons, under trees, or near buildings. Many times in these environments, the signal is strong enough to detect code phases and track but not strong enough to reliably

demodulate the ephemeris data. Transmitting the ephemeris data from the master to the slave unit alleviated this problem.

In 1990, another system [72] transmitted almanac data via an over-the-air message from a master station to many slave units called pseudorangers. The pseudorangers accepted the almanac data, and used it to acquire and track GPS satellites, then transmit back the measured pseudoranges and a time stamp. The master station, remote from the movers, computed position of each mover from the pseudoranger unit–measured pseudoranges. This idea is a precursor of the MS-assisted method of A-GPS employed in mobile phones today.

Approximate position, ephemeris, almanac, and approximate time-assist information was present in a White Sands Missile Range system [73]. The White Sands system used GPS to measure the performance of missiles. When a missile is fired, it has little time to acquire and track GPS satellites and cannot tolerate the 30-second ephemeris acquisition period. A wireless message was sent from a master station to the just-launched missile consisting of approximate position, approximate time, almanac data, and ephemeris data, all of which was used by the missile to acquire GPS signals rapidly and produce position reports while in flight.

13.4.2 Emergency Response System Requirements and Guidelines

In the United States, the original FCC mandate in 2008 allowed for two types of solutions: network-based solutions that work with all legacy (non-GNSS) phones, and handset-based solutions (such as A-GNSS, E-OTD, AFLT) in which the handset must include the location technology (hardware, software, or both) in its design. Location determination for legacy phones must be performed within the cellular infrastructure based on triangulation of time-of-arrival measurements of the handset signal from multiple base stations. For network-based solutions, the accuracy requirement is 100-m 67% of the time and 300-m for 95% of calls. For handset-based solutions, the corresponding accuracies are 50-m and 100-m, respectively. In all cases, a 30-second maximum response time (TTFF) is suggested.

The legislative rulings on accuracy and availability have evolved a number of times, the most recent ruling from 2014 [74] proposes to eliminate the handset-based/network-based technology division, and proposes to add indoor horizontal and vertical location accuracy requirements. Specifically, the proposed rule states the following for horizontal and vertical indoor location and availability requirements:

- Horizontal location (x- and y-axis) information within 50m of the caller for 67% of 911 calls placed from indoor environments within 2 years of the effective date of adoption of rules, and for 80% of indoor calls within 5 years;
- Vertical location (z-axis) information within 3m of the caller for 67% of indoor 911 calls within 3 years of the adoption of rules, and for 80% of calls within 5 years.

Clearly the inclusion of indoor location accuracy requirements are driven by the adoption of mobile phones (everyone has one), and a steady increase of 911

calls originating from cell phones/smart phones [74], this reference reported the following 911 calling statistics over time:

> In January 2011, Consumer Reports reported that 60 percent of 911 calls were placed through wireless phones. More recently, the California Office of Emergency Services indicates that the percentage of 911 calls that came from wireless devices increased from 55.8 percent in 2007 to 72.7 percent as of June 2013. Furthermore, an increasing percentage of wireless calls are placed from indoors. A 2011 study showed that an average of 56 percent of wireless calls were made from indoors, up from 40 percent in 2003. That number is even higher for smartphone users, who represent the majority of wireless phone owners, as 80 percent of smartphone usage occurs inside buildings.

The proposed vertical requirement provides the emergency responder with actionable location information for all 911 calls; it solves the problem of emergency calls that originate inside high-rise buildings. Life-threatening delays have occurred when emergency response personnel arrive on scene; if the dispatched address is a 50-story high-rise building, they presently have no way to know from which floor the call originated. The 3-m z-axis requirement addresses that need. Indeed, there are multiple technologies being evaluated [75] that enable the determination of accurate indoor location including altitude such as micro-cell base stations, Wi-Fi, Bluetooth Beacon, LEO satellite signals, RF pattern matching techniques, and barometric pressure altimeter data [63] (in the phone), some of which will likely evolve to be integrated and combined with A-GNSS so that location can be obtained from the totality of available resource(s). A majority of researchers agree that A-GNSS by itself cannot meet the indoor location accuracy (x, y, and z) and availability requirements, even though there has been a steady increase in performance. A-GNSS performance will continue to increase as GLONASS, Galileo, and BeiDou constellations are included in the solution.

There are similar mandates in Russia and Japan for emergency calling location, while Europe is struggling to define the requirements for E112 beyond its current network-based capability, such as perhaps requiring assisted Galileo in all handsets as opposed to an all-constellation encompassing A-GNSS mandate [76]. Indeed, most European phones already include A-GNSS that is used for location services but not E112 calling [77].

The 50-m accuracy requirement appears relatively easy to meet when using A-GNSS for outdoors and in moderately challenging environments. However, the location determination must ideally be performed wherever a cellular phone emergency call can be made, including indoors. A-GNSS does a good job in most of the use-case environments except for deep indoors, underground, or in high-rise buildings. It is useful to understand how different environments affect the GNSS signal in order to squeeze out all of the availability that can be obtained. This is the subject of the next two sections.

13.4.2.1 Characterization of Environments

A characterization of L-band signal environments was previously reported in [78–83], which summarize data collection campaigns at 1,600 MHz in support of

satellite telephone communications link margin studies. The proximity of the test frequency to the 1,575-MHz GNSS frequency makes this research applicable. In addition, in-building cumulative distribution function (CDF) fade data from GPS field trials was presented in [84, 85]. In all cases, extensive radio propagation data was collected at L-band and analyzed to characterize the shadowing, scattering, and blocking effects of trees, cars, and buildings. Hundreds of hours of test data were collected and analyzed. Table 13.4 lists the environments characterized in the previously mentioned references and summarizes the 50% median fade of the signal due to the environment. The data in the table was extracted from charts showing fade depth versus probability charts presented in the multiple references listed.

Heavy urban with the portable unit and the three in-building environments was chosen for the basis of further calculations as the median attenuation values were large and expected to produce reduced GNSS satellite signal availability. Mobile and in-vehicle data in an open environment were also chosen to show the trivial case where the received signal strength is so high that fix percentage will surely be 100% and the important case of a unit employed inside of a car.

It should be noted that due to the requirement for reasonable transmitter efficiency portable antennas used for the data collection experiments are fairly large

Table 13.4 Environments Characterized for L-Band Signal Transmission

Environment	*Description*	*Median Signal Attenuations in Decibels (Mobile/Portable/ In-Vehicle*)*
Open	Almost no trees or buildings	2.5/0.0/12.0
Rural light	Moderate to large number of trees, very few buildings	3.0/3.5/12.0
Rural moderate	Moderate to large number of trees, very few buildings	8.0/7.0/16.0
Rural heavy	Light to moderate forested area	16.0/10.0/18.0
Suburban light	Scattered trees and building structures (e.g., homes far from mobile receiver or new residential areas with little vegetation)	2.0/1.5/14.0
Suburban moderate	Suburban area with 1 and 2-story homes with moderate amount of trees	3.5/6.5/13.5
Suburban heavy	Older suburban areas with large numbers of trees and homes close to roads (e.g., older subdivisions in a city like Chicago)	7.0/2.5/11.0
Urban light	Small, sparse urban areas (e.g., urban areas of smaller cities)	2.0/2.0/16.0
Urban moderate	Urban areas from moderate sized cities (e.g., Phoenix)	4.0/4.0/15.5
Urban heavy	Steel canyons (e.g., downtown Chicago)	5.0/15.0/16.0
In-building residential	Buildings made of wood or stucco (e.g., Phoenix and California residences)	12.5**
In-building commercial	1 to 3-story motels, airports, and commercial buildings	24.0**
In-building high-rise	High-rise buildings	30.0**

*The numbers in the column correspond to decibels of attenuation for the indicated conditions. The mobile case corresponds to the reception conditions in an automobile with an antenna installed on its roof, while, for the in-vehicle case, the antenna is used inside the car. The portable case corresponds to an antenna from a transportable satellite receiver with a large quad helix antenna, not typical of GPS antennas embedded into cell phones. **These numbers correspond to the portable case.

and mounted to minimize head blockage. The result is some attenuation numbers that are similar to mobile attenuations. Due to the size of these antennas, they are not considered acceptable for the GNSS needs of a cellular handset. Appropriately sized antennas for handsets have significantly less gain than any antennas used in the collection of the data discussed from [84, 85] or, for that matter, than conventional antennas for mobile or automotive applications. As cellular telephones continue to shrink, the problem of integrating adequate performing GNSS antennas becomes even more difficult. To perform well, the antenna needs to present uniform gain in the up direction covering the full hemisphere where GNSS signals emanate. Simple patch antennas are used in automotive applications and can be hidden under the dash or under the rear deck with little effort and provide the ideal RHCP to match the satellite transmitted signal. However, placement of a dedicated antenna in a cell phone forces compromises in performance with regard to antenna efficiency and gain pattern, especially when the user can hold it in many different orientations (handheld next to head, in the dialing position, and using different hand grips). Antenna efficiencies in the 30% to 40% range are typical, with attenuation profiles in the 5- to 15-dB range, dependent on orientation and use pattern.

Figure 13.34 shows a photograph an embedded inverted-L GNSS antenna in a cellular handset—in this case, an iPhone-6. This is a dual-use antenna that performs both GNSS receive and Wi-Fi Tx/Rx functions and illustrates the packaging challenges of modern smart phones.

13.4.2.2 Characterizing Signal Attenuations

This section presents the results of a measurement campaign that was conducted to statistically characterize L-band and signal attenuation in various environments. Preprocessing of the raw, measured signal amplitude data was corrected to remove the effect of the transmit antenna pattern as its angle to the receiver changed. Measurements in high-rise buildings included a reference receiver on the roof. The resulting fade data was a differential measurement from the two receivers. The preprocessed data output that is of interest here is fade magnitude versus time. Traces were typically 4 to 8 minutes long and can be interpreted as signal attenuation versus time relative to an unattenuated outdoor received signal. Plots of two such

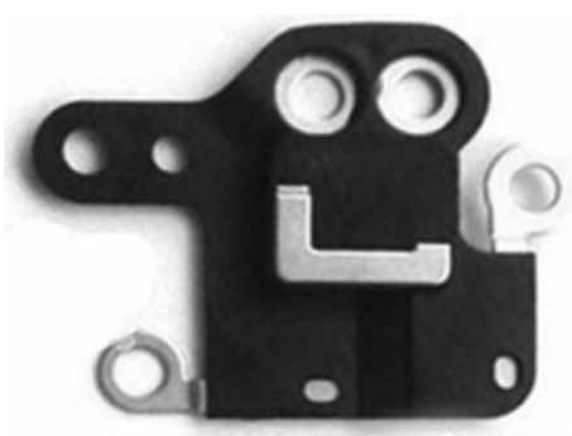


Figure 13.34 Combined GNSS/Wi-Fi antenna used in the Apple iPhone-6.

traces are shown in Figure 13.35. Note the relatively high frequency variation in the traces, corresponding to vehicle motion through its signal environment.

Two CDFs corresponding to the traces in Figure 13.35 are shown in Figure 13.36. The mobile curve on the left has an extremely steep slope with very few fades exceeding 8 dB. The in-vehicle curve on the right has a gentle slope characteristic of a greater standard deviation of the fade value.

An alternate way to look at the signal attenuation profile is to directly use GPS signals detected by a high-sensitivity receiver. In most cases, 8–12 satellite signals are available for measurement at any one time. In order to profile the signal attenuation characteristics in a particular environment, 12 to 24 hours of data needs to be mapped out to capture the effects of the GPS satellite constellation repeat time. The GPS receiver-reported SNR for each satellite (typically in units of dB-Hz) is collected and then translated to an equivalent signal power on the antenna in units of dBm by making an estimate of the receiver noise figure and equivalent bandwidth. If desired, one can map the attenuation profile of the environment as a function of satellite azimuth and elevation angle to provide even more detail on the environment and identify directions of low and high attenuation. Signal power CDF curves are produced and used to predict the availability of location fix within the environment by determining the probability of at least four satellites offering signal power above the receiver's raw detection threshold.

Figure 13.37 shows the signal power CDF curves for the eight strongest satellites detected from a rooftop antenna over a 12-hour period. As can be seen, the 95% probability for the fourth satellite is stronger than about –132 dBm in this open-sky condition. As expected, the signal power spread from the strongest to weakest is only a few decibels.

By contrast, Figure 13.38 shows the same corresponding CDF curves for the strongest eight satellites observed in-building. The environment is the second floor

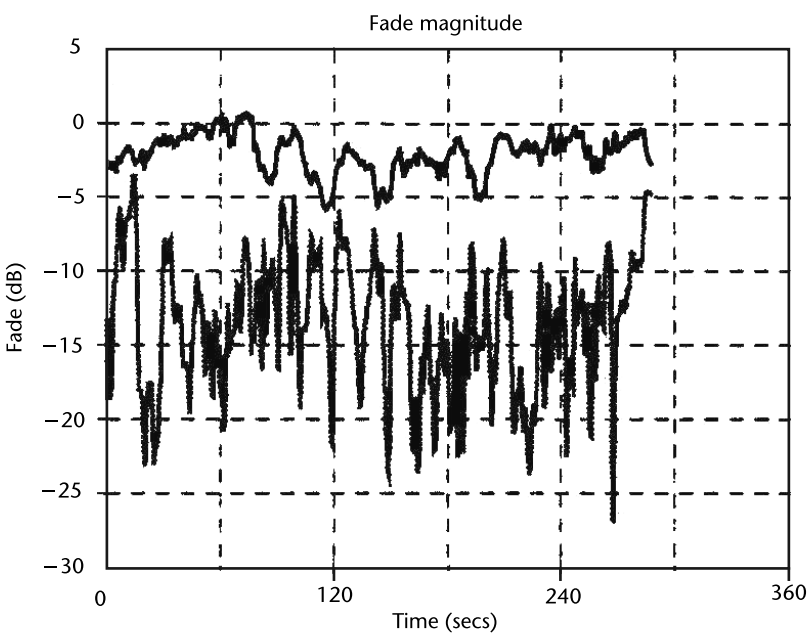


Figure 13.35 Typical fade magnitude versus time for mobile (top) and in-vehicle (bottom).

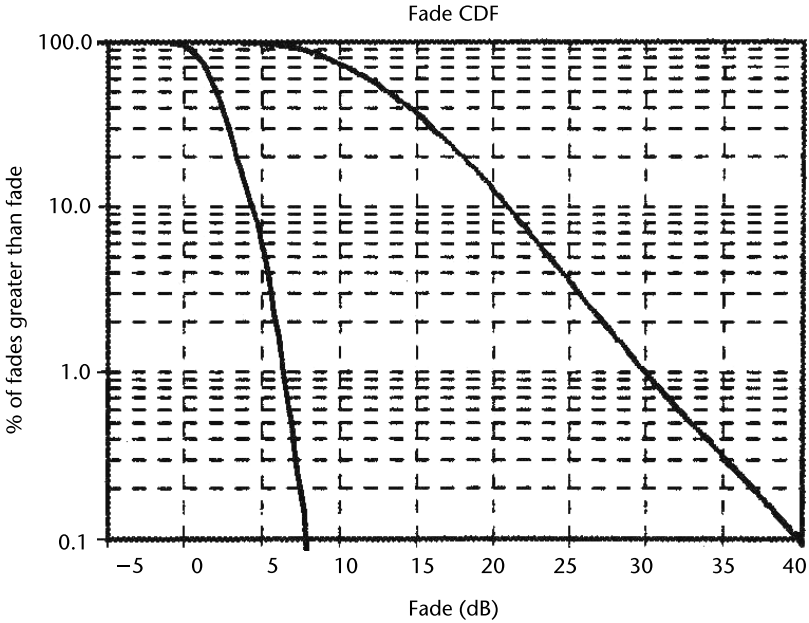


Figure 13.36 Fade CDFs for mobile (left) and in-vehicle (right).

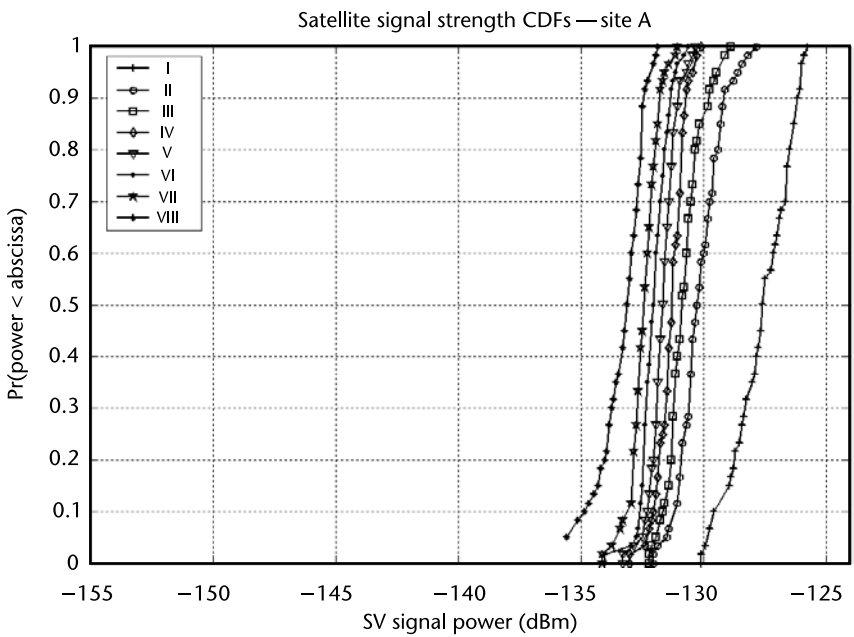


Figure 13.37 Open-sky CDFs for the strongest eight GPS satellites over 12 hours.

of a 3-story apartment building, in the center of the main living room away from windows, with wood and brick construction. In this environment, the 95% probability for the fourth strongest satellite is at >= –152 dBm. Other notable items include that the strongest signal is approximately 10 dB below that from the roof antenna (50% point), and the spread from the strongest to the weakest is much larger, on the order of 20 dB or more in this particular environment. The large

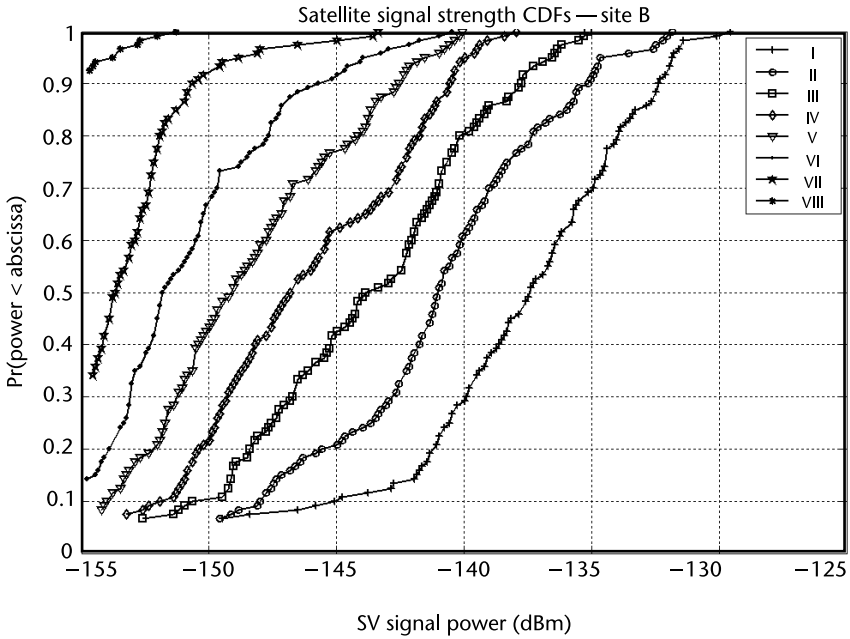


Figure 13.38 In-building (moderate indoor) CDFs for the strongest eight GPS satellites over 12 hours.

spread implies that the algorithm to detect indoor signals should be adaptive as the integration dwell time to detect the stronger signals can be shorter than the dwell time to detect the weaker signals. As will be shown later, common-mode error parameters associated with each satellite signal (code phase error due to time error and Doppler error due to oscillator error) can be exploited to reduce the total search space after one or more satellites are detected; thus, the already detected stronger signals can be used to further reduce the search space in order to detect the weaker signals.

For every environment tested in this way, a unique set of CDF curves will be produced. Thus, it is very difficult to project success or failure in a particular environment based on Table 13.4 or Figure 13.35 through Figure 13.38 without first collecting data and generating CDF curves within the environment in question. The CDF curves include only GPS constellation data, if we project to the GNSS end-state and include a full constellation of Galileo, GLONASS, and BeiDou satellites in the solution, it is clear that the availability of location data from challenged environments improves significantly.

The data in the table and figures should only be used as an example of the specific location tested and should not be used to project other environments, although the trends shown are useful.

With regard to compliance testing, the FCC issued guidelines for testing handsets for compliance to the E911 location mandate [86]. A consortium of CDMA cell phone vendors and suppliers, together with representatives from cell phone carriers, defined a set of minimum performance tests [87] that must be met by A-GPS enabled CDMA phones. These tests define specific signal simulation scenarios and requirements for both position accuracy and TTFF. For example, the sensitivity test of the IS-916 specification requires that the GPS function embedded in a

CDMA telephone acquire four GPS satellite signals at –147 dBm within 16 seconds with success rate of 95% or better. Meeting the minimum performance standard should not be confused with meeting the Phase II location accuracy and availability requirements.

The current Phase II location accuracy rules contain no requirement for testing compliance with the standards or for reporting the results thereof, but it's clear that cellular providers are continuously extracting useful information from their networks as a natural consequence of operating the network. As of 2013, several cellular providers reported 911 location yields ranging from 91% to 95%, which includes emergency calls from indoor locations [74].

With respect to the proposed indoor location requirements, the FCC is considering third-party test-bed compliance testing that would determine actual performance levels of solutions in various real-world conditions and those representative of indoor environments across the country.

Acquiring and using GNSS signals in challenged environments combined with poor performing (small) antennas; and using those measurements for computation of accurate location has been the focus of the research of A-GNSS engineers for the last 15 years. A-GNSS engineers employ three methods of overcoming the losses and meet requirements.

1. Signal processing techniques: Driving the signal acquisition threshold to ever lower levels has increased the GNSS location yield well beyond the capabilities that the original developers of GPS ever imagined. For example, Section 8.5 describes the signal processing gain obtainable by extended coherent and noncoherent integration times, this technique is used by A-GNSS receivers to overcome the signal losses from poor environments and small antennas.
2. Advancing semiconductor technology: By allowing increasing numbers of correlators to be included on-chip and using those to attack the two-dimensional search space (Doppler and code phase) for each satellite of interest, the receiver can find signals faster using longer integration times (i.e., higher sensitivity).
3. Acquisition assistance: Innovative methods of aiding and assisting the receiver that minimize the Doppler and code phase search space (number of search bins) and minimize the receiver on-time (i.e., lower power consumption).

13.4.3 The Impact of Assistance Data on Acquisition Time

The discussion that follows is specific to GPS, but the same methodology is applicable to all GNSS signals with small modifications. The use of the assistance information enables lowering of the number of satellite Doppler and code phase search bins to acquire signals for fix. If a sufficient number of correlators are not available to cover the total uncertainty space in parallel, then some form of sequential processing is required such as searching for each satellite sequentially.

For a given scenario, one can compute the total number of Doppler-code phase search bins needed to be searched, and consequently, the number of correlators

required to cover the entire search space in parallel. The initial parameters of position, time, and frequency uncertainty, along with the particular orientation of the satellite constellation at the time, can be used to compute the total uncertainty search space. Figure 13.39 depicts the two-dimensional search space for a single satellite, the x-axis representing the total Doppler uncertainty and the y-axis showing the total code phase uncertainty. For each satellite, the number of Doppler search bins (N_{dopp}) and code phase search bins (N_{cp}) is computed.

The number of required correlators N_c to cover the search space in parallel is given by:

$$N_c = \sum_{i=1}^{M} N_{dopp_i} \times N_{cp_i} \tag{13.64}$$

where M is the number of visible satellites. For each satellite M, the number of Doppler search bins N_{dopp} is dependent on the total Doppler uncertainty (in hertz), and the coherent integration predetection integration (PDI) period in seconds, which is the same as the PDI period T discussed in Section 8.4.

$$N_{dopp_i} = \frac{\sigma_{dopp_i}}{\left(k/PDI\right)} \tag{13.65}$$

The parameter k is based on the desired overlap of the Doppler search bins and can generally range between 0.5 and 1. The computation of total Doppler uncertainty $\sigma^2_{dopp_i}$ for each satellite is then dependent on the contributions of Doppler uncertainty due to position uncertainty, time uncertainty, reference oscillator uncertainty, and user motion (velocity) uncertainty. Thus, one can write a simple equation for the total Doppler uncertainty per satellite as:

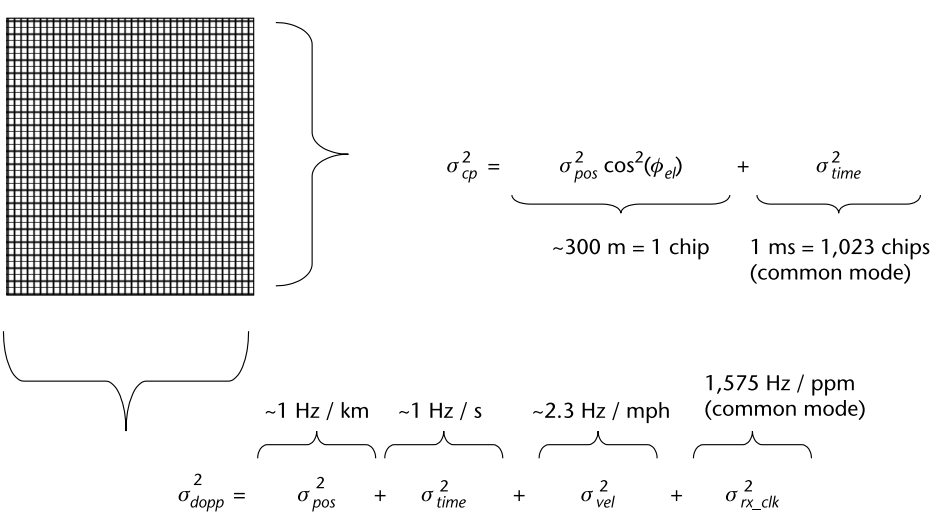


Figure 13.39 Two-dimensional Doppler/code phase search space.

$$\sigma^2_{dopp_i} = \sigma^2_{dopp\_time_i} + \sigma^2_{dopp\_pos_i} + \sigma^2_{dopp\_vel_i} + \sigma^2_{dopp\_oscl} \tag{13.66}$$

The first term is the sensitivity of Doppler uncertainty to time uncertainty, can be computed for each satellite [88], but as a rule of thumb, it is no larger than 1 Hz per second of time uncertainty. Likewise, the second term, the sensitivity of Doppler to initial position error, is about 1 Hz/km in the worst case. In [88], equations are presented for the precise computation of $\sigma^2_{dopp\_pos_i}$ for each individual satellite, which is generally much less than 1 Hz/km. The effects of user platform motion are accounted for by the third term, which represents the Doppler error induced on the GPS signal due to user motion. The term $\sigma^2_{dopp\_vel_i}$ is a maximum for low elevation angle satellites if the user is heading directly at or away from the satellite, and is very small for high elevation angle satellites. In the worst case, $\sigma^2_{dopp\_vel_i}$ contributes no more than 2.3 Hz/mph of user motion, and can be generally multiplied by the cosine of the elevation angle to limit its effect.

$$\sigma_{dopp\_vel_i} \sim 2.3 \times \cos(\phi_{\text{el}}) \text{ Hz/mph} \tag{13.67}$$

The first three terms of (13.66) are dependent on the satellite constellation, the user position, and the user motion. The last term, $\sigma^2_{dopp\_oscl}$ (note no index *i*) is dependent on the reference oscillator and is common-mode with respect to all satellites. $\sigma^2_{dopp\_oscl}$ is typically 1,575 Hz/ppm of reference oscillator frequency uncertainty and is by far the most dominant element of (13.42).

To compute the total code phase dimension uncertainty [see (13.40)], the two dominant terms are proportional to the position uncertainty and time uncertainty. Thus,

$$\sigma^2_{cp} = 4\sigma^2_{pos} \cos^2(\phi_{el}) + \sigma^2_{cp\_time} \tag{13.68}$$

where the term σ_{cp_time} is in units of half-chips by multiplying the time uncertainty (in seconds) by the conversion 2,046 half-chips per millisecond and the first term of (13.68) is computed as shown in Figure 13.40. Figure 13.40 shows the simple relationship of the effect of position uncertainty and satellite elevation angle to transform to the dimension of code phase uncertainty in units of half-chips in the direction of the satellite LOS vector (conversion factor: 1 half-chip = 150m).

The other element of code phase uncertainty is common mode across all satellites and directly proportional to time uncertainty. A 1-ms error in time transforms into a 2,046 half-chip error in code phase. For the typical assisted case in which the approximate position uncertainty is relatively small (e.g., 6 km) as delivered from the network, the largest term of (13.66) is that contributed by the time error.

For the time dimension, we first recognize that the GPS signals are all synchronized in time, which means that, except for the relative drift between the satellite clocks, the first PRN bit and the first navigation data message bit (Subframe 1) leave each satellite at precisely the same time. Each PRN bit and each navigation data message bit is then predictable in time in the following manner:

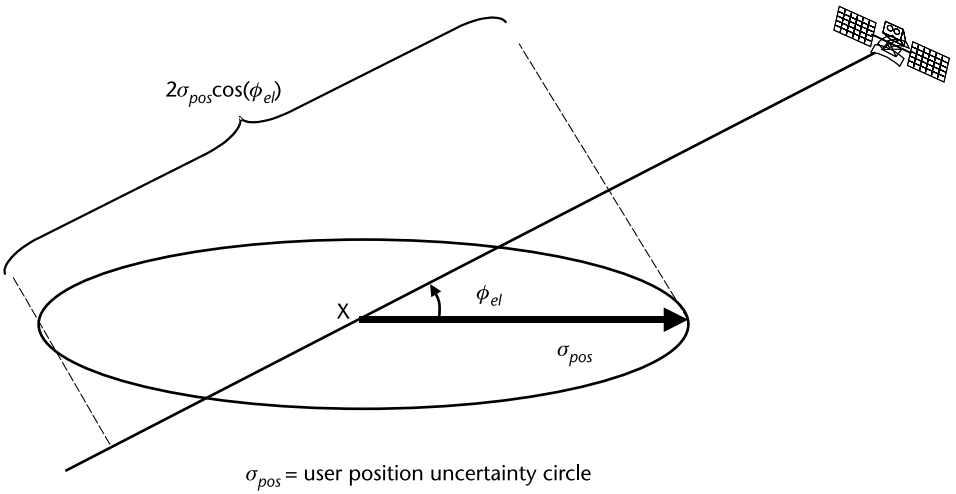


Figure 13.40 Relationship of position uncertainty to code phase uncertainty.

$$\begin{aligned}
&\text{Subframe_Number} = 1 + \text{MOD}(\text{GPS_Time}/6, 5);\\
&\text{Word_Number} = 1 + \text{MOD}(\text{GPS_Time}/(30 \times 0.020), 10);\\
&\text{Bit_Number} = 1 + \text{MOD}(\text{GPS_Time}, 0.020);\\
&\text{Integer_PN_Rolls} = \text{MOD}(\text{GPS_Time}/0.001);\\
&\text{Code_Phase} = \text{MOD}(2046 \times \text{GPS_Time} \times 1000, 2046);
\end{aligned} \tag{13.69}$$

The equations shown in (13.69) enable the user to precisely compute the code-phase, bit_phase, bit_number, word_number, and subframe number of the signal leaving the satellite at any time into the week based on a GPS_Time. The user will observe this code phase on the ground later in time after it propagates from the satellite to the user by the propagation time, dtprop. Dtprop is easily computed as shown in (13.70) based on the geometric range between satellite and user divided by the speed of light. In addition, the signal will be slipped forward or backward in time an amount proportional to the satellite clock correction Tcorr, its range of correction is +/– a few milliseconds. The user will always be able to predict the code phase to all satellites observed on the ground at any instantaneous GPS_Time by the pseudo-code:

$$\text{Code\_Phase\_observed} = \text{MOD}(2046 \times (\text{GPS\_Time} + \text{Tcorr} - \text{Dtprop}) \times 1000, 2046); \tag{13.70}$$

where:

$$\begin{aligned}
&\text{Tcorr} = \text{af0} + \text{af1} \times (\text{GPS_Time} - \text{toc}) + \text{af2} \times (\text{GPS_Time} - \text{TOC})^2;\\
&\text{Dtprop} = |\text{SV_POS-USER_POS}|/\text{Speed_of_light}
\end{aligned}$$

af0, af1, af2 are the zeroth through second-order satellite clock correction terms from navigation data message subframe 1, and T_{OC} is the reference GPS time for the satellite clock correction terms.

Likewise, all equations of (13.69) can be modified to include the user-observed parameters as follows:

$$\begin{aligned}
&\text{Subframe_Number_observed} = 1 + \text{MOD}\big((\text{GPS_Time} + \text{Tcorr} - \text{Dtprop})/6,\ 5\big);\\
&\text{Word_Number_observed} = 1 + \text{MOD}\big((\text{GPS_Time} + \text{Tcorr} - \text{Dtprop})/(30 \times 0.020), 10\big);\\
&\text{Bit_Number_observed} = 1 + \text{MOD}\big((\text{GPS_Time} + \text{Tcorr} - \text{Dtprop}), 0.020\big);\\
&\text{Integer_PN_Rolls_observed} = \text{MOD}\big((\text{GPS_Time} + \text{Tcorr} - \text{Dtprop})/0.001\big);\\
&\text{Code_Phase_observed} = \text{MOD}\big(2046 \times (\text{GPS_Time} + \text{Tcorr} - \text{Dtprop}) \times 1000,\ 2046\big)
\end{aligned} \tag{13.71}$$

In which the "_observed" described by (13.71) represent what a ground-based user located at USER_POSITION on the Earth would observe at the time instant "GPS_Time." Equations (13.71) ignore the small effects of Earth rotation rate as well as tropospheric and ionospheric delay on the signal. However, from a macro level, (13.71) are useful in determining the most likely initial receiver state for the signal detection function to initialize the starting code phase and bit-phase. The computed code phase uncertainty from (13.68) then defines the range of code phase to search over, specifically

$$\text{Code\_Phase\_Search\_range} = \text{Code\_Phase\_Observed} \pm \sigma_{pos} \cos\left(\phi_{el}\right) \tag{13.72}$$

Equation (13.64) can be used to determine the search time for a particular scenario based on signal power, number of satellites, and the required PDI and noncoherent integration dwell time required to positively detect the signal. Section 8.5 previously showed that as signals get weak, the integration dwell time required to positively detect the signal increases substantially. For example, if the signal is –130 dBm (typical clear view of the sky conditions), the signal can be positively detected with a 1-ms PDI and a 2-ms noncoherent integration dwell time. As the signal gets weaker, the required PDI and noncoherent integration time increases substantially (e.g., if the signal is –150 dBm, then a PDI of 10 to 12 ms and a noncoherent integration time of 1 or more seconds is required). Tables 13.5 through Table 13.7 illustrate this effect.

Equation (13.64) describes N_c as the number of total Doppler/code phase search bins for all satellites for a particular scenario. Given that the dwell time per bin is indicated by T_{dwell}, and the number of available correlators for the search is given by N_{corr}, then the maximum total search time is approximately indicated by

$$\text{T}_{\text{search}} = \text{T}_{\text{dwell}} \times \left(N_c / N_{corr}\right) \tag{13.73}$$

For a particular scenario in which the time uncertainty is 1 ms or more, full code phase search of 2,046 half-chips is required to find the first satellite. Given a condition of a 0.5-ppm oscillator, the number of Doppler bins is dominated by the oscillator uncertainty; thus, the column N_{dopp} in Table 13.5 indicates the number of Doppler bins per satellite and N_{cp} indicates the number of code phase search bins. The initial conditions of time, position, and frequency uncertainty are shown.

Table 13.5 Maximum Search Times*

Signal (dBm)	*PDI (s)*	*T_{dwell} per Bin (s)*	*N_{cp} per SV (half-chips)*	*N_{dopp} per SV*	*N_c for 8 SVs*	*T_{search} 12 (s)*	*T_{search} 32K (s)*
–130	0.001	0.002	2,046	2	32,736	5.4	0.002
–145	0.006	0.050	2,046	12	196,416	818	0.3
–150	0.012	1.0	2,046	25	409,200	9.4 hr	13
–155	0.020	5.0	2,046	42	687,456	80 hr	107

*Time uncertainty = 1 ms, frequency uncertainty = 0.5 ppm, 8 satellites, position uncertainty = 30 km, ignoring code-phase and Doppler search range reductions after finding a first satellite.

Table 13.6 Maximum Search Times*

Signal (dBm)	*PDI (s)*	*T_{dwell} per Bin (s)*	*N_{cp} First SV (Half-Chips)*	*N_{dopp} First SV*	*N_c for 8 SVs*	*T_{search} 12 (s)*	*T_{search} 32K (s)*
–130	0.001	0.002	2,046	2	6,240	1	0.002
–145	0.006	0.050	2,046	12	26,700	111	0.05
–150	0.012	1.0	2,046	25	53,300	1.23 hr	1.7
–155	0.020	5.0	2,046	42	90,230	10.4 hr	14

*Time uncertainty = 1 ms, frequency uncertainty = 0.5 ppm, 8 satellites, position uncertainty = 30 km, taking advantage of reduced code-phase and Doppler search range after finding a first satellite. Number of code phase delays and Doppler bins is reduced after finding first satellite by and reflected in total N_c.

Table 13.7 Maximum Search Times*

Signal (dBm)	*PDI (s)*	*T_{dwell} per Bin (s)*	*N_{cp} per SV (half-chips)*	*N_{dopp} per SV*	*N_c for 8 SVs*	*T_{search} 12 (s)*	*T_{search} 32K (s)*
–130	0.001	0.002	~300	1	1,636	0.27	0.002
–145	0.006	0.050	~300	1	1,636	6.8	0.05
–150	0.012	1.0	~300	2	1,841	153	1
–155	0.020	5.0	~300	4	3,682	1,535	5

*Time uncertainty = 100 μs, frequency uncertainty = 0.05 ppm, 8 satellites, Position uncertainty = 30 km, taking advantage of reduced code-phase and Doppler search range after finding a first satellite. Number of code phase delays and Doppler bins is reduced after finding first satellite and reflected in total N_c.

Column N_c indicates the number of Doppler-code phase search bins for an 8-satellite case in which the receiver does not take advantage of the code phase learned from a first detected satellite to reduce the code phase search range on the remaining satellites. Finally, two conditions are highlighted, the total search time T_{search} using (13.73) for two cases: that of a typical automotive-grade receiver containing 12 searchers, and a high-performance flash correlator that can search up to 32,000 bins simultaneously.

As described earlier, when a first satellite is detected, it is possible to substantially reduce the code phase and Doppler search range for the remaining $N_{sv} - 1$ satellites. Table 13.6 illustrates the gain achieved as reflected in reduced N_c and T_{search} cases by using the full code phase and Doppler search space to find the first

satellite, and reducing the remaining seven satellite uncertainties to approximately 300 half-chips in code phase and 100 Hz in Doppler.

Finally, Table 13.7 illustrates further reductions in the search space and search time by changing the reference oscillator to 0.05 ppm (for example, taking advantage of handset AFC tuning) and reducing the time uncertainty to 100 μs (such as taking advantage of precise time transfer).

13.4.4 GNSS Receiver Integration in Wireless Devices

As shown in (13.66), much of the total code/Doppler uncertainty space for N satellites is represented by common-mode error terms of time error and oscillator frequency error. Typical low-cost reference oscillators are in the 0.5 to 1-ppm stability range, and at this level, the oscillator frequency uncertainty is by far the largest element of the total Doppler uncertainty search space. Likewise, a 1-ms or more time uncertainty is common-mode across all satellites and forces full code phase (2,046 half-chips) scan for each satellite, time error being the largest of the possible contributors to code phase uncertainty. There are methods to remedy this common-mode frequency and time problem that are unique in a cellular handset.

With regard to time, some types of handsets such as CDMA have knowledge of precise time (submillisecond) internally as long as the handset is monitoring at least one paging channel. CDMA cell towers are synchronized in time using GPS receivers in each cell tower. The handset uses the precise time information when handing over from one cell tower to another so that it can align the cell signal spreading code phase and maintain seamless communication as the user moves from one cell tower to the next. By transferring the precise time information into the GPS function, it becomes possible to substantially reduce the contribution of time error as it reflects into the code phase uncertainty dimension [the second term in (13.68)], leaving (mostly) the contributions to position uncertainty as shown in Figure 13.40.

Certain types of handsets, such as GSM, do not have precise time information available internally. As such, methods have been devised by which the precise time necessary for the navigation solution and for predicting the submillisecond code phase can be determined in unsynchronized networks. Some of these methods include:

- Synchronizing the unsynchronized network (LMU), in which the network messaging is calibrated by an external device;
- Solving for the time-error with an overdetermined solution;
- Observing the data bits from at least one satellite navigation data message.

With regard to multiple approaches to solving the time problem in unsynchronized networks, the GSM over-the-air protocol is time-division multiplex; each handset is assigned a time-slot in which it receives and transmits packets of data between itself and the network. To accomplish precise time transfer in the asynchronous GSM network, an additional hardware element is installed in the network called a location measurement unit (LMU). The LMU contains a GPS receiver for time synchronization. It also contains a GSM phone receiver that it uses to measure

the absolute timing of certain data packets that it receives from each cell tower that it can "hear," in effect, time-tagging the bits received with GPS time. The LMU measures the time shift or time offset of each cell tower signal that it can hear and makes this time-shift information available to the cell network for delivery to those handsets desiring precise time correction. The handset accepts parameters via a network-to-handset message that allows it to instantiate a particular portion of the network to handset message with a precise time tag. As such, when the handset receives the particular portion of the network-to-handset message, it can associate the event of receiving the bits with the precise time tag (derived from the LMU), thus providing a method of time-transfer that is much better than 1 ms, or 1 GPS PRN-code time period.

Installing LMUs into a GSM network is a rather expensive proposition, so not all GSM networks will have LMUs. Network operators prefer a lower-cost alternative to deliver an approximate time estimate to the handset via a standard network-to-handset message. Network latencies in delivering the message to the handset establish the best possible accuracy of no more than ±2 seconds; thus, approximate time is useful in computing satellite Doppler when satellite ephemeris and approximate position is available, but generally is useless in computing precise code phase estimates for each satellite so as to avoid searching the entire code phase space.

However, all is not lost because, as described earlier, most receivers take advantage of the common-mode nature of time uncertainty once one satellite is detected. After detecting a first satellite generally using a full-code phase scan, the code phase uncertainty region for the remaining satellites is reduced substantially because the measured code phase from the detected satellite can be differenced with the predicted code phase (computed using the 2-second error approximate time) to provide a first estimate of the common-mode time error. This correction represents most of the common-error time contribution to code phase in Figure 13.40; the remaining satellites can be searched for using constrained or limited code phase search space, substantially reducing the size of the total Doppler/code phase uncertainty search space.

With regard to frequency, Figure 13.40 shows that the frequency uncertainty dimension of the satellite search process is dominated by the reference oscillator uncertainty. The other contributions are small with position uncertainties of tens of kilometers are assumed. If one assumes that a 0.5-to-1.0 ppm reference oscillator is used for GPS, it is by far the largest contributor to Doppler uncertainty and is common-mode across all satellites.

A cellular handset also contains a reference oscillator for its communication function, and sharing or reusing the oscillator for GPS offers a compelling cost advantage. Sharing the oscillator also enables substantial reduction in the reference oscillator frequency uncertainty because all modern cellular telephones employ a method of an AFC control loop to correct the oscillator frequency. This is based on a frequency error relative to the cellular BS-to-handset signal as shown in Figure 13.41. The frequency of the cellular BS-to-handset signal is precisely controlled by the network to better than 0.05 ppm within each network tower. As such, the handset AFC control loop adjusts the frequency of the reference oscillator (via VCO in Figure 13.41) until the frequency difference is zero. Thus, the AFC function calibrates the reference VCO oscillator to the same accuracy as the

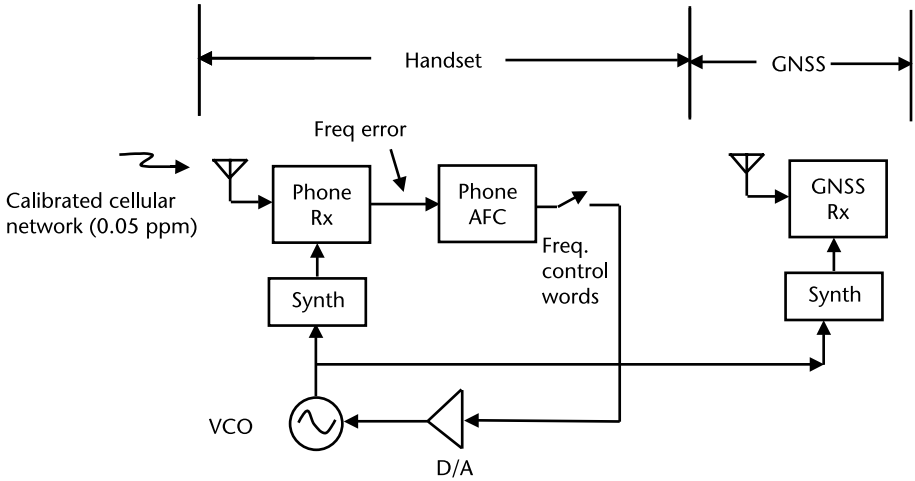


Figure 13.41 Typical handset AFC tuning of reference oscillator.

network-to-handset signal, or 0.05 ppm. Reusing this high-accuracy clock for GPS purposes enables significant reduction in the number of Doppler uncertainty search bins, contributing to lower overall TTFF and minimizing the number of required correlators needed to meet a minimum performance criteria.

Some handsets physically adjust the frequency of the reference oscillator as shown in Figure 13.41. Other handsets do not do so; instead, they let the oscillator free-run and then adjust the control registers on a fractional-N synthesizer so as to produce an adjusted frequency inside the phone receiver. The control registers of the fractional-N synthesizer are translatable into a known frequency of operation of the handset reference oscillator. This frequency is known to better than 0.05 ppm, achieving the same goal as long as the synthesizer tuning parameters are made available to the GPS function. The latter method offers significant advantages over the former, as the discrete jumps in frequency attributed to physically adjusting the reference oscillator frequency can cause data demodulation and tracking problems to the GPS function. If large enough, the instantaneous phase rotation due to the frequency jump cannot be discriminated from the ±180° phase rotations due to signal PRN modulation or navigation data bit modulation, thus confusing the data demodulation process and causing possible loss of lock.

As with time, using handset-based frequency aiding information is not absolutely necessary in order to meet the acquisition time goals of cellular A-GPS. In all cases, the battle can be won by having sufficient correlators available to search out the uncertainty space in sufficient time. However, there is a cost and power consumption penalty associated with a maximum correlator solution that is painful to overcome, at least in the near term until IC technology evolves further. As with time, one can take advantage of the common-mode nature of the reference frequency uncertainty for applications choosing to install a separate GPS reference oscillator. In this case, most of the Doppler uncertainty is due to reference oscillator uncertainty and can be solved for once one satellite is detected. Thus, the total uncertainty search space collapses significantly once a first satellite is detected and a precise Doppler measurement to it is made.

13.4.5 Sources of Network Assistance

Assistance information available from a digital cellular network is governed by applicable standards which vary with the underlying cellular technology, for example, GSM or CDMA. Generally, the standard messaging protocols have evolved to contain similar content across the various standards and have migrated from GPS-only content to include provision for non-GPS GNSS (Galileo, GLONASS, and BeiDou), regional systems such as EGNOS, QZSS, SBAS, and WAAS; and cellular communication link ranging such as enhanced observed time difference (E-OTD), uplink time difference of arrival (UTDOA), advanced forward link trilateration (AFLT), and enhanced cell-ID (E-CID). In order to simplify the discussion, this section will focus on the use of the 3GPP GSM standard for GNSS positioning because it is the most widely deployed specification. Extensions of this specification includes terrestrial location methods but are not discussed here other than illustrating how the hybrid inclusion of cellular communication link ranging, or wireless LAN data from Wi-Fi or Bluetooth, can enhance the GNSS aiding function and improve the position solution.

It is important to develop a set of agreed-upon over-the-air messaging specifications to guarantee interoperability among various handset and location technology developers. As such, telecommunication standards setting organizations, location technology developers, handset manufacturers, and carriers have incorporated provision for A-GNSS in the specifications for GSM, TDMA, CDMA, CDMA2000, and W-CDMA/UMTS.

The process to create a new over-the-air protocol can take years to develop and requires continuous updating to include new features and capabilities while guaranteeing backward compatibility. For example, the process to develop the CDMA protocol IS-801 began in late 1998 (the initial release contained 148 pages), while the 2014 revision of the specification [89], now called C.S0022, has grown to 602 pages.

The development of the standards is contribution driven: interested parties contribute written descriptions of candidate features for presentation at periodic meetings. The contribution is discussed, merits of each idea are judged, and the idea is voted upon for inclusion or exclusion. Needless to say, it is a long process to obtain agreement among all parties, update and publish the final specification, then start the process all over again for the next revision.

With regard to control plane versus user plane interfaces, there are two types of messaging interfaces in the standards to support A-GNSS functionality. The two methods are called control plane and user plane protocols [66].

The control plane is the low-level signaling layer between the cellular base station and the handset that carries the low-level signaling for call setup and traffic channel (voice and data) and is the domain of the cellular technology. It is the primary messaging that is used to support emergency location (E911, E112) in handsets as these messages are not disabled if the handset is not in service and cannot be turned off by the user.

The user plane is an Internet protocol method by which high-level applications within the mobile device can access the Internet and establish application specific messaging to support location based services such as turn-by-turn route guidance

(also known as, Google maps on a smart phone), or to enable application specific features such as find a friend (a function within the Facebook application, among many others). There are literally thousands of third-party applications for smart phones that use location and obtain this information by A-GNSS positioning enabled by user-plane (SUPL) messaging. Location information flowing over the user plane follows the Secure User Plane Location (SUPL) protocols supported by the Open Mobile Alliance (OMA). User plane messaging is sometimes called over-the-top (OTT) messaging.

Three primary organizations establish the modern protocols for handsets as shown in Table 13.8 and include control plane specifications for CDMA, GSM and LTE technologies supported by 3GPP2 and 3GPP, respectively, plus the user plane protocol supported by the SUPL Open Mobile Alliance specifications. These are the primary specifications supporting A-GNSS for mobile devices. The specifications include messaging supporting cellular-based terrestrial location methods (AFLT, E-OTD, OTDOA, and E-CID) that are used to assist A-GNSS acquisition or to combine with A-GNSS measurements enabling hybrid terrestrial and satellite location determination. The terrestrial methods are not discussed here other than how the data can support or assist A-GNSS acquisition.

GNSS systems include GPS, Galileo, SBAS, Modernized GPS, QZSS, GLONASS, and BDS.

As discussed earlier and as shown in Figure 13.33, there are two main types of A-GNSS technology; MS-assisted and MS-based. The GSM over-the-air protocol information elements to support each will be discussed next. A-GNSS assistance information from a network includes a long list of data types; the data requested by or delivered to the mobile device is dependent on the location method employed by that mobile device as depicted in Table 13.9.

Most mobile receivers use a subset of this information to acquire the requisite number of satellites for a fix, depending if it is MS-assisted or MS-based. For example, the MS-assisted handset may use visible satellite list, predicted Doppler, and predicted code phases to acquire signals. The MS-based handset may use approximate position, ephemeris, and approximate time. Both types (MS-based and MS-assisted) can transform these parameters into corresponding Doppler and Doppler uncertainty and code phase and code phase uncertainty, allowing the GNSS receiver to greatly restrict the satellite signal search region (only look where the signal

Table 13.8 Major A-GNSS Location Specifications Per Handset Type

Handset Technology	*Organization*	*Plane*	*Controlling Document*	*Location Technologies*	*Reference*
CDMA2000, CDMA (IS-95)	3GPP2	Control	C.S0022	A-GNSS, AFLT	[89]
GSM, UMTS	ETSI, 3GPP	Control	TS 44.031	A-GNSS, EOTD	[90]
LTE	3GPP	Control	TS 36.355	A-GNSS, OTDOA, E-CID	[91]
Secure User Plane Location (SUPL)	Open Mobile Alliance	User	OMA-TS-ULP-V3	A-GNSS, E-OTD, OTDOA, AFLT, E-CID	[92]

Other specifications not shown in this table describe messaging to support LAN location technologies such as Wi-Fi, Bluetooth, and others.

Table 13.9 Possible Assist Information Dependent on Mobile Location Method

Assist Information	*MS-Assisted*	*MS-Based*
Visible satellite list	X	—
Predicted SV Doppler and optionally rate of change	X	—
SV azimuth and elevation angles	X	—
SV code phase and search window	X	—
Approximate location of mobile device	—	X
Satellite almanac data (course Keplerian parameters)	—	X
Satellite ephemeris data (precise Keplerian parameters)	—	X
Satellite clock correction polynomial	X	X
Approximate time	X	X
Precise time	X	X
Navigation data bit timing (bit number, fractional bit)	X	X
Navigation data bits (sensitivity assist)	X	X

is known to be), reducing the number of correlators needed to find the signals quickly. The typical use of each data type is now discussed.

With regard to the visible satellite list (MS-Assist), the visible satellite list is generated within the cellular network by simply reporting the visible satellites at a GNSS reference receiver within or in the vicinity of the cellular network. The reference receiver should be positioned to ensure an unobstructed view of the sky. Because of the relative proximity of the network and the mobile with which it is communicating (i.e., a maximum separation of 20–30 km is expected), the visible satellite list is virtually the same for the reference and mobile receiver, except possibly for a satellite very close to the horizon (i.e., less than the separation distance divided by the radius of the Earth, or roughly a 0.2° elevation for a 20-km separation) and with an azimuth opposite to the LOS between the reference receiver and the mobile. Knowledge of the satellites that are potentially visible permits the mobile receiver to focus its search and avoid wasting time searching for satellites that are not visible, thus reducing its time to acquire sufficient satellites for a fix.

With regard to Doppler data (MS-Assist), one dimension of the two-dimensional search for a particular GNSS signal is the Doppler space dimension as shown in Figure 13.39. GNSS satellite motion-induced signal Doppler covers a large range (±4.2 kHz for GPS, ±4.5 kHz for GLONASS, ±3.6 kHz for Galileo, ±4.0 kHz for BeiDou); providing a good estimate of the Doppler for each satellite of interest drastically reduces the required number of Doppler bins (and total number of correlators) required to cover the search space and find the signal. For example, assume that a receiver uses a predetection integration time (PDI) of 10 ms in order to acquire weak signals but with no knowledge of time. At 10-ms PDI, the response of the integrate-and-dump coherent integration filter is as shown in Figure 13.42; the width of the peak with less than 1-dB attenuation is 50 Hz; this is the typical Doppler search step size to minimize the probability of missed detections.

With a 50-Hz step size, 168 Doppler search bins (2×4.2 kHz/50 Hz) are required to cover the entire Doppler search space. In contrast, given knowledge of the satellite Doppler, the search range can be restricted to a level that is consistent with maximum expected host velocity, initial position uncertainty and reference oscillator frequency uncertainty [the last three terms of (13.42)], or only about 5 Doppler

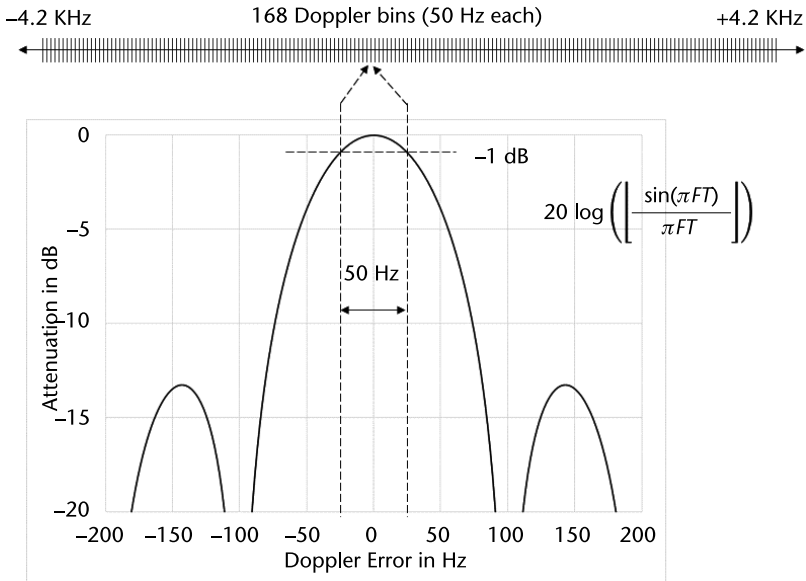


Figure 13.42 Response of integrate-and-dump filter dependent on frequency error F and predetection integration time T (PDI). T = 10 ms shown.

bins. Because satellite Doppler rate of change is relatively small compared to the Doppler itself [88], the Doppler rate-of-change parameter is used to keep the search signal Doppler stationary with the during very long noncoherent accumulations (weak signals), or to enable periodic fixes with only one Doppler assist packet, for example, to predict the current Doppler minutes after receipt of the Doppler assist packet.

With regard to the code phase and code phase search window data (MS-Assist), as shown in Figure 13.39, the second dimension of searching for a GNSS spread-spectrum signal is the code phase dimension, typically in units of spreading code chips. For the case of GPS the code phase is measured in C/A code chips (0 to 1,022) each chip about 300m in size. Other GNSS define code phase commensurate with the structure of its spreading code. The network supplies an estimate of the code phase for each visible satellite based on the approximate position of the mobile, the network knows the mobile approximate position based on a number of options such as CID (the location of the cell tower communicating with the mobile), AFLT or E-OTD location (triangulation of the mobile location from cellular signals), or even LAN location [knowledge of the location of wireless WLAN network(s) that the mobile station can presently receive]. Code phase knowledge for each visible satellite allows the receiver to restrict the search to the predicted code phase delay and search window.

The code phase search window describes the range of code phases around the estimated code phase where the GNSS spreading code is most likely to be found. With knowledge of only the CID, the network will set the search window commensurate with the size of the cellular network coverage for the particular cellular tower. The code phase search window can be greatly reduced if the network can determine a more refined approximate location of the mobile by using AFLT/E-OTD methods or WLAN location.

With regard to azimuth and elevation information (MS-Assist), satellite azimuth and elevation angles can be used by a mobile receiver in its assignment of search ranges. For example, in the previous paragraph, the number of Doppler bins assigned was computed solely as a function of the maximum expected host velocity. This calculation ignores the fact that Earth-borne host velocities are largest in the horizontal plane; a more realistic assignment of search range could therefore have been made using the elevation angle of the to-be-acquired satellite, E, as indicated in the following equation:

$$\Delta D = v_{H\max} \cos E + v_{z\max} \sin E \tag{13.74}$$

Since maximum vertical velocities (i.e., $v_{z\max}$) are expected to be small relative to maximum horizontal velocities (i.e., $v_{H\max}$), smaller search ranges would generally be assigned to higher elevation satellites using (13.74).

With regard to approximate location data (MS-Based), providing the mobile receiver with an approximate location is most useful for MS-based acquisition assistance when combined with either ephemeris or almanac data for the satellites expected to be visible. The position provided by the network is generally either the location of the serving cell tower or the center of the service area; it is therefore expected to be within 20 km of the mobile's actual location. Given this position, and either an ephemeris or almanac representation for each satellite, Doppler and Doppler rate information can be computed by the mobile with satisfactory accuracy (the sensitivity of Doppler prediction error to position error is generally less than 1 Hz/km [88]), the value of which for acquisition assistance has already been discussed.

With regard to ephemeris data (MS-based), as referenced in the preceding paragraph, satellite ephemeris or almanac information enables accurate Doppler prediction, given relatively coarse position information. In addition, if time is known such that the satellite positions can be accurately computed (a 1-second error in knowledge of time translates to 1 km of ranging error in the worst case), an accurate range to the GNSS satellite can be determined. If, additionally, the handset is precise time-synchronized (example, CDMA handsets), prediction of the satellite code phase can be made as described in (13.73) to substantially reduce the range of code phases to search. For example, if the local oscillator has been synchronized, and time is known to 1 second, the relative code phases can be resolved to roughly 140 half-chips (i.e., 21 km of ranging error) after finding a first satellite, representing a significant savings relative to a full code phase search.

The information, provided to assist acquisition, can also increase sensitivity (i.e., enable acquisition of weaker signals). This is because the assistance information is likely to reduce the search ranges in Doppler and code phase such that the receiver has sufficient correlators to cover all cells in a parallel search, spending more time searching the remaining space.

With regard to sensitivity assist, independent of the acquisition assistance types already discussed, the primary form for actual sensitivity-increasing assistance data is the provision of navigation data bits over the cellular network. Given that the navigation data bits can be synchronized with the knowledge of the data bit edges for each satellite for which acquisition is attempted, the PDI can be extended beyond one navigation data bit: each doubling of the coherent integration time

lowers the acquisition threshold by 3 dB. However, each doubling of the coherent integration period requires a correspondingly narrower Doppler size due to the SINC function, and so more Doppler bins (and more correlators) will be required to cover the same uncertainty range.

Some of the over-the-air protocols for A-GNSS have provisioned methods of sending the navigation data message to the handset so it has a priori knowledge of each bit and can subsequently wipe-off the data if needed for additional signal processing gain. The handset has to assemble the total bit sequence through a number of different messages. For example, in the GSM protocol, most of the bits for each GPS satellite from subframes 1–3 (words 3 through 10) are delivered via the Navigation Model assist data message. Most of the bits for subframes 4 and 5 (words 3 through 10) are delivered via the Almanac assist data message. The bits contained in each word's 6-bit parity field are not sent, as these are computable after the handset has the data elements. Each subframe has a constant preamble that does not need to be sent, and the 17-bit HOW word contained in each subframe (word 2) is predictable with time. Thus, the remaining missing bits in the navigation message, primarily the TLM Message (14 bits), the antispoof flag (1 bit), the alert flag (1 bit), and the TLM-reserved bits (2 bits) have been accumulated into one additional garbage collection message and appended at the end of the Reference Time network-to-handset message.

At least two alternatives exist to sending and receiving navigation data bits over the network that achieve most of the benefit: predicting the navigation data bits, as discussed earlier, and guessing the navigation data bits. Estimated bits [93] can substantially increase the required number of correlators for longer coherent integrations. In guessing the navigation data bits, a hypothesis corresponding to each possible bit transition is formulated, and parallel integrations are performed, with the integration resulting in the largest signal correlation peak determined to be the correct bit sequence. For a sequence of n data bits, $2n$ parallel integrations are required, corresponding to each hypothesized bit sequence. This increases the number of correlators dedicated to each satellite for which bits are guessed by $2n$. In [93], a practical limit of 5 estimated bits is imposed, corresponding to 32 parallel integrations. Modernized signals such as the GPS L2C signal offer a dataless component to the signal, which allows for long coherent integration periods without regard to data bit modulation interference, eliminating the need to transmit the bits to the mobile user through a cellular network.

With regard to navigation assistance, in many cases the solution geometry can be significantly degraded relative to open-sky conditions. Hence, each meter of ranging error can be scaled by a large multiplier related to the geometry (e.g., HDOP), resulting in a significant navigation error, which can be reduced through the use of differential corrections.

At least in the United States, as new indoor and vertical position accuracy requirements eventually take hold, the importance of the Z-dimension will drive solutions [74]. Existing GNSS solutions can be extended to include altitude information from multiple sources such as barometric pressure sensors in the handset coupled with calibrating barometric pressure/altitude data sent via the cellular network to the handset. Other sources of altitude could come from WLAN location information such as Wi-Fi or Bluetooth [63, 75].

The approximate altitude of the mobile can be combined with a GNSS solution to provide a hybrid location accurate in the Z dimension. It is reasonably important that an accuracy measure be provided for the altitude, generally represented as a 1-sigma error: The Z measure is then most readily incorporated into a weighted least squares solution for the mobile position. Thus, the altitude is added as an additional measurement to the m pseudorange measurements, z_{m+1}, with an error variance set to the square of the 1-sigma value from the network:

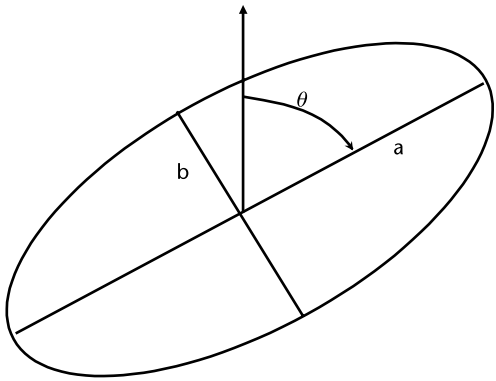
$$\mathbf{x} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z} \tag{13.75}$$

Note that bold letters denote vectors in (13.75), that is, the $m + 1$ dimensional measurement vector **z**, and the four-dimensional vector of state corrections, **x**. The measurement gradient matrix **H** is of dimension $m + 1$ by 4, with its first m rows corresponding to the m pseudorange measurements, and the last row corresponding to the altitude, and **R** is the $m + 1$ dimensional diagonal measurement error variance matrix with each element representing an error variance assigned to the corresponding measurement.

The approximate location which is communicated to the mobile can serve two functions for the navigation solution. The first is simply to initialize the WLS solution, that is, provide a starting point for its iterations, the x value in (13.75), which is defined as a set of corrections relative to this, initial, supplied location. In order for (13.75) to be valid, the approximate location must be sufficiently accurate such that the pseudorange measurements are effectively linearized. A second function is to add horizontal position domain constraints to the WLS solution, in the same way in which the altitude constraint is added. The dimension of the measurement vector, **z**, is then increased to $m + 3$, where m is the number of pseudorange measurements, and the **R** matrix elements corresponding to the position constraints are assigned error variances which reflect the accuracy of the approximate location. As referenced in (13.65), error variances, perhaps in the form of an error ellipse, are communicated with the approximate position. The error ellipse can be communicated (as an orientation angle and one sigma errors in principal axes), as illustrated in Figure 13.43 when the approximate position is determined from a coarse fix (e.g., based upon ranging off the cellular signals). In the figure, σ_e^2, σ_n^2, and σ_{en} denote the elements of the covariance matrix corresponding to East and North position error. In the case of an error ellipse, the East and North position error components will generally be correlated (i.e., a nondiagonal measurement error variance will be needed if the constraints are expressed directly in terms of East and North position error components). Preferably, the measurement error variance matrix can remain diagonal if the measurements are expressed in the principal axes.

In a manner completely analogous to the addition of position constraints as additional measurements in the WLS solution for location, a timing constraint can be added to the clock offset solution, if fine timing information is available and sufficiently accurate (i.e., submillisecond).

Finally, the mobile's navigation solution can be aided by the transmission of satellite clock correction and ephemeris data, which may already be part of the acquisition assistance. However, for a handset-based solution in response to an emer-



$$a = \sqrt{\frac{1}{2}(\sigma_n^2+\sigma_e^2) + \sqrt{\frac{1}{4}(\sigma_n^2+\sigma_e^2)^2+\sigma_{ne}^2}}$$

$$b = \sqrt{\frac{1}{2}(\sigma_n^2+\sigma_e^2) - \sqrt{\frac{1}{4}(\sigma_n^2+\sigma_e^2)^2+\sigma_{ne}^2}}$$

$$\theta = -\frac{1}{2}\tan^{-1}\left(\frac{2\sigma_{ne}}{\sigma_e^2-\sigma_n^2}\right)$$

Figure 13.43 Error ellipse relationship to covariance matrix.

gency call, both are generally required for an accurate solution, since time does not permit decoding of the equivalent information from the navigation data bits.

With regard to A-GNSS data exchanges, it is useful to demonstrate the MS-Assist and MS-Based data exchange for the typical mobile device. The examples that follow show data content specific to the A-GPS case, although the protocols support GPS, Galileo, SBAS, Modernized GPS, QZSS, GLONASS, and the BeiDou Navigation Satellite System. Later, we show examples of how the standards support the non-GPS satellite systems.

The data content of the MS-Assist and MS-Based exchanges among all the varying GNSS systems are similar to the A-GPS case, unique differences between the systems are reflected in the number of bits of certain fields (differing methods of specifying the navigational models as reflected by the slight variation in their constellation orbital parameters and orbit altitudes), and the way in which certain parameters are defined (for example, code phase in C/A code chips for GPS, code phase in units of fractions of a millisecond for Galileo). If the reader becomes familiar with the A-GPS exchange first, it is relatively easy to translate to Galileo, GLONASS, or BeiDou given an understanding of the GNSS system differences reflected in the rest of this book. For purposes of simplifying the discussion, the MS-Assist and MS-Based exchanges as described in the TS-144.031 GSM/UMTS specification [90] are used.

The specification allows for multiple GNSS systems to be included or combined into a hybrid multi-GNSS solution in order to improve performance (accuracy), and availability of measurements when an A-GNSS handset has many more opportunities to find signals in challenging environments compared to A-GPS solution. Given that most GNSS systems have at least one signal on the original GPS L1

frequency (1,575.42 MHz), the RF electronics (antenna, RF gain, downconverter) necessary to process multi-GNSS signals simultaneously is simplified, the increased complexity is limited to the low-cost digital processing and software functions.

With regard to the MS-Assist Exchange, referring to Figure 13.33, recall that the MS-assist method moves the position computation element to the network-based position computation server, called a position determining entity (PDE) in a CDMA network or called a serving mobile location center (SMLC) in a GSM network. Information flows from the network to the handset to enable the handset-based GNSS receiver to acquire, detect, and measure pseudoranges to multiple satellites. The handset then returns measured code phases, Doppler, and signal power estimates for each detected satellite.

In the MS-Assist mode, the exchange begins when the handset requests an Acquisition Assistance message from the network. Tables 13.10 and 13.11 show the information content of the Acquisition Assistance message that is promptly delivered from the network to the handset via a short digital message. The data in Table 13.10 is sent once, while the data in Table 13.11 is sent for each visible satellite data in the assist message set.

The parameters Doppler uncertainty and code phase search window correspond to the network's estimate of Doppler uncertainty and code phase uncertainty depicted in Figure 13.39. The parameters of code phase, integer code phase, and GPS bit number correspond to "Code Phase_observed," "Integer_ PN_Rolls_observed," and "Bit_Number_observed" in (13.71); however, "GPS_ Bit_number" is truncated further to just two bits by a modulo function, such that GPS_Bit_number = MOD(Bit_Number_observed, 4). In Table 13.10, the parameter "GPS TOW" represents the time-tag corresponding to the data contained in Table 13.11 and is analogous to "GPS_Time" in (13.71). In Table 13.10, the parameters BCCH carrier, BSIC, frame number, timeslot number, and bit number represent the LMU-generated parameters that link the asynchronous cellular messaging protocol state at an instant in time to the corresponding GPS-TOW time tag (i.e., the cellular messaging protocol state defined by these parameters existed at the precise GPS-TOW time tag that enables precise time transfer). The use of the time transfer parameters are optional, O (marked as O in the table), which means the handset does not necessarily need to use the data, and, in fact, the parameters may be missing from certain cellular networks if the network operator does not want to deploy LMUs. Those elements marked as M are mandatory.

The parameters of azimuth and elevation provide the MS-Assisted handset with the ability to compute approximate HDOP (see Chapter 11) as it acquires satellites.

Table 13.10 GPS Acquisition Assist: Parameters Appearing Once Per Message

Parameter		*Range*	*Bits*	*Resolution*	*Including*
Number of satellites		0–15	4		M
Reference time	*GPS TOW*	0–604799.92sec	23	0.08 second	M
	BCCH carrier	0–1,023	10		O[1]
	BSIC	0–63	6		O[1]
	Frame #	0–2,097,151	21		O[1]
	Timeslots #	0–7	3		O[1]
	Bit #	0–156	8		O[1]

Table 13.11 GPS Acquisition Assist: Parameters Appearing (Number of Satellites) Times Per Message

Parameter	*Range*	*Bits*	*Resolution*	*Incl.*
SVID/PRNID	1–64 (0–63)	6		M
Doppler (zeroth-order term)	–5,120–5,117.5 Hz	12	2.5 Hz	M
Doppler (first-order term)	–1–0.5	6		O[1]
Doppler uncertainty	12.5–200 Hz [2^{-n}(200) Hz, n = 0–4]	3		O[1]
Code phase	0–1,022 chips	10	1 chip	M
Integer code phase	0–19	5	1 C/A period	M
GPS bit number	0–3	2		M
Code phase search window	1–192 chips	4		M
Azimuth	0–348.75°	5	11.25°	O[2]
Elevation	0–78.75°	3	11.25°	O[2]

Without some form of geometry quality indicator, the MS-Assisted handset does not know when it has detected sufficient satellites for a good fix. Thus, being able to compute HDOP after each subsequent new satellite detection enables the MS-Assisted handset to know when it has detected sufficient satellites for a quality fix and to deliver the pseudorange measurement response message to the network.

When the MS-Assisted handset acquires sufficient satellites for a good fix, it returns the measured pseudorange data to the network via a measurement information element response message, depicted in Tables 13.12 and 13.13. As before, Table 13.12 data is sent one time and Table 13.13 data is sent N-SAT times. Table 13.13 contains the receiver-measured pseudorange data that the network-based SLMC will use to compute the position of the handset.

With regard to the MS-Based exchange, referring to Figure 13.33, recall that the MS-Based method provides the position computation element in the handset, enabling local applications such as personal navigation or mapping to operate within the handset. To do so, the handset will need a fresh copy of satellite ephemeris data as it needs to know precise satellite positions in order to compute range residuals and update its local estimate of user position. Thus, one of the data elements the MS-Based handset will need from the cellular network is the real-time (current) ephemeris data. The ephemeris is also useful in computing local acquisition assist data given that the handset also knows approximate position and time. Two other data elements that are obtainable by the handset include this additional data.

The MS-Based handset has a number of things that it can request from the cellular network. Table 13.14 describes the suite of assistance data elements that can be requested by the handset. The handset can individually select each or all data elements listed in the table in one uploaded request to the network; thus, the handset that is *cold* (i.e., no time, position, ionosphere correction, or ephemeris data) can request the entire load in one uploaded message and then accept each *assist* data element as it is delivered from the network message.

Each data element is formatted into a unique data message before being sent to the handset. For a detailed description of each data element message, refer to [90]. The handset uses the assist data, transforms it into Doppler, code phase estimates and uncertainties as described in Figure 13.39. The signals are acquired and

Table 13.12 GPS TOW Field Contents

Parameter	*No. of Bits*	*Resolution*	*Range*	*Units*
Reference frame	16	—	0–65,535	frames
GPS TOW	24	1 ms	0–14,399,999	ms
N_SAT	4	—	1–16	—

Table 13.13 Measurement Parameters Field Contents

Parameter	*No. of Bits*	*Resolution*	*Range*	*Units*
Satellite ID	6	—	0–63	—
C/N_0	6	1	0–63	dB-Hz
Doppler	16	0.2	±6,553.6	Hz
Whole chips	10	1	0–1,022	chips
Fractional chips	11	2^{-10}	$0–(1–2^{-10})$	chips
Multipath indicator	2	4 levels		—
Pseudorange RMS error	6	3-bit mantissa, 3-bit exp	0.5–112	m

position determined using the locally stored ephemeris and ionospheric correction constants. The handset then can return the position to the network via one of five different digital messages [94]. The messages contain the user position data along with optional uncertainty and altitude. The optional messages include:

- Ellipsoid point;
- Ellipsoid point with uncertainty circle;
- Ellipsoid point with uncertainty ellipse;
- Ellipsoid point with altitude;
- Ellipsoid point with altitude and uncertainty ellipse.

The most general option, ellipsoid point with altitude and uncertainty ellipse from [95], is detailed in Table 13.15.

Table 13.14 Fields in the GPS Assistance Data Element

Parameter	*Presence*
Reference time	O
Reference location	O
DGPS corrections	O
Navigation model	O
Ionospheric model	O
UTC model	O
Almanac	O
Acquisition assistance	O
Real-time integrity	O

One example of how the cellular over-the-air protocol can be used to solve a particular handset application problem is demonstrated by the Real-Time-Integrity acquisition assist data element. In Section 11.4, the importance of ensuring integrity for GNSS is discussed, since GNSS satellite clocks can fail, resulting in significant error in unprotected receiver solutions. A GNSS receiver embedded in a cellular handset cannot generally be expected to perform its own RAIM (see Section 11.4.3.1) function, since signal reception conditions may be poor, and the luxury of redundant measurements may not exist. For this reason, the cellular standards have allowed for integrity information to be communicated to the handset, since a network-based GPS receiver will certainly be able to perform the RAIM function and identify which satellites are failed or failing. It should be noted that the historical failure rate of GNSS satellites or ephemeris uploads to those satellites has been very low, approximately one event every 18 to 24 months. However, when GNSS is used for high-frequency emergency location function, it is certain that someone will need the system at precisely the time a satellite fails. Consequently, a real-time integrity function was added to the Radio Resources Location Services Protocol (RRLP) to prevent such failures.

An MS-Based handset can be particularly vulnerable to satellite failures. The handset can request real-time ephemeris data from the cellular network and then subsequently use the data for several hours. One mode that may be used by the handset is a periodic fix mode, in which the handset accepts the ephemeris assist data and then computes position at some periodic rate (e.g., once per minute). The handset only needs to get current ephemeris for each satellite at the start, as its

Table 13.15 Position Response Data Element

Information Element/ Group Name	*Type and Reference*	*Semantics Description*
Latitude sign	Enumerated (North, South)	
Degrees of latitude	Integer (0...2^{23} – 1)	The IE value (N) is derived by this formula: $N \leq 2^{23}\ X\ /90 < N+1$ with X being the latitude in degree (0°.. 90°)
Degrees of Longitude	Integer (–2^{23}...2^{23} – 1)	The IE value (N) is derived by this formula: $N \leq 2^{24}\ X\ /360 < N+1$ with X being the longitude in degree (–180°..+180°)
Altitude direction	Enumerated (height, depth)	
Altitude	Integer (0..2^{15}–1)	The IE value (N) is derived by this formula: $N \leq a < N+1$ with a being the altitude in meters
Uncertainty semi-major	Integer (0...127)	The uncertainty r is derived from the uncertainty code k by $r = 10 \times (1.1^k – 1)$
Uncertainty semi-minor	Integer (0...127)	The uncertainty r is derived from the uncertainty code k by $r = 10 \times (1.1^k – 1)$
Orientation of major axis	Integer (0..89)	The IE value (N) is derived by this formula: $2N \leq a < 2(N + 1)$ with a being the orientation in degree (0°.. 179°)
Uncertainty altitude	Integer (0..127)	The uncertainty in altitude, h, expressed in meters is mapped from the IE value (K), with the following formula: $h = C((1+x)^\kappa – 1)$ with $C = 45$ and $x = 0.025$
Confidence	Integer (0..100)	In percentage

useful life is ±2 hours around the time of ephemeris (TOE). In an assisted mode, the handset may never observe the satellite-broadcast real-time integrity data that is available in the 50-bps satellite navigation data message. Thus, if a particular satellite fails between the time the handset accepts ephemeris and the time it wants to use it for a position solution, the handset will not have knowledge of the failed state and could produce erroneous position data.

To combat this potential problem, a short real-time integrity message was added to the RRLP to inform the handset when a particular satellite has failed. The real-time-integrity message is requested at the start of each location attempt and consumes only a few bits of the available bandwidth. The network generated real-time-integrity message is then sent to the handset. For the case of no failed satellites, this message returns one 0 bit. For the case of a failed satellite or group of failed satellites, the satellite IDs of the failed satellite(s) are returned to the handset; the handset excludes those failed satellites from any subsequent position solution. As such, the MS-Based handset needs to request real-time integrity information at the start of each location attempt to ensure solution integrity.

With regard to GNSS support in the standards, as mentioned earlier, the MS-Assist and MS-Based data exchanges described above are examples specific to GPS. The over-the-air standard for GSM [90] (and others) have been augmented to support all the other modern GNSS systems using GPS-like MS-Assist and MS-Based messaging for Galileo, SBAS, Modernized GPS, QZSS, GLONASS, and BeiDou as well as continuing to support the E-OTD terrestrial positioning method. The standard uses the term GANSS to describe the non-GPS GNSS systems; GANSS means "Galileo and Additional Navigation Satellite Systems." To preserve backward compatibility to earlier versions of the specification, the GPS assist and response messages are not changed, new messaging was added to the standard to support the other GNSS systems.

Using the MS-Based positioning method as an example, any or all of the various GNSS constellations can be used in the handset-computed position response message. The handset informs the network which system(s) it is capable of using, and tells the network which ones were used to create the position response message using a 16-bit-wide bit field, each bit representing one of the GNSS constellations, such as:

- Bit 0: E-OTD;
- Bit 1: GPS;
- Bit 2: Galileo;
- Bit 3: SBAS;
- Bit 4: Modernized GPS;
- Bit 5: QZSS;
- Bit 6: GLONASS;
- Bit 7: Biedou;
- Bits 8 to 15: These bits are reserved for future use.

Any combination of systems can be used in determining the position data; for example, E-OTD + GPS + Galileo can be combined into a single hybrid solution to

improve accuracy and availability. As noted earlier, the standards are under continuous update and the authors have left room for GANSS expansion by leaving 8 additional reserve bits for future use.

There is considerable commonality between the satellite navigation systems that the assist information can be described in terms of a common set of GANSS assist data, including:

- Reference time;
- Reference location;
- Ionospheric model;
- Additional ionospheric model;
- Earth orientation parameters;
- Reference time extension.

These apply to any and all GANSS. The unique aspects of each GNSS constellation is described with a set of GANSS-generic assist messages, which includes unique messages set for each particular GNSS constellation. For example some GNSS systems describe the satellite orbit model using standard Keplerian orbit parameters (e.g., GPS, Galileo, BeiDou), while GLONASS prefers to represent the satellite orbit model using a second-order curve fit in the Cartesian coordinate system.

The full set of Generic Assistance Messaging is shown in Table 13.16 and includes a description of the message content and use. Within each of the assist data type (e.g., the GANSS Navigation Model), the standard includes a detailed description the assist data unique to each satellite navigation system.

The standards document consumes 35 pages to describe the details of every parameter and field shown in Table 13.16; we will not go into such detail here. For further details of all other parameters, refer to [90].

To illustrate how the specification handles GNSS uniqueness, let us look at how the navigation model is described. The navigation model includes the unique ways to describe the precise orbit of the satellite (i.e., position and velocity vector as a function of time), as well as how to describe the satellite clock error. Here, for illustrative purposes, we focus only on the orbit models.

The orbit models are defined using six different models unique to each GANSS, and are labeled Model-1 through Model-6. Model-4 and Model-5 are unique to GLONASS and SBAS and use Earth-centered Earth-fixed curve fit parameters to describe the satellite orbit. The name for each parameter, number of bits, scale factor, and units of measure that are native to the GNSS system are used to specify these two models. Nine parameters describing the zeroth-, first-, and second-order parameters for the X, Y, and Z coordinate as shown in Table 13.17 are needed to specify satellite as a function of time, the receiver reconstructs the satellite location coordinates with three simple equations:

$$\begin{aligned} X(t) &= X_0 + \dot{X}\left(t - T_{ref}\right) + \ddot{X}(t - T_{ref})^2 \\ Y(t) &= Y_0 + \dot{Y}\left(t - T_{ref}\right) + \ddot{Y}(t - T_{ref})^2 \\ Z(t) &= Z_0 + \dot{Z}\left(t - T_{ref}\right) + \ddot{Z}(t - T_{ref})^2 \end{aligned} \tag{13.76}$$

Table 13.16 Generic Assist Data Content and Summary of the Data Use

Generic Assist Data Type	*Applicable to the GANSS System(s) in Use, What the Data Are Used For*
GANSS ID	Defines which satellite navigation systems are in use
GANSS Time Model	Defines the time using the time coordinate system model for the constellation(s) of interest
DGANSS Corrections	Differential range and range-rate corrections
GANSS Navigation Model	Precise satellite orbit position and velocity model and satellite clock models
GANSS Real-Time Integrity	Real-time information regarding current satellite health and usability
GANSS Data Bit Assistance	Sensitivity enhancement: navigation data bits needed to wipe off the data modulation, thus enabling lower signal detection
GANSS Reference Measurement Information	The MS-Assist message containing Doppler, code phase, code phase search window, and other information for the GANSS in use; the generic form of code phase and search window is in terms of fractions of a millisecond, not using the GPS method of a number of spreading code chips
GANSS Almanac Model	Coarse satellite position data unique to the GANSS in use
GANSS UTC Model	Corrections from the GANSS time reference to UTC
GANSS Ephemeris Extension	Orbit and satellite clock model: method to extend period of applicability of the orbit model beyond the traditional applicability of the broadcast ephemeris, using delta additions to a baseline ephemeris set
GANSS Ephemeris Ext Check	Defines the applicability of the ephemeris extension data
SBAS ID	If GANSS ID indicates SBAS, this field further defines which SBAS is used
GANSS Additional UTC Model	Additional parameters for UTC for constellations not included in the UTC model above
GANSS Auxiliary Information	Additional information dependent on the GANSS ID; provided together with other satellite dependent GANSS assistance data
DGANSS Corrections Validity Period	Period of applicability for the differential corrections
GANSS Time Model Extension	An extension to the time model, if needed
GANSS Reference Meas. Extension	An extension to the GANSS Reference Measurement Information (e.g., higher resolution satellite azimuth and elevation angles)
GANSS Almanac Model Extension	A single bit indicating if the full almanac model was provided above
GANSS Almanac Model Extension-R12	Unique extension if the Almanac model is for Galileo
GANSS Reference Meas. Extension-R12	MS Assist data extension defining the Doppler uncertainty search window
DBDS Corrections	BeiDou system unique differential corrections and reference time
BDS Grid Model	Parameters are used to estimate the ionospheric distortions on BeiDou pseudoranges

As shown in Table 13.18, Models-1, 2, 3, and 6 use traditional Keplerian orbit parameters similar to GPS and consist of between 16 and 19 parameters. Model-1 and-2 use the same Keplerian formulation except for the addition of URA Index

Table 13.17 Model-4 (GLONASS) and Model-5 (SBAS) Orbit Model Parameters

Parameter	*Model-4 (GLONASS)*			*Model-5 (SBAS)*		
	Bits (each)	*Scale*	*Units*	*Bits (each)*	*Scale (each)*	*Units*
X_0, Y_0, Z_0	27	2^{-11}	km	30, 30, 25	0.08, 0.08, 0.4	m
$\dot{X}, \dot{Y}, \dot{Z}$	24	2^{-20}	km/sec	17, 17, 18	1/1,600, 1/1,600, 1/250	m/s
$\ddot{X}, \ddot{Y}, \ddot{Z}$	5	2^{-30}	km/sec^2	10, 10, 10	1/80,000, 1/80,000, 1/16,000	m/sec^2

and Fit Interval Flag (in Model 2) and different scale factors for t_{oe}. Model-1 is used for Galileo. Model-2 is used for NAV parameters in Modernized GPS. Model-3 is also used for Modernized GPS for L2C and L5, while Model-6 is unique to the BeiDou system. These models allow the receiver to determine the position and velocity vector of the corresponding satellite as a function of time over a predetermined period of applicability.

13.5 Hybrid Positioning in Mobile Devices

This section examines hybrid positioning systems found in mobile devices including smart phones and tablets. Low-cost sensors, alternate positioning systems, and methods used to augment GNSS solutions are presented, and example systems are discussed.

13.5.1 Introduction

Since the early 2000s, small, low-cost GPS receivers have been integrated into cellular phones for emergency location in 911 calls (E-911). As smart phones emerged, they utilized more powerful GPS receivers that could also be used for continuous positioning. More recently, GLONASS, BeiDou, and then Galileo tracking capabilities have been integrated to improve coverage and accuracy in dense urban areas. These phones now dominate the market and are used for a wide range of location-based services (LBS) such as finding nearby services (gas stations, restaurants, and shops), personal navigation, tracking workers, locating friends and family members, health and fitness monitoring, social media updates, and gathering location history for mobile marketing. Such phones also incorporate various low-cost MEMS sensors for context awareness.

Users expect their smart phones to work in all environments including indoors, parking garages, and in dense urban canyons. Wireless signal coverage is the most important service to have ubiquitous coverage; however, users also desire and expect the accurate positioning function to work well in all environments. Outdoors, height is presumed to be on the ground, but indoors, a user might be on any level and hence floor-level determination is an important capability for indoor positioning. In the case of emergency response, accurate positioning indoors including the right floor level is vital to finding where to send needed help. Recent advancements

Table 13.18 Keplerian Orbit Models Specified for Model-1, 2, 3, and 6 for Use with Galileo (Keplerian), Modernized GPS (NAV, CNAV), and BieDou, Respectively

Parameter	*Model-1*	*Model-2*	*Model-3*	*Model-6*
1	t_{oe}	URA Index	T_{op}	AODE
2	ω	Fit Interval flag	URA_{oe} Index	URA_Index
3	Δn	t_{oe}	ΔA	t_{oe}
4	M_0	ω	A_dot	$A^{1/2}$
5	OMEGAdot	Δn	Δn_0	e
6	e	M_0	Δn_{0_dot}	ω
7	Idot	OMEGAdot	$M_{0\text{-}n}$	Δn
8	sqrtA	e	e_n	M_0
9	i_0	Idot	ω_n	Ω_0
10	$OMEGA_0$	sqrtA	$\Omega_{0\text{-}n}$	Ω_0 dot
11	Crs	i_0	$\Delta\Omega_{\_dot}$	i_0
12	Cis	$OMEGA_0$	$i_{0\text{-}n}$	Idot
13	Cus	Crs	$i_{0\text{-}n}$ dot	C_{uc}
14	Crc	Cis	Crs-n	C_{us}
15	Cic	Cus	Cis-n	C_{rc}
16	Cuc	Crc	Cus-n	C_{rs}
17	—	Cic	Crc-n	C_{ic}
18	—	Cuc	Cic-n	C_{is}
19	—	—	Cuc-n	—

in Assisted-GNSS technology have enabled improved positioning indoors (as seen in Section 13.4), but GNSS receivers are still not sensitive enough to determine position everywhere that users carry their devices nor can they perform as accurately indoors as outdoors. Any solution to improving coverage and accuracy indoors must also be low cost and low power.

Modern smart phones are equipped with an increasing array of MEMS sensors including accelerometers, magnetometers, gyroscopes, and barometers. These sensors have been discussed in Section 13.3.2. The main differences between sensors used in vehicle and mobile phone applications are the cost, performance, and power characteristics. A mobile phone has to have sensors that are smaller, lower-power, and less expensive than a vehicle can accommodate. As a result, the performance is also lower. Nevertheless, the sensors are available on the platform and can be leveraged for positioning.

Smart phones today also have access to Wi-Fi, Bluetooth, and NFC signals. In the future, there may be other RF radios that become common in phones. Each of these wireless communications technologies can also be used for positioning given appropriate location data is available about the individual transmitters. Such a database is typically created through surveying target environments or through anonymous crowd sourcing [96].

Other potential positioning solutions for mobile devices include adaptations to mobile phone transmitters to support more accurate positioning, the NextNav network [97], and other dedicated beacons for indoor positioning. Each of these technologies has advantages and disadvantages for use as an indoor solution.

The observations from all of these sources can be fused together using a Kalman filter to maximize positioning coverage and accuracy. Care must also be taken to manage these various sensors to that they are only used when necessary in order to keep power draw to a minimum. The result is continuous position availability in indoor environments.

13.5.1.1 Target Use Cases

Mobile phones can be located indoors using methods including A-GNSS, AFLT, or Wi-Fi positioning. Tablets that do not have cellular service rely on A-GNSS and Wi-Fi positioning. Typically, it takes several seconds to determine a fix indoors and the accuracy is not as good as a GNSS fix outside. It is also not practical to get continuous position updates for use in tracking, fitness, or navigation systems.

Wi-Fi positioning on its own has improved the availability of position fixes indoors and also the time to get an initial fix. However, the positioning of the wireless access point (WAP) transmitters is somtimes based solely on surveys that have been done using GNSS positioning from the outside of a building where GNSS is available, so the determined positions tend to also be outside even when the mobile device is indoors.

Consumers use their handsets, tablets, and other mobile devices for myriad applications, many of which use location. Typical indoor uses include [97]:

- Find where I am currently located and show the position on a map.
- Find where my friend or family member is located and show the position on a map.
- Find the nearest restaurant, store, restroom or other point of interest (POI).
- Show walking directions to the chosen location or POI.
- Give voice prompted turn-by-turn walking guidance to the chosen location or POI.
- Show my journey progress on a map.
- Record my position (geotag) on a photo so that I can sort or plot based on a position later.
- Check in at a location on Facebook, Yelp, or other social media service.
- Receive context and location-aware messages and promotions from advertisers and venue owners.
- Let the emergency operator know my position and floor level for emergency assistance.

In order for the mobile device to be able to reliably perform these functions, the positioning system must be able to do the following:

- Determine position quickly, within 2 to 3 seconds.
- Determine position including floor level accurately, within 5 to 10m (CEP 50%).

- Determine position updates at 1-Hz rate when needed for journey or tracking.
- Minimize impact on battery life of the device.

Fitness products use location for recording distance traveled, speed, elevation, and calorie counting and for showing a track of running or cycling workouts. Users value accuracy and a fast start-up time when they are about to begin a workout. The positioning system needs to be able to determine position continuously, but not necessarily show the position updates continuously in real time. Wearable fitness products have more limited battery capacity so the power consumption is even more of a design challenge.

Another important use case for indoor positioning is asset tracking. Typically, a small battery-powered device is placed in or on an important asset so that it can be located and tracked when necessary. Battery life for such systems is extremely important, as is the ability to locate the asset in any environment, however continuous position updates are not needed. A typical feature of asset tracking systems is the ability to set a geofence boundary that is used for generating alerts. The asset position is periodically compared with the geofence and if a rule is violated an alert is generated. For example, a user may wish to know when a particular asset leaves a geographic area. The positioning system needs to determine position periodically and compare it with the relevant geofence. If the position is outside the geofence, an alert is sent to the user.

13.5.2 Mobile Device Augmentation Sensors

13.5.2.1 Multi-GNSS Receiver

Modern GNSS receivers have positioning algorithms that combine range measurements from all visible GNSS satellites including QZSS and SBAS. High sensitivity has improved performance indoors. However, the utility of reception sensitivities below –165 dBm has been found to have limited value for all but static cases, due to the very long integration times required to make reliable measurements. Increasing the number of independent range measurements by using multiple constellations helps improve indoor positioning coverage and accuracy marginally.

Multipath delays for indoor environments are typically much shorter than outdoors, and hence conventional mitigation methods cannot be applied without a very wide RF bandwidth. The shorter delays therefore result in lower signal levels due to phase cancellations and pseudorange bias errors. While the advantage of augmenting GPS measurements with GLONASS is typically 20% to 40% improvement in position accuracy in urban canyon environments, it shrinks to only 7% to 15% indoors [96] due to the lower signal strength and associated multipath effects. While the use of multiple constellations improves the accuracy and availability of the GNSS fixes, additional position sources are needed to achieve suitable availability and accuracy for continuous indoor positioning. There are diminishing returns when adding BeiDou and Galileo signals leaving many indoor areas out of reach of GNSS signals.

13.5.2.2 MEMS Pedestrian Dead Reckoning (PDR)

As was discussed in Section 13.3.2.1, inertial systems are effective at bridging coverage gaps in GNSS and also in smoothing the position output when GNSS signals are weak or noisy. MEMS devices have emerged in smart phones to support context detection for controlling screen rotation. Smart phones are typically equipped with a 3-axis accelerometer, a 3-axis gyro, and a 3-axis magnetometer. Some phone models now also include a barometric altimeter to sense changes in altitude. All of these sensors are smaller, lower cost, and lower performance than those typically found in vehicles, however they are still very useful for inertial positioning albeit with different approaches.

A land vehicle has somewhat constrained movements as the acceleration is governed by the fact that it is a large object and rolls on wheels in a plane. Mobile devices are not so constrained and so the inertial algorithms employed must account for this three-dimensional movement freedom. Also, the accuracy of the sensors is limiting in that the drift rate is so high that inertial results are only reliable for a few minutes before they need correction with absolute position or some other constraint.

In order to determine the optimum approach for utilizing the MEMS output for positioning, it is first useful to ascertain what the context of the movement is. In fact, the inertial sensors can be used to detect whether a user is stationary, walking or running, or if the user is climbing or descending stairs, an elevator or an escalator [98]. For these modes of motion, a pedestrian dead reckoning (PDR) approach can be used. Other modes of operation such as cycling, skateboarding, or riding some other wheeled vehicle would require a different integration approach more like the land vehicle scenario. Once the data is processed to determine the dynamic mode of the user, appropriate position constraints, algorithms, and motion parameters can be assigned.

The PDR algorithm is designed to detect individual steps, calibrate a step length, and track the direction and vertical displacement. The generalized navigation equation [7] can be written as:

$$\dot{\mathbf{v}}_e^n = \mathbf{C}_b^n \mathbf{f}^b - [2\boldsymbol{\omega}_{ie}^n + \boldsymbol{\omega}_{en}^n] \times \mathbf{v}_e^n + \mathbf{g}_l^n \tag{13.77}$$

where $\mathbf{v}_e^n$ is ground velocity in navigation frame, $\mathbf{C}_b^n$ is a direction cosine matrix relating body reference frame to navigation frame, $\mathbf{f}_b$ the is specific force, $\boldsymbol{\omega}_{ie}^n$ is the turn rate of the Earth, $\boldsymbol{\omega}_{en}^n$ is the body rate, and $\mathbf{g}_l^n$ is the local gravity vector expressed in navigation frame. This equation (in navigation frame) relates the ground speed of an object to measured specific force and measured body rate. The generalized navigation equation when integrated twice, transforms from the acceleration of the platform into position represented in North and East reference frame, results in:

$$\begin{aligned} E(t) &= E(0) + \int_0^t s(t)\sin(\psi(t))dt \\ N(t) &= N(0) + \int_0^t s(t)\cos(\psi(t))dt \end{aligned} \tag{13.78}$$

where $s(t)$ is displacement and $\psi(t)$ is heading. In the case of pedestrian motion, velocity and heading can be assumed to be constant during the interval when a step is taken. With this assumption, the integral form of (13.78) can be rewritten as a difference equation with piece-wise linear approximation.

$$\begin{aligned} E_t &= E_{t-1} + \hat{s}_{[t-1,t]} \sin \psi_{t-1} \\ N_t &= N_{t-1} + \hat{s}_{[t-1,t]} \cos \psi_{t-1} \end{aligned} \tag{13.79}$$

This equation describes a method of dead reckoning, which is based on step counting rather than integration of acceleration and angular rate. This PDR process consists of three important components: the previously known absolute position of the user at time $t - 1$ (E_{t-1}, N_{t-1}), the stride length or distance traveled by the user since time $t - 1$ ($\hat{s}_{[t-1,t]}$), and the user's heading (ψ) since time $t - 1$. The coordinates (E_t, N_t) of a new position with respect to a previously known position (E_{t-1}, N_{t-1}) can be computed as shown in (13.79). The position initialization of the PDR process can be accomplished using any or a combination of absolute positioning technologies including GNSS, Wi-Fi and other RF positioning methods.

PDR solutions can be implemented with floor level awareness by utilizing the gyroscope and accelerometer data to determine change in height. A barometric altimeter increases the accuracy and reliability of height determination, but not all mobile devices have this sensor. To determine floor level, altitude is combined with floor level elevations in a building map. Automatic floor level detection can be done by monitoring PDR and the altimeter to detect elevation changes that approximate the elevation differences of the floors. This technique requires actual or estimated elevation difference of the floors and an initial level assumption either from the level of entry or from user input.

Performance of PDR algorithms is dependent on obtaining calibrated MEMS inertial sensor data continuously. Calibration of sensors is accomplished through collecting and processing sensor data for user motion of device in Earth's gravity and magnetic field. Accelerometer and gyroscope calibration logic utilize the knowledge of device being in a stationary condition. Magnetic sensor calibration logic requires that various axes of the sensor are exposed to Earth's magnetic field vector at the user location. This can be done by requesting that the user hold the mobile device in various orientations for calibration, however this is not desirable. Normal use of a mobile device would result in rotations in various Euler planes thereby applying Earth's magnetic field to various axes of magnetic sensor. With the given time and location estimate, the Earth's magnetic field parameters are computed using the World Magnetic Model [99]. Earth's magnetic field parameters are also used to detect occurrences of magnetic disturbances. Magnetic sensor measurements are de-weighted for the PDR process during such magnetic disturbances.

The essential logic components which impact the performance of PDR positioning system are: calibration of sensors, step detection, determination of walking direction, positioning fusion logic, and orientation of the mobile device while walking. Typical phone users will have the phone in a pocket, in a belt clip, in a purse or bag, in their hands looking at it, or up to their ear in a conversation. The PDR algorithms need to be able to perform robustly in any of these orientations [100, 101].

With PDR, an absolute position can be propagated forward in time as a user moves on foot. Due to the error growth characteristics of MEMS used in mobile devices, the estimated path deviates from the actual path as a function of distance traveled from the last absolute position fix. The error growth is typically on the order of 10% of distance traveled, and is especially high in the presence of magnetic disturbances. This level of error growth makes MEMS PDR unsuitable as a sole positioning solution when moving indoors. Periodic absolute positioning updates are required to correct the path and to allow additional calibration.

13.5.2.3 Wi-Fi Positioning

Opportunistic positioning using observed Wi-Fi signals is a well-established method of absolute positioning in GNSS-denied environments. Most existing Wi-Fi Access Point (WAP) transmitters are not well suited to positioning using timing observations as there is not an encoded time stamp in the signal. Instead, mobile devices can use the Received Signal Strength Indicator (RSSI) from a WAP transmitter to estimate range to the transmitter location. The broadcast Basic Service Set IDentifiers (BSSIDs) or Media Access Control (MAC) Address is a unique identifier for each WAP transmitter so that the mobile device can know the source of each signal. Note that the BSSID is broadcast on an open signal and does not require any authentication to obtain. If the location of each WAP transmitter in range is known, then the location of the mobile device can be trilaterated if range estimates to at least 3 WAP transmitters can be observed.

Signal strength information is by its nature asymmetric. A strong observation of a Wi-Fi Access Point (WAP) indicates that one is near it, but it is not safe to infer from a weak observation that you are far away. This is because weak observations may be due to, for example, occlusion, fading, or antenna orientation. This means that the performance of Wi-Fi positioning varies considerably with location and time, especially in areas with many pedestrians.

Methods and standards are emerging to support round-trip travel time (RTT) measurements between a mobile device and a WAP transmitter which will improve the range accuracy to the 1–4-m range [102, 103]. The IEEE 802.11mc FTM (Fine Time Measurement) standard enables RTT measurements between enabled WAPs and mobile devices. Chipsets and products that support this standard are emerging now and it is expected that over the next 3 to 5 years as WAPs are replaced in buildings, the RTT will be possible in most venues.

There are several limitations to Wi-Fi positioning. First, since it is opportunistic, there is no guarantee of consistent performance or coverage. Fortunately, WAP density is typically highest in just the areas where Wi-Fi positioning is most needed, namely, deep indoors and in dense urban areas where there are lots of visitors. Second, walls, objects, and even people in the environment have a large effect on the transmitted signals causing variation in the received signals strengths which then affects estimated range. Hence, typical range measurements using RSSI can be 10–20m in error [102]. Third, the location of the WAP transmitters is not controlled on installation and not known a priori and hence must be learned through surveying or observations so that a database of WAP transmitter locations can be created for positioning. Fourth, there is also no guarantee that WAPs will remain in the same locations. WAPs may be attached to mobile devices or WAP equipment

may simply be moved. This leads to a requirement for the database of WAP locations to be dynamically monitored and continuously improved.

To learn the location of WAP transmitters, there are three distinct approaches: WAP surveying, WAP fingerprinting, and crowdsourcing. A manual WAP survey is a process of having a field tester visit a venue and walk through the building stopping at identifiable locations within the building. At each identifiable location, the tester indicates on a map where they are located at that time. A stamped measurement is taken gathering the RSSI data to each available WAP. This data is recorded and then used to estimate the location of each WAP transmitter. For positioning mobile devices, the coordinates of each WAP transmitter in range is used along with range estimates from the RSSI in order to trilaterate a position. This WAP location method is effective and can deliver reasonable accuracy within the venue; however, the disadvantages are the requirement to send a tester to the field to do the survey initially and also periodically to update the survey since WAP installations change over time.

Fingerprinting is a process that also requires a manual survey to be completed. In this technique, at each survey location, the RSSI of each WAP transmitter is also captured and recorded. A database of RSSI values at each location is created and is later used to compare with the data captured by mobile devices in the area; hence, the fingerprint of signal strengths is matched to the database so that location can be determined. In this technique, the location of the WAP transmitter is not estimated; rather the observed RSSI is stored at various locations [104]. Fingerprinting techniques typically have a higher demand on data stored in the database and also data transferred between the mobile device and server. The field of fingerprinting continues to advance with new techniques for handling the temporal variations of signals to improve accuracy [105]. Like the WAP location method, fingerprinting also has the requirement of an initial and periodic manual survey to create and maintain the database used for device positioning.

The crowdsourcing technique is a method of determining WAP transmitter locations without the need of manual surveying. In this type of system, anonymous data from user devices is sent to a server for learning. This data consists of the BSSIDs and RSSI measurements as well as the best estimated position and associated accuracy. When the user device is outdoors, position is determined using GNSS. As the user moves indoors, MEMS is used to propagate the position for a short period of time in order to maintain a position estimate for WAP learning to take place [96, 106]. This approach has the advantage of not requiring any costly or time consuming a priori survey work in the field. However, the initial performance of the system has limited coverage since it requires some time for the WAP transmitter locations to be learned. The time it takes for learning is a function of the number of people using the system and providing learning data. A manual survey has the advantage of higher initial accuracy, but the crowdsourcing method has the advantage of faster correction when WAP transmitters are moved or replaced. Some venue owners may be willing to fund a manual survey for their venue in order to provide highest possible accuracy to their visitors. However, there are many locations where it is not feasible to have a manual survey performed and without crowdsourcing, these areas would be left as a coverage gap.

13.5.2.4 Bluetooth and Other RF Transmitters

Bluetooth Low Energy (BLE) transmitters are emerging in venues for use as proximity beacons for advertising and consumer interaction. These BLE transmitters can be used in the same manner as Wi-Fi transmitters for positioning. The locations of the transmitters have to be determined and the same techniques of manual learning or crowdsourcing are effective. BLE transmitters are not ubiquitous in indoor spaces like Wi-Fi transmitters are, but BLE adoption is growing. BLE has a typical range of 50m, which is not as far as typical Wi-Fi, but long enough that it is well suited for positioning once sufficient transmitters are in use in an area.

Other wireless transmitters could also be used for positioning in the same manner, but Wi-Fi has the advantage of having the greatest in-building installed base and 100% penetration in mobile phones and tablets. BLE is following a similar path towards high adoption on mobile devices and is gaining wider installation base in various venues. The only other wireless technology that is in a large percentage of mobile devices is near-field communication (NFC). NFC is very low cost to use as a beacon or tag, as the transmitter does not require a power source. However, the range is very short at about 20 cm, so it is better suited to proximity-based positioning than trilateration. This method requires knowing the position of the NFC beacon and then when a device is near enough to transmit, the position can be determined with an uncertainty equivalent to the range of the transmitter. Unfortunately, when you are not in range of an NFC reader, there is no knowing what your position is.

Existing cellular transmitters can be used to determine position indoors using techniques such as AFLT, U-TDOA, and O-TDOA. The accuracy of these methods depends on the density of cell towers in the area and ranges from several tens of meters to several hundred meters. This is not accurate enough to support indoor navigation and locating to the store level in a mall. New small cell transmitters offer an opportunity to improve the accuracy, but require installation of a few transmitters in the vicinity.

Another approach to using wireless transmitters for indoor positioning is to create and deploy a whole new network of transmitters. NextNav is building such a wide-area network of ground-based transmitters called a Metropolitan Beacon System (MBS) in the 902–928-MHz band. Current mobile phones do not have a radio that can receive NextNav signals; however, Release 13 of the 3GPP includes messaging specifications to support MBS location technology. This new standard is aimed at improving emergency response services to persons while indoors by providing reliable positioning across an entire metro area. The transmitted signal penetrates buildings and can be used to determine horizontal position to 20-m accuracy and vertical to 2m. If adopted by the FCC for use in E-911 calls, this technology could become a standard for phones sold in the United States [107].

13.5.2.5 Other Positioning Methods

With the emergence of smart phones with application development capabilities, virtually every sensor available on the phones has been used to aid positioning. The magnetometer is used to capture the magnetic field the phone is in [108, 109]. The camera sensor is used to capture images for image recognition [110] or light pattern

matching [111]. The microphone is used to listen to an audio signature of the environment [112]. Each of these techniques can be used for positioning by correlating the phone data captured against a database of previously recorded measurements on the assumption that each position has a unique signature of measured values.

To capture the database of recorded measurements, a fingerprinting process is used to record the magnetic data, image, light, or audio data with the device throughout the target area. The data is processed to generate a continuous model that matches the submitted data points. For positioning, the database is searched to find the location that best matches the snapshot of data recorded by the device. This process is quite similar to the fingerprinting process described for Wi-Fi positioning, except that the type of data recorded is magnetic fields, imagery, light, or sounds. These techniques have been shown to work effectively for positioning; however, they have a high data transfer load, they all require a manual survey to create the initial fingerprint data model, and in the case of image and light recognition, the phone has to be outside of a pocket or purse so that the camera has a view of the surroundings. These techniques are well suited for projects with limited geographic reach or limited number of venues so that the manual survey step is practical.

13.5.2.6 Indoor Map Databases

As was seen in Section 13.3.2.2, although a digital map database is not a standalone positioning system, a map that is spatially and positionally accurate can be a strong aid to a positioning system. Indoor maps provide the same opportunity to enhance the positioning system performance through map aiding.

Several of the established digital mapping companies have started to publish maps of indoor venues such as shopping malls, airports, convention centers, and train stations. In the early 2010s, Google and HERE started adding indoor venues to their digital road maps. Another company, Micello, was formed specifically to produce indoor maps concentrating on malls and other popular venues throughout the world. More recently, Apple and TomTom have started to produce indoor digital maps to add to their digital road maps.

Indoor venue maps have similar characteristics as digital road maps in that they have attribution to enable search, routing, and map display. As mentioned, these maps should be positionally accurate to correlate with the rest of the world, and be spatially accurate with properly indicated walls, stairs, escalators, elevators and entryways. Each service and place of business should be attributed for searching. For walking instructions, it is very important that each level be modeled separately and the connections between levels be attributed.

Typically, an indoor map is created by the venue owner first providing a printed map or digital blueprint of the building. These maps are then digitized and calibrated through site survey or comparison with existing maps and air photos. Map rendering tools provided by the map companies or other third parties such as eeGeo or VisioGlobe enable the maps to be displayed on computers or mobile devices. Overhead views and three-dimensional perspective views allow users to visualize the environment and orient themselves to find their way. A rendering of Westfield Valley Fair Mall in San Jose, California, is shown in Figure 13.44. This map is published by Micello and rendered using eeGeo tools and is provided

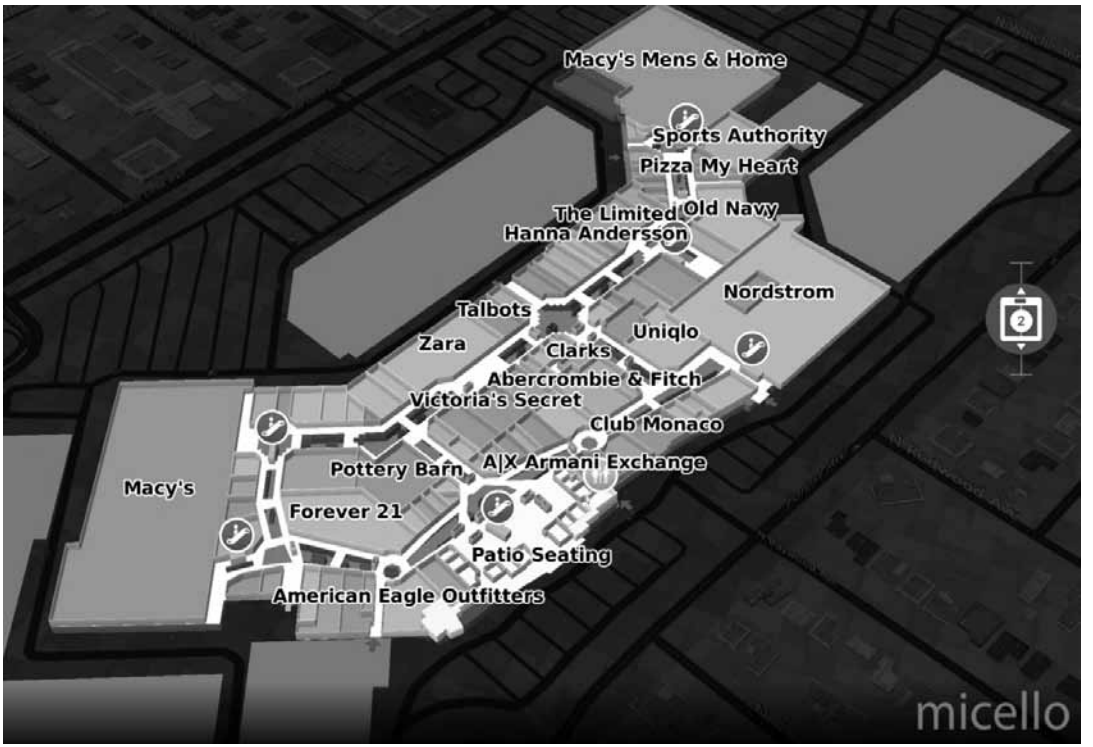


Figure 13.44 Westfield Valley Fair Mall three-dimensional perspective view of indoor map. (Courtesy Micello, eeGeo, Westfield.)

courtesy of Micello and eeGeo, and used with permission from Westfield. The indoor geometry is shown in a perspective view, with a floor level indicator/selector on the right and with stores, escalators, and restrooms clearly marked.

Aside from search, visualization, and walking directions, digital indoor maps serve an important role in positioning. Similar to digital road maps, an indoor map can be used for map matching by using the structure geometry to determine where it is possible for a person to be and what paths are possible to traverse. Indoor maps do not constrain a solution to the extent that a digital road map does because pedestrians are able to move around much more freely than vehicles on a road network. It is possible to attribute hallways and corridors, but people are able to stop, change direction, and resume moving without restriction. Therefore, map matching must accommodate this by allowing the displayed position more freedom to move about. Map-aiding is also possible with indoor maps, but the constraints are similarly loosened. One area that can be exploited for positioning assistance is the presence of elevators, stairs, and escalators. As mentioned in Section 13.5.1, absolute position fixes are required indoors to help initiate and reinitiate user position. When the motion sensors are able to determine a change in elevation, it is also possible to determine if the user mode of movement is riding an escalator, riding an elevator, or using a stairway. If the motion sensor processing determines that the user is riding up an escalator for example, the map can be used to identify if an up escalator is nearby. If there is only one up escalator within the region of uncertainty of the last known position, then it is highly likely that the user is in fact using that escalator and the positioning system can reinitialize the position to the coordinates

of the top of the escalator. In this manner, the indoor position can be periodically calibrated to known locations in the mall through modal detection and probability analysis.

13.5.3 Mobile Device Sensor Integration

GNSS, Wi-Fi, MEMS PDR, and other positioning solutions offer varying levels of accuracy, coverage, and reliability. As seen in Sections 13.2 and 13.3.3, a Kalman filter can be used to combine all of these position inputs to determine a single best estimate of position and confidence to the user.

This technique is known as sensor fusion in the mobile device context and the major components are shown in Figure 13.45 [96]. A fusion filter takes as input absolute positions from GNSS, Wi-Fi, and/or other solutions and also relative positioning data derived from the MEMS PDR subsystem. The positioning data is then fused together continuously to determine the best estimate of position even when an absolute position cannot be computed.

In order to determine how to weight and smooth the different inputs, it is crucial that the individual input technologies provide reliable estimates of their confidence and correlation. As an example, it was mentioned earlier that the quality of Wi-Fi positioning is variable and is best when strong WAP signals are received. A high quality Wi-Fi position, signified by a high confidence value, will cause the fusion filter to be strongly biased towards the Wi-Fi position solution. When the Wi-Fi position quality subsequently deteriorates, it is reflected in a lower position confidence and hence the fusion filter down-weights the influence of Wi-Fi. In turn, this allows dominance of the MEMS PDR input until another sufficiently high quality absolute position allows the filter to correct. The net effect of this behavior is that the MEMS bridges the position output smoothly between high-quality absolute position fixes and to a first approximation, any low-grade information is ignored. Another benefit is that the MEMS smooths the individual Wi-Fi positions, which can be noisy due to the considerable variation in the received WAP signals. Ultimately, the aim of the fusion filter is to provide a continuous position trajectory and hence a more satisfying user experience.

Another function of the fusion filter is to transition smoothly from outdoors where GNSS dominates, to indoors where Wi-Fi and MEMS PDR dominate, and vice versa. A properly tuned fusion filter will be able to handle this transition

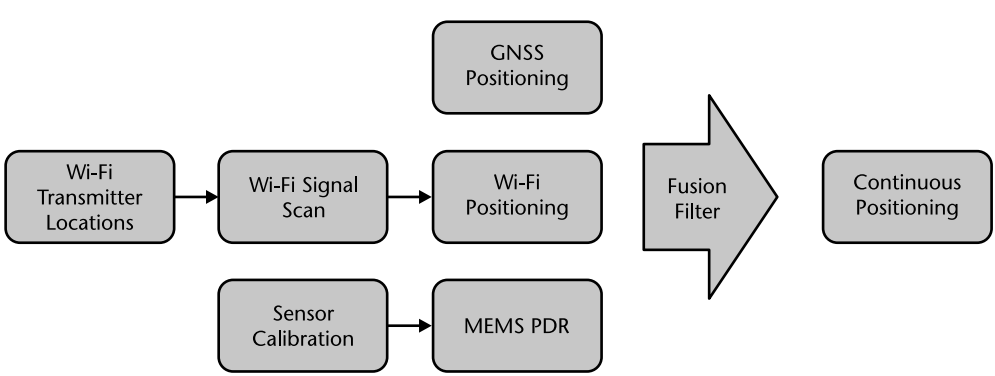


Figure 13.45 Major components of sensor fusion.

automatically, as the GNSS accuracy drops the Wi-Fi positioning becomes the more reliable absolute position source and is weighted accordingly. Conversely, the Wi-Fi position accuracy will typically decrease outdoors and the GNSS position will gradually dominate the solution.

An important consideration in the design of a hybrid positioning system with a fusion filter is power consumption. To get the overall best positioning performance all of the available positioning systems should be on all the time to get all possible positioning solutions for use in the fusion filter; however, this would consume far too much power. To balance battery life with positioning accuracy, the system should be designed to leverage the motion sensors to determine when there is movement and hence when a position update may be necessary. If GNSS is powered on and no position is obtained, the GNSS should be powered down for some time so as to not waste power searching for signals that are not there.

Sensor Fusion Performance

In [96], a series of tests were carried out by CSR Technology at Tokyo Station in Tokyo, Japan, to assess the performance of using a fused solution of Wi-Fi positioning and MEMS PDR for indoor positioning. The tests were done on the B1F level in the shopping area adjacent to the station. This area is two levels below the tracks and is below ground level. There are no windows and there is no GNSS reception. The environment also has a lot of magnetic anomalies due to tracks, trains, elevators, and escalators, and it also has many people in motion, which affects Wi-Fi signal transmission.

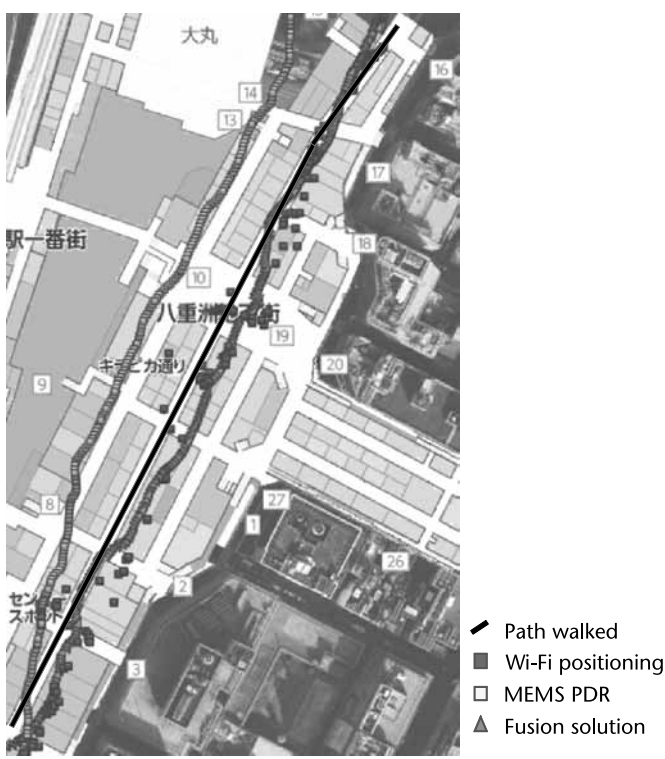


Figure 13.46 Indoor positioning test in Tokyo Station. (Courtesy CSR.)

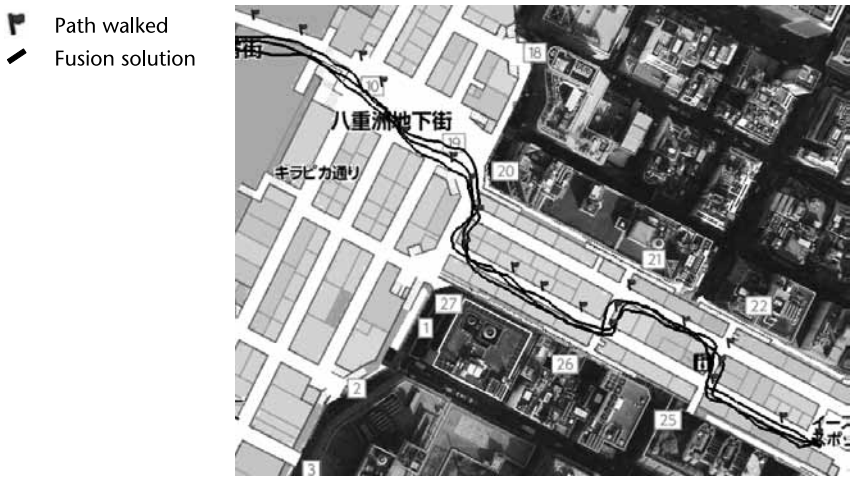


Figure 13.47 Tokyo Station test repeatability. (Courtesy CSR.)

Figure 13.46 shows an indoor map of the station superimposed on the Google Earth image of the area. The narrow aisles in the map are about 5m wide. The map is used for presenting results only; it was not used to do map-aiding or map-matching. Prior to this test, the area was calibrated with a manual process of marking identifiable points on the map and collecting the Wi-Fi signal information to construct a database of WAP transmitter locations.

The route walked is shown by the straight line, starting in the lower left corner and finishing near the top right. Each computed Wi-Fi position is shown as a dark square, the series of light squares is the MEMS PDR solution, and the triangles show the fusion solution that is combining Wi-Fi and PDR. The Wi-Fi position is not available every second and at times has discontinuities of several meters due to the signal variability as discussed previously. The PDR solution shows a gradual drift that is more than 25m off track in places due to the drift of the MEMS sensors. The fusion solution combines the noisy absolute positions from the Wi-Fi with the smooth but drifting PDR path and the result is a smooth continuous output that has a maximum cross-track error of about 7m.

Figure 13.47 shows another path through the corridors with several turns that takes about 7 minutes to walk. The test was repeated three times with a phone reset before each walk to clear the positioning system. The fusion solution shows each of the turns correctly and in this case, the maximum cross-track error is about 5m. The results of the three trials agree closely showing high repeatability between test runs.

References

[1] Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, Vol. 82, 1960, pp. 35–45.

[2] Hemesath, N. B., et al., "Anti-Jamming Characteristics of GPS/GDM," Collins Division of Rockwell, Cedar Rapids, IO, *Proc. of the National Telecommunications Conference*, Dallas, TX, November 1976.

[3] Greenspan, R. L., "Inertial Navigation Technology from 1970-1995," *NAVIGATION: Journal of the Institute of Navigation*, Vol. 42, Spring 1995.

[4] U.S. Patent Number 5,983,160, "Increase Jamming Immunity by Optimizing Processing Gain for GPS/INS Systems," November 9, 1999.

[5] Ohlmeyer, E. J., "Analysis of an Ultra-Tightly Coupled GPS/INS System in Jamming," *Proc. of IEEE/ION PLANS 2006,* San Diego, CA, April 2006, pp. 44–53.

[6] Lawrence, A., *Modern Inertial Technology: Navigation, Guidance, and Control*, New York: Springer-Verlag, 1998.

[7] Titerton, D., and J. Weston, *Strapdown Inertial Navigation Technology*, 2nd ed., Stevenage, UK: The Institution of Electrical Engineers, 2004.

[8] Britting, K. R., *Inertial Navigation System Analysis*, New York: Wiley-Interscience, 1971.

[9] Pinson, J. C., "Inertial Guidance for Cruise Vehicles," in *Guidance and Control of Aerospace Vehicles*, C. T. Leondes, (ed.), New York: McGraw-Hill, 1963.

[10] Rogers, R. M., *Applied Mathematics in Integrated Navigation Systems*, 2nd ed., Reston, VA: AIAA Education Series, 2003.

[11] Gelb, A., et al., *Applied Optimal Estimation*, Cambridge, MA: MIT Press, 1992.

[12] Carlson, N. A., and M. P. Beraducci, "Federated Kalman Filter Simulation Results," *Journal of the Institute of Navigation*, Fall 1994.

[13] Vallot, L., "Vibration Compensation for Sensors," U.S 6,448,996, December 24, 2002.

[14] Brown, R. G., and P. Hwang, *Introduction to Signal Processing and Applied Kalman Filtering*, New York: John Wiley & Sons, 1992.

[15] Brown, R. G., and P. W. McBurney, "Proper Treatment of the DR Measurement in Integrated GPS/INS," *Proc. of the National Technical Meeting of the Institute of Navigation*, January 1987.

[16] Thornton, C. L., and G. J. Bierman, *UDU^T Covariance Factorization for Kalman Filtering*, New York: Academic Press, 1980.

[17] Manry, C. W., et al., "Advanced Mini Array Antenna Design Using High Fidelity Computer Modeling and Simulations," *Proc. of The Institute of Navigation ION GPS-2000*, Salt Lake City, UT, September 2000, pp. 2485–2490.

[18] Tseng, H. -W., et al., "Test Results of a Dual Frequency (L1/L2) Small Controlled Reception Pattern Antenna," *Proc. of The Institute of Navigation ION GPS-2002*, San Diego, CA, January 2002.

[19] Kunysz, W., "Advanced Pinwheel - Compact Controlled Reception Pattern Antenna (AP-CRPA) Designed for Interference and Multipath Mitigation," *Proc. of The Institute of Navigation ION GPS-2002*, Portland, OR, September 2002.

[20] Klemm, R., *Principles of Space Time Adaptive Processing*, London, U.K.: Institute of Engineering and Technology, 2006.

[21] Cox, D. B., "Integration of GPS with Inertial Navigation Systems," *Global Positioning System: Papers Published in NAVIGATION, Volume I*, Fairfax, VA, Institute of Navigation, 1980.

[22] Carroll, R. W., et al., "Velocity Aiding of Non-Coherent GPS Receiver," *Proc. of the 1977 National Aerospace Conference*, Dayton, OH, 1977.

[23] Widnall, W. S., "Alternate Approaches for Stable Rate Aiding of Jamming Resistant GPS Receivers," *NAECON Proceedings*, Dayton, OH, 1979.

[24] Copps, E. M., et al., "Optimal Processing of GPS Signals," *NAVIGATION: Journal of The Institute of Navigation*, Fall 1980.

[25] Sennott, J. W., et al., "Navigation Receiver with Coupled Signal Tracking Channels," U.S. Patent 5,343,209, May 1992.

[26] Leimer, D., "Receiver Phase Noise Mitigation," U.S. Patent 6,081,228, SiRF Technology, September 1998.

[27] Beser, J., et.al., "TRUNAV: A Low Cost Guidance/Navigation Unit Integrating a SAASM Based GPS and MEMS in a Deeply Coupled Mechanization," *Proc. of The Institute of Navigation ION-GPS 2002*, Portland, OR, September 24–27, 2002.

[28] Abbott, T., et al., *Ultra-tight GPS/IMU Coupling Method*, The Aerospace Corporation, TOR-2001(1590)-0846e, El Segundo, CA, April 10, 2001.

[29] Gautier, J. D., et.al., "Using the GPS/INS Generalized Evaluation Tool (GIGET) for the Comparison of Loosely Coupled, Tightly Coupled, and Ultra-Tightly Coupled Integrated Navigation Systems," *Proc. of The Institute of Navigation 59th Annual Meeting*, Albuquerque, NM, June 2003.

[30] Gelb, A., et al., *Multiple Input Describing Functions and Nonlinear System Design*, New York: McGraw-Hill, 1968.

[31] Yazdi, N., et al., "Micromachined Inertial Sensors," *Proc. of the IEEE*, Vol. 86, No. 8, August 1998, pp. 1640–1659.

[32] Peng, K., "A Vector-Based Gyro-Free Inertial Navigation System by Integrating Existing Accelerometer Network in a Passenger Vehicle," *Proc. IEEE Position Location and Navigation Symposium* (*PLANS*), Monterey, CA, April 26–29, 2004, pp. 234–242.

[33] Schuler, A. R., "Measuring Rotational Motion with Linear Accelerometers," *IEEE Trans. on Aerospace and Electronic Systems*, Vol. AES-3, 1967, pp. 465–471.

[34] Weinburg, H., "MEMS Sensors Are Driving the Automotive Industry," *Sensors*, Vol. 19, No. 2, February 2002, pp. 36–41.

[35] Mostov, K. S., A. A. Soloviev, and T. J. Koo, "Accelerometer Based Gyro-Free Multi-Sensor Generic Inertial Device for Automotive Applications," *Proc. IEEE Conference on Intelligent Transportation Systems*, Boston, MA, November 1997, pp. 1047–1052.

[36] Chen, T. H., "Gyroscope Free Strapdown Inertial Measurement Unit by Six Linear Accelerometers," *Journal of Guidance, Control, and Dynamics*, Vol. 17, No. 2, 1994, pp. 286–290.

[37] Stephen, J., "Development of a Multi-Sensor GNSS Based Vehicle Navigation System," M.Sc. Thesis, UCGE Report No. 20140, Department of Geomatics Engineering, University of Calgary, Canada, August 2000.

[38] U.S. Department of Defense, "Micromachined System Opportunities," Department of Defense Dual-Use Technology Industrial Assessment, 1995.

[39] Helsel, M., et al., "A Navigation Grade Micro-Machined Silicon Accelerometer," *Proc. IEEE Position Location and Navigation Symposium* (*PLANS*), Las Vegas, NV, April 11–15, 1994, pp. 51–58.

[40] Lemkin, M. A., et al., "A Three Axis Surface Micromachined Sigma-Delta Accelerometer," *ISSCC Digest of Technical Papers*, February 1997.

[41] Gustafson, D., et al., "A Micromechanical INS/GPS System for Guided Projectiles," *Proc. ION 51st Annual Meeting,* Colorado Springs, CO, June 5–7, 1995, pp. 439–444.

[42] Warren, K., "High Performance Silicon Accelerometers with Charge Controlled Rebalance Electronics," *Proc. IEEE Position Location and Navigation Symposium* (*PLANS*), Atlanta, GA, April 22–26, 1996, pp. 27–30.

[43] Le Treon, O., et al., "The VIA Vibrating Beam Accelerometer: Concept and Performance," *Proc. IEEE Position Location and Navigation Symposium* (*PLANS*), Palm Springs, CA, April 20–23, 1998, pp. 25–29.

[44] Hulsing, R., "MEMS Inertial Rate and Acceleration Sensor," *Proc. of The Institute of Navigation National Technical Meeting*, Long Beach, CA, January 1998, pp. 353–360.

[45] Clark, W., R. Howe, and R. Horowitz, "Surface Micromachined Z-Axis Vibratory Rate Gyroscope," *Proc. Solid-State Sensors and Actuators Workshop*, Hilton Head, SC, June 13–16, 1996, pp. 283–287.

[46] Kourepenis, A., et al., "Performance of MEMS Inertial Sensors," *Proc. IEEE Position Location and Navigation Symposium* (*PLANS*), Palm Springs, CA, April 20–23, 1998, pp. 1–8.

[47] Barbour, N., "Operational Status of Inertial," *Proc. of The Institute of Navigation National Technical Meeting*, Santa Monica, CA, January 22–24, 1996, pp. 7–15.

[48] Park, M., "Error Analysis and Stochastic Modeling of MEMS Based Inertial Sensors for Land Vehicle Applications," M.Sc. Thesis, UCGE Report No. 20194, Department of Geomatics Engineering, University of Calgary, April 2004.

[49] Xu, Y., "The European Digital Road Map MultiMap and ITS Applications," *International Archives of Photogrammetry and Remote Sensing*, Vol. XXXI, Part B4, Vienna, 1996, pp. 982–986.

[50] Kang, J. M., J. K. Park, and M. G. Kim, "Digital Mapping Using Aerial Digital Camera Imagery," *ISPRS Commission IV, WG IV/9*, pp. 1275–1278, 2008.

[51] Biagioni, J., and J. Eriksson, "Inferring Road Maps from Global Positioning System Traces," *Transportation Research Record: Journal of the Transportation Research Board*, No. 2291, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 61–71.

[52] Bullock, J. B., and E. J. Krakiwsky, "Analysis of the Use of Digital Road Maps in Vehicle Navigation," *Proc. IEEE Position Location and Navigation Symposium* (*PLANS*), Las Vegas, NV, April 11–15, 1994, pp. 494–501.

[53] Zavoli, W. B., and S. K. Honey, "Map Matching Augmented Dead Reckoning," *Proc. IEEE Position Location and Navigation Symposium* (*PLANS*), Las Vegas, NV, 1986, pp. 359–362.

[54] Honey, S. K., et al., "Vehicle Navigation System and Method," U.S. Patent 4,796,191, Etak Incorporated, January 3, 1989.

[55] Mathis, D. L., et al., "Combined Relative and Absolute Positioning Method and Apparatus," U.S. Patent 5,311,195, Etak Incorporated, May 10, 1994.

[56] French, R. L., "Map Matching Origins, Approaches and Applications." *Proc. Land Vehicle Navigation,* Verlag TUV Rheinland GmbH., Koln, Germany, July 4–7, 1989, pp. 91–116.

[57] Harris, C. B., "Prototype for A Land Based Automatic Vehicle Location and Navigation System," M.Sc. Thesis, Department of Geomatics Engineering, University of Calgary, Canada, 1989.

[58] Bullock, J. B., "A Prototype Portable Vehicle Navigation System Utilizing Map Aided GPS," M.Sc. Thesis, Department of Geomatics Engineering, University of Calgary, Canada, 1995.

[59] Ribbens, W. B., "Understanding Automotive Electronics," *SAMS*, 1992, pp. 138–143.

[60] Zavoli, W. B., et al., "Method and Apparatus for Measuring Relative Heading Changes in a Vehicular Onboard Navigation System," U.S. Patent 4,788,645, Etak Incorporated, November 29, 1988.

[61] Carlson, C. R., J. C. Gerdes, and J. D. Powell, "Error Sources When Land Vehicle Dead Reckoning with Differential Wheelspeeds," *NAVIGATION: Journal of The Institute of Navigation*, Vol. 51, No. 1, Spring 2004, pp. 13–27.

[62] Honeywell silicon pressure sensors, http:content.honeywell.com/sensing/prodinfo/pressure.

[63] Vannucci, G., "Inclusion of Atmospheric and Barometric Pressure Information for Improved Altitude Determination," *TIA TR45 Cellular Network Standards Proposal TIA 45:1.1.1 LocTaskG*, March 2000.

[64] Wald, M., "An Automobile Option for Self-Navigating Car," *New York Times*, January 5, 1995.

[65] Geier, G. J., et al., "Integration of GPS with Dead Reckoning for Vehicle Tracking Applications," *Proc. 49th Annual Meeting of the Institute of Navigation*, Cambridge, MA, June 21–23, 1993, pp. 75–82.

[66] Van Diggelen, F., *A-GPS: Assisted GPS, GNSS, and SBAS*, Norwood, MA: Artech House, 2009.

[67] Monteith, K. A., "Wireless E911: Regulatory Framework, Current Status and Beyond," *IBC Mobile Location Services Conference*, McLean, VA, April 2001.

[68] Taylor, R. E., et al., "Navigation System and Method," U.S. Patent 4,445,118, May 1981.

[69] *RTCM Recommended Standards for Differential GPS Service*, Version 2.0, January 1990.

[70] Brown, A. K., et al., "Vehicle Tracking System Employing GPS Satellites," U.S. Patent 5,225,842, NAVSYS Corporation, May 1991.

[71] *Motorola EAGLE GPS Receiver Users' Manual*, January 1986.

[72] Bryant, R.C., Australian patent number AU-B-634587, "Position Reporting System," Auspace Limited, November 1989.

[73] "White Sands Missile Range (WSMR) Interface Control Document (ICD)," ICD 3680090, April 28, 1994.

[74] *Third Further Notice of Proposed Rulemaking, In the Matter of Wireless E911 Location Accuracy Requirements*, PS Docket No. 07-114, Federal Communications Commission, Washington D.C., February 20, 2014.

[75] The Communications Security, Reliability and Interoperability Council –III, Working Group III (CSRIC-III), "Leveraging LBS and Emerging Location Technologies for Indoor E911," March 14. 2013.

[76] Proctor, A., "Expert Advice: Taking Up Positions – Galileo and E112," *GPS World Magazine*, March 31, 2015.

[77] "Using Mobile Phone GNSS Positioning for 112 Emergency Calls," *PTOLEMUS Consulting Group presentation to public hearing by the European Commission*, Brussels, May 7, 2014.

[78] Vogel, W. J., G. W. Torrance, and N. Kleiner, "Measurement of Propagation Loss into Cars on Satellite Paths at L-Band," *Proc. of EMPS '96*, Rome, Italy, October 1996.

[79] Vogel, W. J., and N. Kleiner, "Propagation Measurements for Satellite Services into Buildings," *European Mobile/Personal Satcoms Conference*, Rome, Italy, October 1996.

[80] Vogel, W. J., "Satellite Diversity for Personal Satellite Communications—Modeling and Measurements," *10th International Conference on Antennas and Propagation*, Edinburgh, U.K., April 14–17, 1997.

[81] Vogel, W. J., and R. Akturan, "Elevation Angle Dependence of Fading for Satellite PCS in Urban Areas," *Electronic Letters*, Vol. 31, No. 25, December 7, 1995.

[82] Vogel, W. J., and J. Goldhirsh, "Mobile Satellite System Fade Statistics for Shadowing and Multipath from Roadside Trees at UHF and L-Band," *IEEE Trans. on Antennas and Propagation*, Vol. 37, No. 4, April 1989.

[83] Vogel, W. J., and G. W. Torrence, "Propagation Measurements for Satellite Radio Reception Inside Buildings," *IEEE Trans. on Antennas and Propagation*, Vol. 43, No. 7, July 1993.

[84] "Update of Indoor Measurement Results," Phillips Contribution number R4-040285 3GPP TSG RAN WG4 (Radio) Meeting #31, Beijing, China, May 2004.

[85] "GPS Satellite Signal Strength Measurements in Indoor Environments," Motorola Contribution number R4-040310, TSG-RAN WG4 meeting #31, Beijing, China, May, 2004.

[86] "OET Bulletin 71 – Guidelines for Testing and Verifying the Accuracy of Wireless E-911 Systems," Federal Communications Commission, Washington D.C., March 2000.

[87] TIA Standard TIA-IS-916 – "Recommended Minimum Performance Specification for TIA/EIA/IS801-1 Spread Spectrum Mobile Stations," Telecommunications Industry Association, Arlington, VA, April 2002.

[88] Smith, C. A., et al., "Sensitivity of GPS Acquisition to Initial Data Uncertainties," *ION GPS Papers*, Vol. III, 1986.

[89] 3GPP2 Technical Specification, C.S0022-B, Version 3.0, Third Generation Partnership Project 2, "Position Determination Service for cdma2000 Spread Spectrum Systems," September 2014.

[90] 3GPP Technical Specification TS 144 031, Third Generation Partnership Project, "Digital Cellular Telecommunications Systems (Phase 2+), Location Services (LCS), Mobile Station

(MS), Serving Mobile Location Centre (SMLC), Radio Resource LCS Protocol (RRLP)," 3GPP TS 44.031 version 12.3.0 Release 12. July 2015

[91] 3GPP Technical Specification, TS 36.355, v12.4.0, Third Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); LTE Positioning Protocol (LPP) Release 12; March 2015. http://www.3gpp.org/DynaReport/36355.htm. October 2015.

[92] Open Mobile Alliance Technical Specification, OMA-TS-ULP-V3_0-20150126-D, "User Plane Location Protocol," Candidate Version 3.0, January 26, 2015. http://www.openmobilealliance.org/. October 2015.

[93] Ziedan, N., et al., "Unaided Acquisition of Weak GPS Signals Using Circular Correlation or Double Block Zero Padding," *IEEE PLANS 2004,* Monterey, CA, April 26–29, 2004.

[94] 3GPP Technical Specification TS 23.032, Third Generation Partnership Project; Technical Specification Group Core Network; Universal Geographic Area Description (GAD), Release 4, 2001.

[95] 3GPP Technical Specification TS 25.331, Third Generation Partnership Project; Technical Specification Group Radio Access Network; Radio Resource Control (RRC) Protocol Specification, Release 1999.

[96] Bullock, J. B., et al., "Continuous Indoor Positioning Using GNSS, Wi-Fi, and MEMS Dead Reckoning," *Proc. of ION GNSS 2012 Conference*, Nashville, TN, 2012.

[97] http://www.nextnav.com.

[98] Chowdhary, M., et al., "Context Detection for Improving Positioning Performance and Enhancing User Experience," *Proc. of ION GNSS 2009 Conference*, Savannah, GA, 2009, pp. 2072–2076.

[99] http://www.ngdc.noaa.gov/geomag/WMM/DoDWMM.shtml.

[100] Chowdhary, M., M. Jain, and R. K. Srivastava, "Test Results for Indoor Positioning Solution Using MEMS Sensor Enabled GPS Receiver," *Proc. of ION GNSS 2010 Conference*, Portland, OR, 2010, pp. 565–568.

[101] Chowdhary, M., et al., "Robust Attitude Estimation for Indoor Pedestrian Navigation Application Using MEMS Sensors," *Proc. of ION GNSS 2012 Conference*, Nashville, TN, 2012.

[102] Malkos, S., and A. Hazlett, "Enhanced WIFI Ranging with Round Trip Time (RTT) Measurements," *Proc. of the 27th International Technical Meeting of the ION Satellite Division, ION-GNSS+ 2014*, Tampa, FL, September 8–12, 2014.

[103] Bahillo, A., et al., "Distance Estimation Based on 802.11 RTS/CTS Mechanism for Indoor Localization," in *Advances in Vehicular Networking Technologies*, M. Almeida, (ed.), University of Valladolid, Spain: InTech, April 2011.

[104] Bahl, P., and V. N. Padmanabhan, "RADAR: An In-Building RF-Based User Location and Tracking System," *Proc. of IEEE 9th Annual Joint Conference of the IEEE Computer and Communications Societies*, Tel Aviv, Israel, March 26–30, 2000, pp. 775–784.

[105] Chen, L., et al., "An Improved Algorithm to Generate a Wi-Fi Fingerprint Database for Indoor Positioning," *Sensors*, 2013.

[106] WO/2011/077166, "Locating Electromagnetic Signal Sources," June 30, 2011.

[107] Meiyappan, S., A. Raghupathy, G. Pattabiraman, "Positioning in GPS Challenged Locations - The NextNav Terrestrial Positioning Constellation," *Proc. of the 26th International Technical Meeting of The Satellite Division of the Institute of Navigation* (*ION GNSS+ 2013*), Nashville, TN, September 2013, pp. 426–431.

[108] Namiot, D., "On Indoor Positioning," *International Journal of Open Information Technologies*, Vol. 3, No. 3, 2015.

[109] U.S. Patent Number 9,078,104, "Utilizing Magnetic Field Based Navigation," July 7, 2015.

[110] Kim, J., and H. Jun, "Vision-Based Location Positioning Using Augmented Reality for Indoor Navigation," *IEEE Trans. on Consumer Electronics*, Vol. 54, No. 3, 2008, pp. 954–962.

[111] Yang, S. H., et al., "Indoor Three-Dimensional Location Estimation Based on LED Visible Light Communication," *Electronics Letters*, Vol. 49, No. 1, 2013, pp. 54–56.

[112] Rossi, M., et al., "RoomSense: An Indoor Positioning System for Smartphones Using Active Sound Probing," *Proc. of the 4th Augmented Human International Conference*, March 2013, pp. 89–95.

CHAPTER 14

GNSS Markets and Applications

Len Jacobson

14.1 GNSS: A Complex Market Based on Enabling Technologies

14.1.1 Introduction

The only thing more difficult than describing the current GNSS market is predicting its future growth. Until there are fully deployed Galileo and BeiDou satellite constellations (circa 2020), the GNSS market will consist largely of the value of receivers and applications using GPS and GPS+GLONASS signals and various space-based and ground-based augmentations. BeiDou use is slowly gaining a small market share primarily in China but also in other Asian countries.

Today's GNSS receivers commonly access two or more constellations; GPS +GLONASS receivers predominate, but many thousands of BeiDou receivers are already appearing in China. A GLONASS+BeiDou receiver is also likely to appear in quantity soon. BeiDou+Galileo+GLONASS+GPS receiver capabilities are available now from numerous receiver manufacturers, especially those that produce equipment for mobile devices or high-precision (e.g., surveying) applications.

Except for military applications, the market potential is more tenuous for a non-GPS receiver like a BeiDou+GLONASS receiver despite a formal agreement between the United States and Russia to foster cooperation in their respective national satellite navigation systems. Widely used civil GPS receivers are already rife in those countries, particularly in smartphones and automobiles. U.S./European and Russian cooperation on GNSS interoperability has prevailed despite political tensions. Regional systems like the QZSS/MSAS and NavIC (IRNSS)/GAGAN are just beginning to show their utility. From practical and technical points of view, it is likely that most multiconstellation receivers will always include GPS capability in addition to other GNSS functionality [1]. The prevalence of particular receiver types and their ability to track multiconstellations is shown in Figures 14.1 and 14.2 [2].

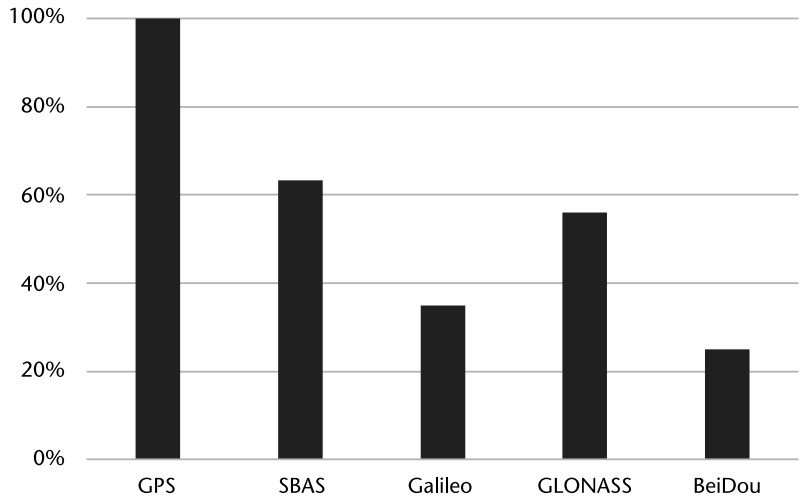


Figure 14.1 Capability of GNSS receivers: all market segments. (Courtesy of GSA.)

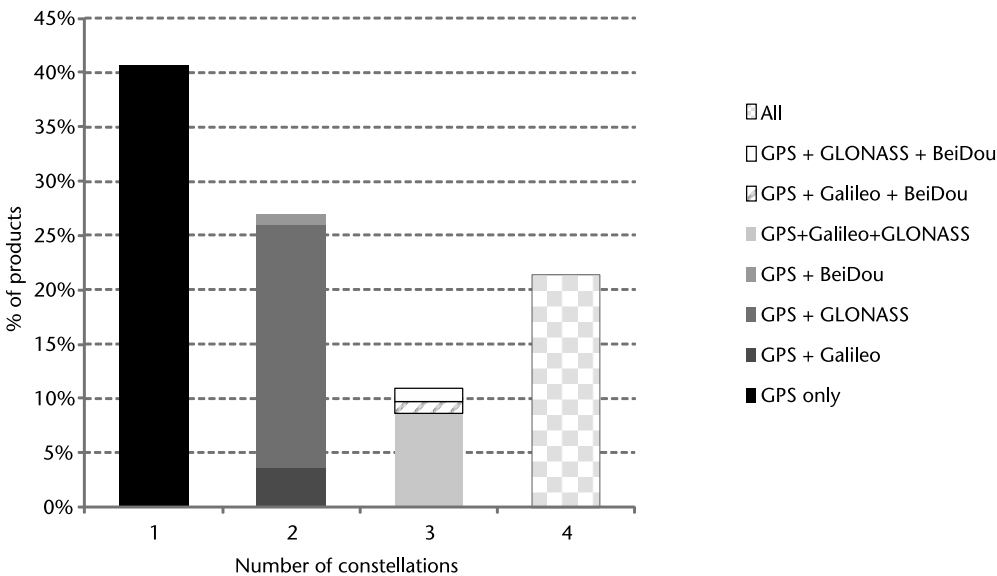


Figure 14.2 Supported constellations by receivers: all market segments. (Courtesy of GSA.)

14.1.2 Defining the Market Challenges

Market definitions usually start by counting the sales of the goods and services loosely associated with a technology, but how does one aggregate and quantify an ensemble of goods such as GNSS receivers that range from $1 chips that are components of a GNSS receiver for use inside cell phones to large, $300,000, nuclear-hardened navigation sets, deployed inside a submarine or space qualified for use in a spacecraft? How does one account for all the value-added applications enabled by GNSS? Are they part of the GNSS market?

The European Global Navigation Satellite Systems Agency (GSA) [2] in Prague has made an admirable attempt at describing the civilian GNSS market. They projected today's 6 billion GNSS deployed devices to grow to over 9 billion by 2023 (Figure 14.3). That is more than one unit for every person on Earth (Figure 14.4). While the U.S. and European markets will grow at 8% per year, Asia and the Pacific Region will grow at 11% per year [2]. The total world market is expected to grow about 8% over the next 2 years due primarily to GNSS use in smart phones and location-based services [2]. 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 was expected to rise from approximately €85 billion ($90 billion) (at the time of this writing, 1 Euro = $1.06) in 2017 to just over €100 billion ($106 billion) by 2021. Enabled revenue should stay fairly flat at €260 billion ($276 billion) over the period but was estimated to rise after 2020 as Galileo and BeiDou reach full operational capability. Figure 14.5 shows the global GNSS market size in billions of Euros [2].

Figure 14.6 shows that GNSS revenue growth between now and 2023 will be dominated by mobile users and location-based services [2].

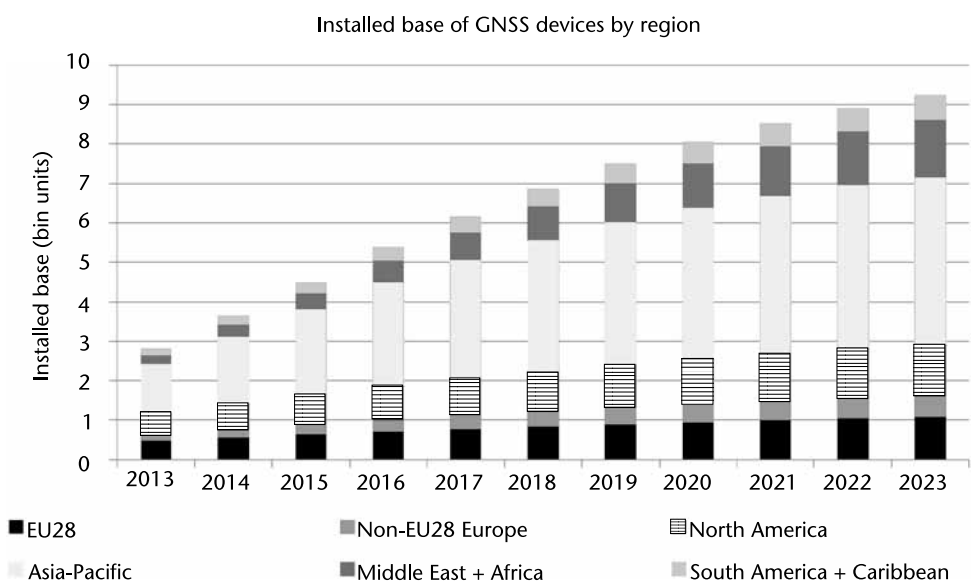


Figure 14.3 Installed base of GNSS devices by region. (Courtesy of GSA.)

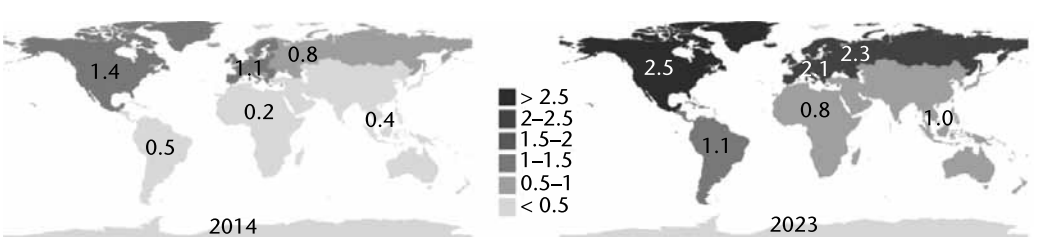


Figure 14.4 GNSS devices per capita: 2014 and 2023. (Courtesy of GSA.)

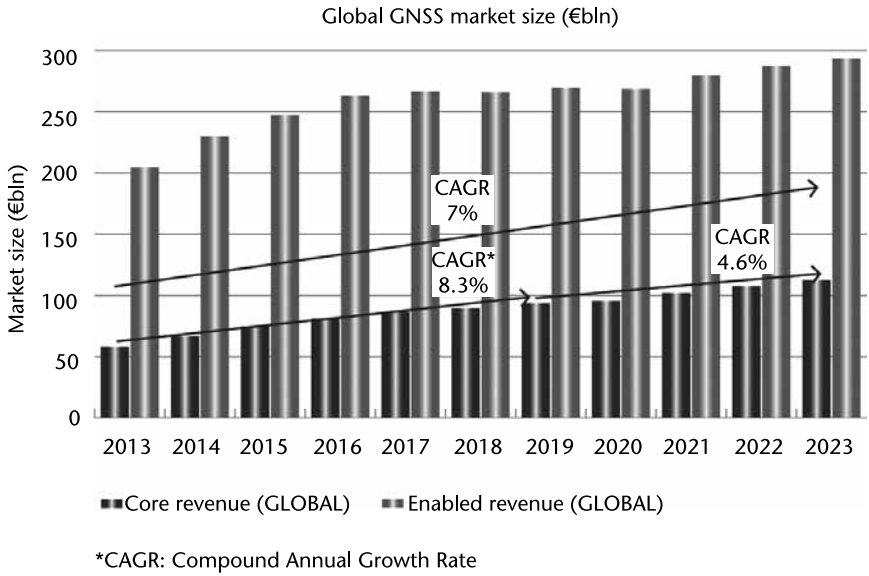


Figure 14.5 Global GNSS market size in billions of Euros. (Courtesy of GSA.)

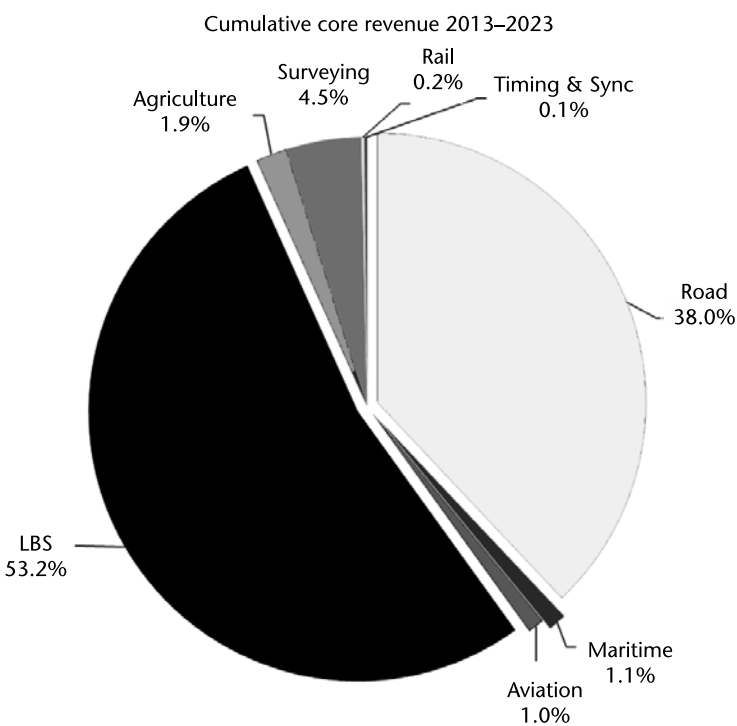


Figure 14.6 Cumulative core revenue 2013–2023 by market segment in billions of Euros. (Courtesy of GSA.)

A July 2015 study conducted by RNCOS Research, a business consulting service based in India and known for a history of GNSS forecasts, predicted that the global core GNSS market would grow at a compound annual growth rate (CAGR)

of 9% from 2015 to 2020 [3]. These forecasts were based on a detailed analysis of the number of users in the various market segments and are considered to be highly credible. However, note that, in addition to the above markets, there is the military market for GNSS receivers and services and for GNSS infrastructure (i.e., satellites, reference receivers, control segments).

14.1.3 Predicting the GNSS Market

All GNSS market projections depend on signals from the various systems being available per the promised deployment schedule while the likely future of GPS is fairly predictable as the USAF modernization program is underway. There is also high confidence that GLONASS will maintain its operational status as evidenced by the Russian government's continued support. Additionally, the Chinese, Europeans, and Indians are making considerable progress with their systems. Deployment of the QZSS space segment has been scheduled through 2023. The previously mentioned market predictions are based on a variety of data including realization of these deployment schedules.

However, defining and quantifying the enabled market segment for GNSS services remain a challenge. Counting smart phones, ships, and aircraft, is fairly straightforward, while delineating services is somewhat amorphous. Consider services such as developing receivers for the government, designing filtering software to integrate GNSS with other sensors in a commercial or military aircraft, testing the products, and installing and integrating them into vehicles and aircraft, and services such as surveying or precision agriculture that rely on GNSS information, vehicle tracking, and location-based services.

Classic definitions of the market have first split it into military and commercial (or civil, as these two terms will be used interchangeably) segments. Others break down the nonmilitary market into consumer and professional segments and note that the professional segment has some similarities with the military segment (e.g., tight accuracy specifications, rugged environmental requirements). Researchers versed in consumer electronics and professional or military markets usually perform market studies that focus on one of these segments.

Organizations performing market studies can count users, rely on sales projections of similar products, draw upon earlier experiences with those products, use existing modeling, and make educated guesses as to the potential for growth. In most cases, these studies are weak in one or more areas (e.g., aviation and marine), but strong in others like consumer products or mobile location services. This is not surprising, as most of these research firms are likely to specialize more in some particular market segments than in others. They do a great job in a micro sense with demographics, historical data, focus groups, surveys, and competitive analyses. Their results are used to decide on investments in new products and new ventures, but in a macro sense they just cannot accurately describe, let alone forecast the totality of something as multifaceted as the GNSS market. It is doubtful that anyone could perform a comprehensive forecast with high confidence in its outcome beyond a year or two.

Almost all previous studies have relegated the military market to a small fraction of the civil market. While it is true that the total dollars expended in the

military market is small compared to the total for civil markets, it is nonetheless significant with approximately $40 billion spent just on GPS to date and with another $15 to $20 billion more expected through the implementation of GPS III [3]. Defense budgets provide for operational funding and seed money for developments that often lead to new or enhanced civilian applications. Even more importantly, the military value of GPS as a force multiplier is the primary reason why it was conceived and remains funded, supported, and sustained. This investment enables the civil market.

Similar considerations are likely prevalent in the defense establishments of Russia and China and, to some extent, in Europe because all of these GNSS have secure, encrypted signals and these entities would like to have an alternative to GPS that does not rely on U.S. Defense Department control. The civil component has become important; therefore, there is no doubt that civil GPS services will be maintained even if the military eventually migrates to some new technology to satisfy its navigation, positioning, and timing needs. Furthermore, the U.S. military is planning on using GPS at least until 2030 [3].

While there are significant differences between commercial and military markets, consider that in the commercial marketplace:

- The market size varies smoothly with supply and demand.
- The seller bears the development risk.
- There are many buyers.
- There are many competitors for market share.
- There are many similar products.
- Prices are set by marginal utility.

While in the military market:

- There is erratic buying behavior due to changing requirements and budgets.
- The government usually bears any development risk.
- There are relatively few buyers.
- In most cases, there are few competitors for market share.
- Product requirements vary significantly among customers.
- Performance is more important than price.

The most important difference may be that in the military market there is a substantial return on investment (ROI) because a company's investment is relatively low. Profitability is certainly also lower in military markets as the allowed amount of profit is usually limited by legislation. Still, the real ROI can be much higher than in civil markets as the risk associated with the investment is much lower for the military market. Yet many military products and technologies eventually find their way into the commercial market. These are called dual-use systems. After the Internet, GPS is likely the second greatest, modern dual-use military system in terms of impact on our civilization.

14.1.4 Changes in the Market over Time

In the first edition of this book in 1996, GPS was described as an enabling technology. It certainly is that, but it is also a ubiquitous technology. With the hindsight of recent history, one can see how GPS has not only enabled new applications heretofore unknown, but it has permeated almost all aspects of commerce, agriculture, leisure, travel, and warfare (e.g., GPS-equipped smart bombs and drones).

GPS has become a critical piece of the United States and other nations' infrastructural underpinning as increasingly more people and functions depend on it for positioning and timing. Subsequent to the issuance of the second edition of this book in 2006, GPS became the mainstay technology for almost all U.S. and Allied nations' weapons systems. Many other nations' militaries adopted civil GPS receivers for their weapon systems. Now some of these nations are switching to the other SATNAV systems as the maturity of these systems allow. Russian and Chinese militaries, having used GPS, have begun equipping their forces with their own GNSS hardware [4]. Civil adoptions of the other SATNAV systems for receivers that heretofore just used GPS is occurring at a rapid pace but more for technical reasons such as to obtain increased availability and accuracy. Today, most of the receiver chipsets on the market are at a minimum GPS+GLONASS capable.

14.1.5 Market Scope and Segmentation

The definition of the GNSS market that is used here is the dollar value of all the goods (such as GNSS receivers, antennas, chipsets) and services (such as software development, testing, integration, location-based services) provided to users of GNSS or to applications, which incorporate GNSS receivers. We cannot logically include such things as flight management systems or the total value of an integrated GNSS/INS, but the GNSS receiver and integrating software is included. In any case, the companies that benefit from this market segment (i.e., space and control segment development and fielding) generally are not the same companies that serve the market segments that deal with equipment or services for users of GNSS.

14.1.6 Dependence on Policies

The GPS component of the GNSS market is obviously global since users are all over the world, yet much of the potential for global GPS market growth is dependent on U.S. government actions and policy. Policies such as the E911 mandate from the U.S. Federal Communications Commission (FCC) that require cell phone operators to pinpoint their users who call 911 (112 in Europe) has spurred the growth of GPS chips for cell phones as the primary way to satisfy the mandate.

This has led to the myriad of location-based services that rely on today's cell phones knowing their locations. Because of the vast difference in number of users, the civil market value will always be far greater than the military market value. For example, a study presented to the U.S. National Space-Based Positioning, Navigation and Timing Board in 2015 asserted that GPS contributed $68 billion to the U.S. national economy in 2013, presaging continuing funding for it by the U.S. government [5].

14.1.7 Unique Aspects of GNSS Market

Markets can be thought of in a hierarchical way with the total market subsuming an addressable market subsuming an achievable market. A company interested in entering the market or concerned with forecasting possible sales will start with the total market, which includes all the goods and services described above. It includes both military and civilian markets and as noted, is global in nature. From that is derived an addressable market that fits the company's business and capabilities. Within that, an achievable or, as some might call it, an expected market becomes their annual sales goal. An example might be the market addressed by a civil GNSS chipmaker. This addressable market would eliminate the military market but consider all civil receiver manufacturers, and chipset adapters, like cell phone manufacturers as potential customers. Figure 14.7 describes the approach.

Another approach is to come at it from the number of possible users of the technology. This is done by just counting the ships, aircraft, hikers, autos, trucks, laptops, pads, smartphones, wearables, and effectively anything that moves. Afterwards, an educated guess is used to quantify what portion of these users of these products will need a GNSS chipset.

The GNSS chipset will most likely incorporate the majority of the SATNAV constellations discussed within earlier chapters. With the present flexible software-based digital signal processing, it is well within the state of the art to develop products that can utilize any and all signals in view. Today's receivers not only can process all satellite signals in view but in many applications also make use of terrestrial signals such as those emanating from cell towers and WiFi. This is particularly important for any indoor applications or other environments where satellite signal reception is degraded. With the advent of wearable receivers, product developers will have new challenges for antenna and battery configurations.

14.1.8 Sales Forecasting

Much research and development activity, employment, and capital expenditures of a company are driven by the sales forecast, which is a best guess as to how many

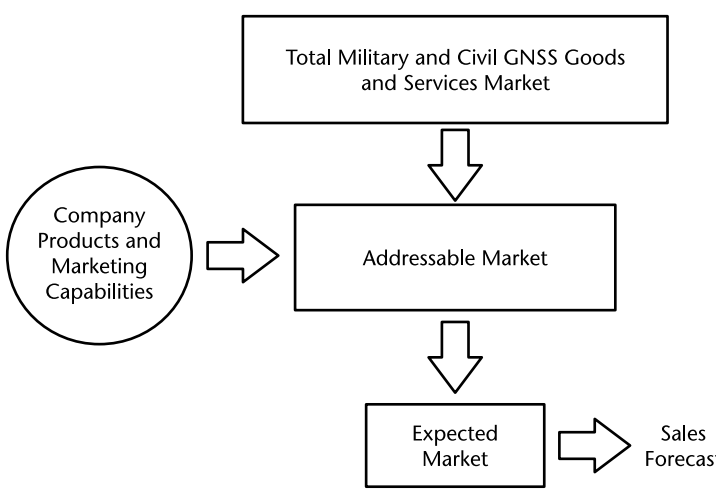


Figure 14.7 Segmenting the GNSS market.

sales can be made from a market definition that is not an exact science. Yet that is the best data from which to start. Fortunately, there is usually historical and competitive information that lends credence to such a forecast. Forecasting a year or even two into the future is usually successful, but any longer-term forecast is more likely to be highly inaccurate. The longer-term forecasts made in this chapter of the second edition of this book in 2006 and an expanded version of it in 2007 [6–8] were proven wrong due to even longer deployment delays of the various GNSS than were assumed at the time of their publication and by the 2008–2009 recession.

In terms of GPS, forecasting in the military is much simpler although not built with any more confidence. The data provided by U.S. government budgets provide a starting point that in general is fairly accurate, at least in the short term. Budgets traditionally cover 5 or more years, so the military GPS equipment forecast is available to a potential supplier. In the United States, Congress and defense priorities often change these forecasts but usually not more than annually. Projects and procurements span several years so there is a built-in inertia that keeps the forecast somewhat stable.

14.1.9 Market Limitations, Competitive Systems and Policy

14.1.9.1 Changes in Market Growth Projections

As mentioned above, GNSS market growth is highly dependent on U.S. and other government actions and policies. Some possible changes are:

- Time to deploy all the new signals including for L5 as other civil signals: Compatibility and interoperability, if any, of the signals common to aforementioned SATNAV systems are all subject to outcomes of government-to-government negotiations.
- Export rule changes and regulatory demands: While U.S. export limits on GPS may never be any more stringent; Galileo receivers or hybrid GPS/Galileo receivers could be mandated for use in Europe. There could also be charges in the form of tariffs or royalties imposed on these receivers, thus limiting the market for them, although recent U.S./EU agreements would militate against that. Similar regulations are possible in Russia and China.
- Expansion of the E911 mandate and its equivalent in Europe to other countries: These have increased the market worldwide as it has done in the United States.
- Regulatory changes that allow for terrestrial transmitters that could interfere with GNSS [e.g., Ligado (formerly LightSquared)].
- Court decisions: Those regarding privacy issues arising from the use of GNSS by law enforcement for tracking suspects and criminals may have a small negative effect on the overall market, but potential liability issues may have a bigger impact.
- Negotiations: Those between the United States and the European Union regarding FCC approval of the use of Galileo's signals over the United States and U.S. Department of Defense access to Galileo's PRS.

In late 2004, then U.S. President Bush issued a new policy on space-based PNT. The policy stressed the military value of GPS to the United States by mentioning the importance of navigation warfare (Navwar) training, testing, and exercises several times. However, it also maintained the commitment to discontinue the use of SA. That policy is still in effect.

The GNSS markets can only expand as they mature. Besides full deployment of Galileo and BeiDou, there is now a more robust GLONASS. The NavIC space segment is fully deployed with GAGAN operational. Japan's QZSS is in development, with one SV used for test purposes while MSAS is operational. Although some of this added market potential will be related to SBAS applications, there will be new combined receiver applications business for many of the world's chipmakers and receiver suppliers.

14.1.9.2 Market Risks

Like any venture, there are always risks to success. The GNSS market looks extremely promising, but there are concerns that any prudent entrepreneur should be aware of. As GNSS receivers embed themselves in our cars, cell phones, laptops, watches, cameras, and wearables, and they become wedded to wireless communications links, a potential backlash from consumers could limit market growth. More and more, we are becoming a society where privacy rights are being eroded by fear of crime and terrorism abetted by technology that fosters the erosion. Telematics, or the provision of services to mobile users, particularly automated vehicles, is one area where the line between location awareness by the service provider can easily become location awareness by unwelcome persons doing surveillance and hackers.

14.2 Civil Applications of GNSS

As shown in Figure 14.6, the major sectors of civil GNSS applications are:

- Location-based services;
- Road;
- Surveying;
- Agriculture;
- Maritime;
- Aviation;
- Rail;
- Timing and synchronization;
- Space.

Based on the information provided in [2, 9], key applications are discussed with examples provided for each of the above sectors. Also, indoor GNSS usage is treated.

14.2.1 Location-Based Services

At over 53% of the GNSS market, location-based services (LBS) applications have permeated most of our daily lives. This is due to the embedding of a GNSS receiver into a smartphone, tablet, camera, and/or wearable device. The most common LBS usage is personal vehicle or pedestrian navigation via the use of a smartphone providing turn-by-turn directions coupled with a digital map. In 2014, 3.08 billion smartphones dominated LBS devices.

In addition to personal navigation, today's GNSS devices are used for a plethora of applications including:

- Safety and emergency assistance via E112, E911, and similar services;
- Keeping track of children, teenage drivers, and patients with Alzheimer disease;
- Helping blind people navigate;
- When looking for carpools or a ride via UBER, LYFT, or Didi Chuxing in China;
- Runners, bikers, and joggers keeping track of their location, speed, and distance;
- Playing games such as Pokémon Go;
- Golf aids (e.g., location on the course, distance to the hole);
- Social networking by exchanging location information among friends and associates;
- Geo-tagging pictures within a camera;
- Conducting a self-guided tour without any external signs or references in an outdoor park;
- Retailers providing personalized offers to potential customers within a particular geographic area.

14.2.2.1 LBS Wearables

Broadcom, OriginGPS+, and other chip manufacturers are offering new GNSS chips specifically for wearables. These chips do not load heavy software into its processor. They concentrate on minimizing power while maintaining accuracy. However, wearable technology is not limited to the wrist. New flexible materials can conform to the body and become part of clothing or in one's shoes. In addition, a person's movements using piezoelectric-coated film on nickel film encapsulated in Kapton tape can charge the batteries of any electronics embedded therein [10]. Figure 14.8 shows a sampling of GNSS wearables.

14.2.2.2 The Internet of Things (IoT)

The Internet is continuing to swallow up the entire planet full of users as electronics continues to shrink GNSS, especially via its inclusion in smart phones, follows along. Everything that moves or needs accurate time could be a potential user [11]:

Garmin EPIX GPS/GLONASS Navigator

Whistle GPS Pet Tracking Collar

Shoes with GTX Corp GPS Tracking Insoles

Figure 14.8 Sampling of GNSS wearables. (EPIX GPS/GLONASS Navigator Copyright © Garmin Ltd.)

"The Internet of Things (IoT) – the integration of uniquely identifiable devices on the Internet – is one of the main current global technology themes and GNSS is integral to its success. Location based services and timing data are essential to IoT applications in particular as a means to control and monitor mobile IoT devices."

14.2.2 Road

Road users are the second largest market for GNSS receivers with over 1 billion cars and 130 million trucks in the world. These users employ a portable navigation device (e.g., Tom Tom, Garmin) or an in-vehicle technology (i.e., built-in dashboard).

The value of current fleet information provided by GNSS is evident for delivery, emergency vehicle, and scheduled service fleet dispatch and control. Automatic vehicle location systems (AVLS) have been developed and installed in many of the world's trucking and emergency fleets. Qualcomm pioneered fleet tracking with over 500,000 trucks and other fleet vehicles tracked via its OmniTRACS System, before selling the business to private investors in 2013. Many of these are GNSS-equipped, and used primarily outside of the United States and particularly in Mexico and South America. One concept employed is called geo-fencing, in which a vehicle's GNSS is programmed with a fixed geographical area and alerts the fleet operator whenever the vehicle violates the prescribed fence.

There are tariff systems where total road usage is tracked using GNSS and taxed rather than just on given roadways. An example is SkyToll, which is used in Slovakia. At 17,741 km, the Slovak Electronic Toll System is the longest tolled roadway system in the European Union. The system was started in 2010 and uses EGNOS and Galileo, to track a vehicle's movements and provide related vehicle data to a tolling authority [12].

Advanced Tracking Technologies Inc. (ATTI) has developed a GNSS-based system to improve efficiency and reliability for transportation systems, such as public buses and private taxis. By monitoring fleet assets, dispatchers can provide rerouting information as well as determine how long a vehicle was idling and take corrective actions. Both of these result in reduced fuel costs. In addition, the system allows riders with text messaging, Twitter, or Facebook to receive messages and

tweets keeping them updated on the location of their bus. Riders no longer have to wait out in the cold and rain for a bus. They just wait inside until they get the alert that the bus is near their stop [13].

The largest operator of a GPS-based land navigation service is OnStar, a General Motors subsidiary. In 2014, over 5 million vehicles were equipped with GPS receivers that communicate with OnStar operators via cell phone to provide either voice commands or map guidance to the driver. Other automobile manufacturers have similar services. Many new businesses based on smart phone applications like UBER and Lyft would not be possible without the embedded GNSS receivers in the phones.

Rental car companies have a strong incentive to offer navigation information to their customers. Hertz relies on the NeverLost system based on a Magellan Roadmate receiver (at the time of this writing, the sixth generation of NeverLost, NeverLost Gen6, was available), while the AVIS where2 system uses a GARMIN solution.

Autonomous road vehicles will likely utilize a GNSS receiver integrated with speed, heading, and other sensors such as stereo vision, radar, lidar, ultrasonic sensors, and IMUs.

For example, Tesla vehicles manufactured after October 2016 have eight surround cameras providing 360° visibility around the car at up to 250m of range. There are also 12 ultrasonic sensors complementing this vision information. These allow for detection of both hard and soft objects. These sensors are coupled with a forward-facing radar that provides additional data about the world on a redundant wavelength, capable of seeing through heavy rain, fog, dust, and even the car ahead [14].

14.2.3 GNSS in Surveying, Mapping, and Geographical Information Systems

GNSS receiver technology owes much to its early application in the business of land surveying. The production of maps and charts and the georeferencing of data using GNSS are natural outgrowths of the accurate and reliable techniques developed for the land-survey market.

Utilizing the DGNSS and PPP techniques described in Chapter 12, the applications excerpted from [2] are realized:

- Cadastral survey aims to establish property boundaries. Fiscal policies such as land taxation rely heavily on cadastral surveying.
- Construction surveying covers the different construction stages of a building or civil engineering project, whereas machine control applications automate construction activities:
  - Machine control applications use GNSS positioning, for example, to automatically control the blades and buckets of construction equipment using information provided by three-dimensional (3-D) digital design.
  - Person-based applications enable many positioning tasks, including making surveys, checking levels, performing built checks, and staking out reference points and markers.

- In mapping, GNSS is used to define specific location points of interest for cartographic, environmental, and urban planning purposes.
- Mine surveying involves measurements and calculations at each stage of mine exploitation, including a safety check.
- Marine surveying encompasses a wide range of activities (seabed exploration, tide and current estimation, offshore surveying, and so forth), all of whose outcomes are important for maritime navigation.

14.2.3.1 Geographical Information Systems

A geographical information system (GIS) is a computer system designed to allow users to collect, manage, and analyze large volumes of spatially referenced information and associated attribute data. As such, it is an organized collection of computer hardware, software, and geographic data, designed to efficiently capture, store, update, manipulate, integrate, analyze, and display all forms of geographically referenced information.

Specific locations recorded may be annotated with location-specific information, such as street address, elevation, or vegetation type, location of utility control boxes, sewers, and power lines. This type of data collection is the building blocks of data for GIS. Personnel equipped with handheld GNSS units with onboard data storage or with a communication link for direct transfer to a central storage point can collect the raw data. Vehicles, ships, and aircraft in addition to people on-foot collect some data for these types of systems.

14.2.4 Agriculture

Both the agriculture and farming industry make heavy use of GNSS and GIS as part of a modern precision farming system. Whether it is mapping where soil samples are taken, spraying fertilizer, seed, or insecticide or directing combine machines exactly where to go to harvest a crop, the application of these materials has become an exact science. Many farm implement manufacturers are producing variable-rate application equipment that is controlled by sophisticated electronics coupled to an information system. It has resulted in lower material input costs and higher yields.

Furthermore, harmful effects of the runoff of unneeded fertilizers are mitigated. For this reason, it is possible that the variable application of fertilizers might be legislatively controlled.

As stated in [9], other agricultural and farming applications include:

- Precision soil sampling, data collection, and data analysis enable localized variation of chemical applications and planting density to suit specific areas of the field.
- Accurate field navigation minimizes redundant applications and skipped areas, and enables maximum ground coverage in the shortest possible time.
- The ability to work through low visibility field conditions such as rain, dust, fog, and darkness increases productivity.

- Accurately monitored yield data enables future site-specific field preparation.
- The elimination of the need for human "flaggers" increases spray efficiency and minimizes over-spray.
- There are tractor guidance and crop spraying.

14.2.5 Maritime

Like aircraft, marine users can usually see the open sky. However, even they have to compete with other electronics devices for antenna placement near the top of the mast on their vessels. While clearly not the largest market segment, marine navigation was the first to embrace satellite navigation. Knowing one's position on the open ocean is a primary requirement for vessels navigating to a destination as they transit the seas and/or inland waterways.

Even submarines can use GNSS whenever they can get their antennas close to or above the surface. Since the early 1980s, sea-level users need only three satellites in view to get a two-dimensional fix; GPS has been used to fix positions on the ocean. Today the market is mature. Along with radios and radar, a GNSS receiver is a piece of standard equipment on any boat operating far from shore. Most can obtain differential corrections from an SBAS. Many others use corrections provided by a radio beacon-based system (e.g., the NDGPS) if available.

Figure 14.9 shows a marine navigator with database management capability and graphical display of position and speed information. In this market, ease of use and the ability to manage a large database of waypoints and sophisticated cartography are key requirements.

Figure 14.9 Typical GNSS marine navigator. (Courtesy of Furuno Inc.)

Fisheries management is a worldwide mandate requiring swift action by governments when a sea boundary is intruded upon. Dwindling fish stocks have prompted the establishment of strict guidelines for fishermen and the closure of entire grounds. The situation is also making countries that share sea boundaries more sensitive to foreign fishing in their waters. These tensions engender the need for accurate position determination and recording to prove or disprove a boundary violation, particularly in the South China Sea.

GNSS can aid in the berthing and docking of large vessels, by means of position, attitude, and heading reference systems (PAHRS). These installations use multiple antennas aboard the vessel along with DGNSS corrections to determine an accurate representation of the ship's orientation and position. Combined with appropriate reference cartography, this can be an immense aid in the handling of large vessels in close quarters. Vessels worldwide are candidates for this type of system.

There is a market for extremely accurate positioning for marine seismic survey and oil exploration activities as well as in dredging, buoy laying, and maintenance. Dredge operators are paid based on the amount of material that they remove from a harbor or shipping channel, so accurate measurement of position can optimize the operation, reducing cost and wasted effort.

The availability of GNSS and accurate differential services has proven to be a boon to the development of precise seismic maps and location of drill sites with respect to identified geologic structure, especially in the offshore case, where exploration teams have paid significant revenue per day for accurate satellite positioning services. The availability of such accurate systems for navigation has enabled much resurveying of published marine chart information. A good portion of the data currently represented on marine charts is over 60 years old and hydrographic services are involved in the production of digital databases to an agreed-upon international format (IHO S-57).

The rise of worldwide terrorism and piracy has spurred the development of means of tracking of large container ships as they ply the seas. GNSS plays an important role in these kinds of systems, which also rely on satellite communications and electronic tagging.

Recreational vessels make good use of basic GNSS for navigation, and the acceptance of differential GNSS bodes well for the health of that sector. The huge number of vessels and the value of GNSS in marine navigation, fishing, and waterway maintenance, coupled with strong economic activity, will allow steady growth to a level of near $1.1 billion by 2020. However, this segment has a fairly low growth rate due to the maturity of the market [2].

14.2.6 Aviation

If it moves above the Earth and it has an associated GNSS receiver, it is an air application. From birds to drones, to airplanes to satellites and even to space-based vehicles, GNSS is widely used aloft for navigation, tracking, aviation operations, sensor integration, science, and recreational activities. Many of the more sophisticated applications marry GNSS receivers to inertial units as well as to communications capabilities. It is a key sensor for flight management systems.

The big need for navigation by GNSS was primarily for over ocean operations where there were no VHF Omnidirectional Range/Distance Measuring Equipment

(VOR/DME) stations and in parts of the world where radio NAVAIDS were sparse and primitive. Just as cell phones became rife in these developing countries due to the impossibility of providing landline phones, GPS was thought of as a technological leap from basic radio beacons.

GNSS could be applied to all phases of flight operations if only its accuracy, integrity, and continuity of service could be assured to the acceptable levels demanded for safety-of-life applications. However, introducing GPS into the U.S. national airspace caused some major issues. Over the United States, the en-route VOR/DME system was adequate at least until the traffic load swamped the Air Traffic Management (ATM) system. Nonetheless with GPS, aircraft would not have to stay on these fixed highways and thus could fly great circle routes and/or optimum fuel consumption routes. Capacity limitations of the present system and skyrocketing fuel costs eventually overcame airlines' resistance to new equipment installations as long as the cost benefit of using GPS could be shown to be positive.

The FAA was also faced with the growing number of aircraft clogging the skies. Approach and landing operations became a critical bottleneck as airports also reached capacities. Many airports with runways totally without instrumentation could potentially minimize their unavailability problems (largely due to inclement weather) by utilizing a GNSS solution for approach and landing. Using GNSS for approach and landing requires a very high level of integrity as well as accuracy, availability, and continuity of service. To reach the specified integrity level, a continuous check on the performance and the quality of the information being derived from GNSS required an independent system. Thus, the first SBAS, the FAA's WAAS, came to be. Yet even WAAS could not provide the required integrity for landing in all categories of weather and visibility conditions. For the most stringent requirements, a LAAS now denoted as the GBAS is needed and is slowly being deployed [15]. (See Section 14.2.6.1.)

Worldwide, this capability continues with EGNOS and Galileo over Europe, by Russia's GLONASS, India's GAGAN, and Japan's MSAS. Modernized GNSS will accommodate greater civil aviation use with its L5 signals. In the 2016–2020 time periods, it is anticipated that there will be a seamless, next-generation system as far as air traffic management is concerned so that aircraft can use their standardized equipment to fly safely in any civil airspace with the same level of confidence in their navigation and positioning.

The GA market for GNSS capability is seeing a surge of activity after publication of GPS nonprecision approaches at the busiest airports and at most of the others. SBASs are being fielded to provide services equivalent to WAAS in other regions of the world. Figure 14.10 is representative of a GA aircraft navigator.

Many airlines routinely check on their aircraft in flight. Those equipped with GNSS can accurately report their position. Over broad ocean areas, they must utilize a leased communications satellite channel. One airline that did not choose to enter into such a lease was Malaysia Air, resulting in a lack of location information when of one of their aircraft disappeared over the Indian Ocean in 2014. While the incident may not have been preventable, at least the area of the search for the aircraft should have been much smaller. China's Civil Aviation Administration has announced that they will be testing a tracking system for general aviation aircraft and then cargo and passenger aircraft with BeiDou [16].

Figure 14.10 Typical general aviation navigator. (Copyright © Garmin Ltd.)

14.2.6.1 Precision Approach Aircraft Landing Systems

Most instrument approaches carried out by commercial air carriers are precision approaches. Unlike nonprecision approaches, these procedures give glideslope guidance to the aircraft on approach. The lack of signal integrity among other performance parameters precludes the use of unaided GPS for demanding aviation applications. These applications require the use of either code differential and/or kinematic carrier-phase tracking techniques. The FAA's WAAS provides warning and sufficient accuracy to perform close to Category-I precision landing requirements. This allows about 90% of the airline approaches currently performed to use a GPS approach augmented in this way. Similar to WAAS operation, there is operational usage of EGNOS.

Approaches, involving lower weather minima, also require improved accuracy and integrity warnings, which will be provided by airport-based differential stations broadcasting GNSS corrections directly to the aircraft on approach (i.e., GBAS) [15].

Some GBAS are just now being deployed with a special dispensation from the FAA. Many airports have very difficult approaches due to surrounding terrain such as high mountains and narrow valleys or regular poor visibility. Airports such as New Jersey's Liberty International in Newark, Houston Intercontinental, and in Zurich, Switzerland and some in Alaska are benefiting from a GBAS. Another example of such a deployment is in Sydney, where Qantas aircraft now have GBAS landing capability. However, replacing ILS with GBAS is a slow process that requires new avionics. Boeing 787 and 747-8 have GBAS avionics as standard equipment. Such receivers are available as options on the 737 and on various Airbus models [17].

14.2.6.2 Other Enterprises and Uses of Air Application

Beyond the primary air application for navigation, there are many enterprises and users that rely on GNSS inputs to perform other missions. They are described here because they occur in the air. For example, an airborne survey such as for mapmaking or resource determination or crop spraying requires precise positioning of the aircraft or precise annotation of a picture or other sensor data with aircraft's exact position and time when the picture was taken or the data was recorded. In terms of aircraft flight testing, there may be a GNSS receiver on board (separate from any navigation receivers) as part of a black box to be used to reconstruct the test

aircraft's PVT. There may also be a GNSS receiver as part of the black box use in accident investigations.

Weather balloons and radiosondes are air applications that also make use of GNSS, as do parachutists, hang glider pilots, and remotely piloted and unmanned aerial vehicle (UAV) and drone operators. Most of the latter are found in military surveillance and reconnaissance missions and increasingly in combat operations. Civil use of drones is on the rise for such applications as fire reconnaissance and real estate marketing. These unmanned applications are almost impossible to conceive of without employing some sort of GNSS guidance.

14.2.7 Unmanned Aerial Vehicles (UAV) and Drones

Today, UAVs and drones are flying everywhere in spite of the fact that governments have been only recently issuing regulations regarding their use in the airspace. (Herein, a UAV is defined as a remotely piloted unmanned aerial vehicle while a drone flies a preplanned flight path.) Devolving from military technology, there are a myriad of civilian drones performing professional and recreational missions. Militaries and intelligence agencies, especially the U.S. Department of Defense have been the most active operating hundreds of UAVs and drones for reconnaissance, surveillance, bomb damage assessment, and so forth and a small number for weapons delivery against suspected terrorist targets. Civilian uses include airborne surveillance by police, photogrammetric survey, hobby flying, and so forth [18]. At the time of this writing, Amazon had started a business unit denoted as Amazon Prime Air that is planning to use drones to deliver packages directly to consumers. Figure 14.11 shows a typical drone equipped with GNSS.

14.2.8 Rail

Rail applications generally utilize the DGNSS techniques described in Chapter 12, these applications excerpted from [2] are:

- High-density command and control systems assist train command and control on main lines, referring primarily to the European Train Control System (ETCS) in Europe and some regions in the rest of the world, as well as positive train control (PTC) in North America. GNSS can also be a source of additional input (e.g., for enhanced odometry in ETCS or to support PTC).
- Low-density line command and control systems provide full signaling capabilities supported by GNSS on lines with small to medium traffic. These lines are usually located in rural areas, where cost savings can be vital for the viability of a service.
- Asset management includes such functions as fleet management, need-based maintenance, infrastructure charges, and intermodal transfers. GNSS is increasingly seen as a standard source of positioning and timing information in these systems.
- Passenger information systems on-board trains show the real-time location of a train along its route. Increasingly, the GNSS location of a train is also supporting platform and online passenger information services.

Figure 14.11 Typical GNSS-equipped drone. (Courtesy of Trimble Navigation Limited.)

14.2.9 Timing and Synchronization

GNSS disseminates time within the Coordinated Universal Time (UTC) timescale. It provides atomic standard-based time to users worldwide and enables precise synchronization of a number of applications including cellular base station handover, power grids, time slot management, and network time protocol. It also enables frequency reference control, calibration of test instruments, time and frequency distribution, and time stamping of financial transactions. These timing applications are critical for a functioning modern economy.

As far back as 1998, there was ongoing work to combine GPS and GLONASS observations to obtain even more accurate timing accuracy than was achievable with GPS alone. Predictions of stabilities of 100 ps per day down to tens of picoseconds per day were expected [19].

It is likely that there will be many more GNSS-based timing applications as satellite AFS technology matures. For example, the hydrogen MASER-based frequency standards on some Galileo satellites are providing a more stable time reference than the cesium- and rubidium-based standards on other GNSS satellites.

Since many receivers make measurements from multiple GNSS constellations some timing applications will have to reconcile any time differences between the SATNAV systems in use. These time differences may be broadcast as part of the SATNAV system data message or may be computed using an additional satellite in the PNT solution. (See Section 11.2.5 for details.)

Microsemi in the United States and SPECTRACOM in Europe are just two of the companies specializing in products that use GNSS signals for timing and synchronization [20].

14.2.10 Space Applications

GNSS has multiple applications for space-based operations. As stated in [9], benefits of using [GNSS] include:

- Navigation solutions—providing high precision orbit determination, and minimum ground control crews, with existing space-qualified GPS units.
- Attitude solutions—replacing high cost on-board attitude sensors with low-cost multiple GPS antennae and specialized algorithms.
- Timing solutions—replacing expensive spacecraft atomic clocks with low-cost, precise time GPS receivers.
- Constellation control—providing single point-of-contact to control for the orbit maintenance of large numbers of space vehicles such as telecommunication satellites.
- Formation flying—allowing precision satellite formations with minimal intervention from ground crews.
- Virtual platforms—providing automatic "station-keeping" and relative position services for advanced science tracking maneuvers such as interferometry.
- Launch vehicle tracking—replacing or augmenting tracking radars with higher precision, lower-cost GPS units for range safety and autonomous flight termination.

14.2.11 GNSS Indoor Challenges

Achieving GNSS usage indoors remains a challenge because of their relatively weak signals and inability to penetrate structures. Most commercial GPS receivers a decade ago did not function well when the antenna did not have a clear view of the sky. This limitation had been addressed in a variety of ways such as with improved signal acquisition performance, additional satellite signal power, new civil signals and signal aiding from augmentations and cell tower transmissions. (Chapter 13 provides details of signal aiding from augmentations including cellular networks.) Even with adequate solutions to the indoor location problem through the use of terrestrial aiding signals, there is still the lack of availability of indoor maps. A particular exception to this is in shopping malls where potential customers' locations would be of great value to retailers who could push ads and specials to them when they were nearby.

14.3 Government and Military Applications

Since their inception, both GPS and GLONASS were designed to satisfy military requirements for worldwide PNT services. Only satellite-based systems could ensure continuous global coverage. The signals had to enable very accurate fixes yet be resistant to enemy jamming. Thus, both the United States and Russia developed user receivers that relied on their own signal called the P-code. The GPS P-code was later encrypted to be today's widely used Y-code. Authorized users such as NATO

forces and other countries with agreed-to access are using the GPS Y-code for their military activities. Details on the GLONASS P-code are contained in Section 4.7.5.

In terms of GPS, the first military applications utilized man-operated receivers on ships and other vehicles. As coverage increased with every new satellite launch, additional applications emerged until GPS became not only a useful tool, but also an essential capability for modern, network-centered warfare. GPS showed its military potential in the first Gulf War and was used prolifically in the second one, in Afghanistan and in ongoing Middle Eastern conflicts.

Modernized GPS satellites transmit the Y-code for existing military receivers and also transmit the new M-code for receivers denoted as Military GPS User Equipment (MGUE). M-code is an even more robust signal than Y-code with dispersed spectrum properties that allow for Allied forces to jam in the band center to interfere with adversary receivers that are trying to use L1 C/A-code and L2 C-code signals, without disturbing their own use of M-code.

14.3.1 Military User Equipment: Aviation, Shipboard, and Land

The original development of GPS receivers was accomplished at the Magnavox Research Laboratories (later acquired by Hughes Aircraft and subsequently by Raytheon). Some typical receivers were produced for aircraft first in a standard avionics package known as a 3/4 Air Transport Rack (ATR) size (Rockwell-Collins 3A) shrinking its width in half later to a 3/8 ATR [Rockwell-Collins and Raytheon Miniature Airborne GPS Receiver (MAGR)]. Man-portable units like the Rockwell-Collins Precision Lightweight GPS Receiver (PLGR) and the Defense Advanced GPS Receiver (DAGR) which is still several times larger than the size and weight of today's commercial handheld receivers. Figure 14.12 shows an airborne GPS Y-code military receiver and a GPS M-code receiver card. At the time of this writing, M-code receivers were in the completion stage of development and government certification.

In addition to GPS and GLONASS, BeiDou, Galileo, and NavIC all have restricted services for authorized users.

While the Galileo system is under civilian control, the PRS signals will be encrypted, and access to the service will be controlled through a government-approved secure key distribution mechanism. The PRS will only be accessible through receivers equipped with a PRS security module loaded with a valid PRS decryption key. (See Chapter 5 for PRS signal characteristics.)

In a similar manner as the design of the GPS Y-code and M-code, the PRS signals have been designed for robustness in the presence of jamming and interference. Per [21], it is stated that "… the new Galileo GNSS …is primarily a civil system which may be exploited by authorised military users. For those with access to both GPS-PPS and PRS, resilience can be increased further by combining the information from the two services into a single PNT solution."

An example of a receiver that combines Galileo PRS and GPS-PPS signals is the Q35 developed by QinetiQ (Figure 14.13). The Q35 is a multiconstellation, multifrequency PRS-enabled GNSS (Galileo-PRS + GPS-PPS) receiver. This receiver was developed for the UK PERMIT project which was a joint QinetiQ + Rockwell

Figure 14.12 Military GPS receivers. (MAGR-2000 Courtesy of Raytheon and GB-GRAM-SM © Rockwell-Collins.)

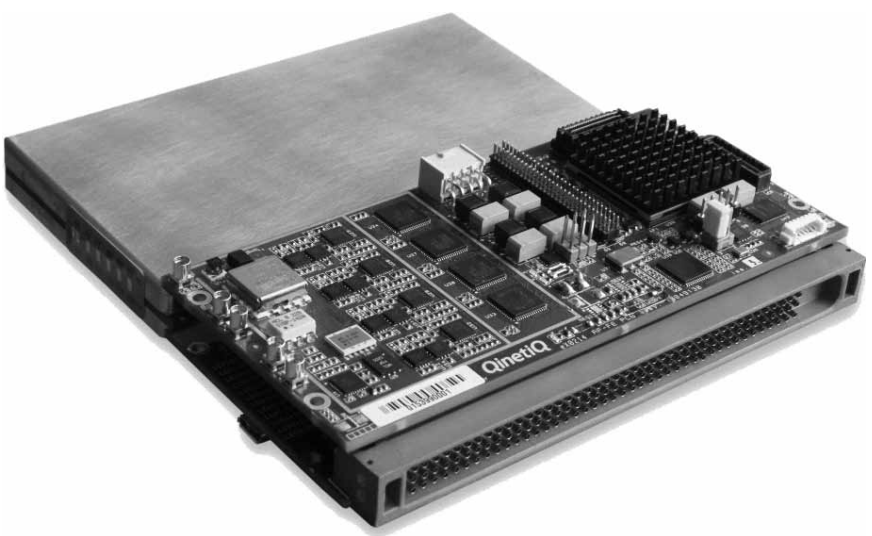


Figure 14.13 Q35 Dual mode Galileo-PRS + GPS-PPS GNSS receiver. (Courtesy of QinetiQ.)

Collins UK project. This project was sponsored by the U.K. Space Agency and Innovate UK. The UK PERMIT project goal was to investigate the challenges regarding the use of both GPS-PPS and Galileo-PRS together in a dual mode receiver and perform the first demonstration of dual mode positioning using the early Galileo satellite deployment. Reference [21] contains details of this demonstration.

As many military aircraft already had inertial navigation systems installed, work began to marry the long-term stability of GPS by virtue of its atomic timing to the short-term stability of the inertial system to create integrated navigation systems that could maintain very accurate solutions regardless of short outages to GPS caused by signal interference or vehicle dynamics and/or antenna shading. The integrations became even more symbiotic as technology allowed for faster processors, smaller receivers and lower cost, and strapped-down inertial measurement units.

14.3.2 Autonomous Receivers: Smart Weapons

Modern warfare attempts to minimize civilian casualties while maximizing their effectiveness in destroying intended targets. This requires pinpoint accuracy, on the order of a few feet in some cases. GNSS is once again the enabling technology. By combining GNSS measurements with those of an on-board inertial sensor and possibly some type of seeker (e.g., infrared), a weapon can provide the required probability of kill with a smaller warhead than would otherwise be necessary. This reduces the number of sorties required to kill a target.

GNSS receivers have found their way into ballistic missiles, guided missiles like the French SCALP EG, smart bombs like the Russian KAB-500S-E, artillery shells, and autonomous air, land, and sea vehicles, particularly for UAV and drone reconnaissance, weapons delivery and bomb damage assessment. However, the use of GNSS in combat begs the question about jamming vulnerability. For these applications anti-jam techniques are employed such as nulling antennas and ultra-tight coupling of the GNSS and the inertial sensors. In terms of GPS, increased military signal power from the forthcoming GPS III satellites further mitigates the possibility of disruption due to enemy jamming.

14.4 Conclusions

Over the next several years, the users of GNSS can look forward to increased accuracy, faster fixes and more integration of functions in their equipment. By the time that all the GNSS are fully operational, user equipment will have evolved into unimaginable complexity of function, simplicity of use, and increased cost-effectiveness for the many applications described herein. Exactly when that will happen is still subject to likely changes in national budgets, schedule impacts, and contractor performance. Such has been the history of these systems, and there is little reason to think the future development performance will be any different than the past.

Whether receiver developers can deliver new products successfully will depend on various factors, including their ability to:

- Accurately predict market requirements and evolving industry standards for the GNSS-based applications industry that they are addressing;
- Anticipate changes in technology standards, such as wireless technologies;
- Develop and introduce new products that meet market needs in a timely manner;

- Attract and retain engineering and marketing personnel and raise the required capital investment.

The next few years are the critical ones that will determine just how accurate all the market projections will turn out to be, but there is no doubt that the GNSS market for receivers, services, and applications is a fabulous growth area for the foreseeable future.

References

[1] van Diggelen, F., "Who's Your Daddy," *INSIDE GNSS Magazine*, March/April 2014.

[2] Source: GNSS Market Report, Issue 4 copyright © European GNSS Agency, 2015

[3] "Global Navigation Satillite System Market Outlook 2022," RNCOS Business Consultancy Services, Noida, India.

[4] "Putting Precision in Operations: Beidou Satellite Navigation System," *Jamestown Foundation Publication: China Brief*, Vol. 14, No. 16, August 22, 2014.

[5] "Study: GPS Contributes More Than $68B to US Economy," *INSIDE GNSS Magazine*, July/August 2015.

[6] Jacobson, L., "The Business of GNSS," *Navtech Seminars, ION-GNSS 2004*, Long Beach, CA, September 2004.

[7] Kaplan, E., et al., *Understanding GPS Principles and Applications*, Norwood, MA: Artech House, 1996, 2006.

[8] Kaplan, E., et al., *Understanding GPS Principles and Applications*, Norwood, MA: Artech House, 2006.

[9] U.S. government GPS information Web site, www.gps.gov.

[10] Pretz, K., "Health Monitors Get More Personal," *The IEEE Institute Newsletter*, Vol. 39, No. 2, June 2015.

[11] Cameron, A., "The Internet of Things and a Galileo/Copernicus Interface," *GPS World Newsletter*, December 28, 2015.

[12] Slovakia's Satellite Tolling System Receives International Recognition," *European GNSS Agency*, October 6, 2015.

[13] https://www.advantrack.com/transit-systems/.

[14] https://www.tesla.com/blog/all-tesla-cars-being-produced-now-have-full-self-driving-hardware.

[15] "What Is GBAS and Its Goal in the National Airspace System?"

[16] Press Trust of India, "China to Deploy Beidou Navigation System to Track Flights," *Economic Times of India*, July 12, 2015.

[17] Croft, J., "On the Fence," *Aviation Week Magazine*, July 5, 2015.

[18] Cosyn, P., "The Range of UAVs Across Civil Applications," *GPS WORLD Magazine*, May 2014.

[19] Lewandowski, W., and J. Azoubib, "GPS + GLONASS," *GPS World*, Vol. 9, No. 1, November 1998.

[20] Spectracomm Corporation Web site, https://www.sprectracomcorp.com.

[21] Davies, N., et al., "Towards Dual Mode Secured Navigation Using the Galileo Public Regulated Service (PRS) and GPS Precise Positioning Service (PPS)," *Proceedings of The Institute of Navigation ION GNSS 2016*, Portland, OR, September 2016.

Least Squares and Weighted Least Squares Estimates

Chris Hegarty

Let $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_M \end{bmatrix}^T$ be a column vector containing M unknown parameters that are to be estimated and $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}^T$ be a set of noisy measurements that are linearly related to $\mathbf{x}$ as described by the expression:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{A.1}$$

where $\mathbf{n} = \begin{bmatrix} n_1 & n_2 & \cdots & n_N \end{bmatrix}^T$ is a vector describing the errors corrupting the N measurements and $\mathbf{H}$ is an $N \times M$ matrix describing the connection between the measurements and $\mathbf{x}$.

The *maximum likelihood* estimate of $\mathbf{x}$, denoted as $\hat{\mathbf{x}}$, is defined as (see, e.g., [1]):

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{y} / \mathbf{x}) \tag{A.2}$$

where $p(\mathbf{y}/\mathbf{x})$ is the probability density function of the measurement $\mathbf{y}$ for a fixed value of $\mathbf{x}$.

If the measurement errors, $\{n_i\}$, for $i = 1, \ldots, N$, are identically Gaussian distributed with zero-mean and variance σ^2 and furthermore if errors for different measurements are statistically independent, then (A.2) becomes:

$$\begin{aligned} \hat{\mathbf{x}} &= \arg\max_{\mathbf{x}} \frac{1}{(2\pi\sigma)^{N/2}} e^{-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|^2} \\ &= \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 \end{aligned} \tag{A.3}$$

The solution to (A.3) can readily be found by first differentiating $\|\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}\|^2$ with respect to $\hat{\mathbf{x}}$:

$$\frac{d}{d\hat{\mathbf{x}}}\left\|\mathbf{y}-\mathbf{H}\hat{\mathbf{x}}\right\|^2 = 2\mathbf{H}^T\mathbf{H}\hat{\mathbf{x}} - 2\mathbf{H}^T\mathbf{y} \tag{A.4}$$

and then setting this quantity equal to zero to obtain:

$$\hat{\mathbf{x}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y} \tag{A.5}$$

where it is assumed that the matrix inverse involved exists (i.e., that $\mathbf{H}^T\mathbf{H}$ is not singular).

The estimate described by (A.5) is referred to as a *least squares estimate* because, as shown in (A.3), it results in the minimum square error between the measurement vector $\mathbf{y}$ and $\mathbf{Hx}$, where the latter is the expected measurement vector based upon the estimate of $\mathbf{x}$.

Next consider the more general case where the measurement errors are still Gaussian distributed with zero-mean, but are not necessarily identically distributed or independent of each other. In this case, the maximum likelihood estimate can be expressed as

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\max_{\mathbf{x}} \frac{1}{(2\pi)^{N/2}\left|\mathbf{R}_\mathbf{n}\right|^{1/2}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{Hx})^T\mathbf{R}_\mathbf{n}^{-1}(\mathbf{y}-\mathbf{Hx})} \\ &= \arg\min_{\mathbf{x}} (\mathbf{y}-\mathbf{Hx})^T\mathbf{R}_\mathbf{n}^{-1}(\mathbf{y}-\mathbf{Hx})\end{aligned} \tag{A.6}$$

where $\mathbf{R}_\mathbf{n}$ is the covariance matrix associated with the measurement errors and $|\mathbf{R}_\mathbf{n}|$ is its determinant.

Proceeding as before, (A.6) can be solved to yield:

$$\hat{\mathbf{x}} = \left(\mathbf{H}^T\mathbf{R}_\mathbf{n}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}_\mathbf{n}^{-1}\mathbf{y} \tag{A.7}$$

The estimate in (A.7) is referred to as a *weighted least squares* solution.

Reference

[1] Stark, H., and J. W. Woods, *Probability, Random Processes, and Estimation Theory for Engineers*, Englewood Cliffs, NJ: Prentice-Hall, 1986.

APPENDIX B

Stability Measures for Frequency Sources

Lawrence F. Wiederholt and Willard A. Marquis

B.1 Introduction

The principle of employing satellite navigation systems for position and time determination requires the satellite clocks to be in synchronism to a common time base.

High-accuracy atomic frequency standards (AFSs) are required to meet the stringent stability and drift rates requirements so that the common time base can be maintained. Stability is also important for the less accurate crystal-based oscillators that are typically employed in user equipment.

Frequency sources are subject to systemic errors such as frequency offsets, aging, and random frequency errors. Random frequency errors are a primary concern, especially when characterizing the performance of an AFS. There are a number of important random frequency noise processes (i.e., frequency fluctuations): random walk frequency modulation, flicker frequency modulation, white frequency modulation, flicker phase modulation, and white phase modulation, as described in [1].

B.2 Frequency Standard Stability

The stability of a frequency source can be described by starting with an oscillator whose output voltage $V(t)$, is given by

$$V(t) = \left(V_0 + \varepsilon(t)\right)\left(\sin(2\pi\nu_0 t + \phi(t))\right) \tag{B.1}$$

where V_0 and ν_0 are the nominal amplitude and frequency, respectively, with corresponding errors $\varepsilon(t)$ and $\phi(t)$.

The instantaneous phase is defined by

$$\Phi(t) = 2\pi\nu_0 t + \phi(t) \tag{B.2}$$

and the instantaneous frequency is defined by

$$\nu(t) = \nu_0 + \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{B.3}$$

A common method used to measure oscillator stability is based upon the instantaneous fractional frequency deviation from the nominal frequency ν_0 given by

$$y(t) = \frac{\dot{\phi}}{2\pi\nu_0}$$

The power-law spectral densities of the five random frequency noise processes mentioned in Section B.1 can be represented in the frequency domain by the sum of five independent noise processes as [1]:

$$S_y(f) = \sum_{\alpha=-2}^{+2} h_\alpha f^\alpha \qquad for \quad 0 < f < f_h$$

$$= 0 \qquad for \quad f \geq f_h$$

where h_α is a constant, α is an integer, and f_h is the high-frequency cutoff of an infinitely sharp lowpass filter.

This power spectral density is visually represented in Figure B.1 for the five random frequency noise processes: random walk frequency, flicker frequency, white frequency, flicker phase, and white phase.

B.3 Measures of Stability

Two basic approaches can be taken to analyzing the stability of an oscillator: a frequency-domain approach and a time-domain approach. One can map from one to the other. The time-domain approach is more commonly used for stability analysis.

The interest in oscillators and the common measurement of their stability became such an item of interest that the IEEE Standards Committee 14 developed a standard in the 1980s. With this standard in place, oscillator stability evaluations could be performed on a common basis using standard definitions and evaluation technique. The latest revision of this standard was published in 2009 [1].

B.3.1 Allan Variance

One common measure of oscillator stability based on the instantaneous fractional frequency deviation is the Allan variance [2, 3], $\sigma_y^2(\tau)$, defined by

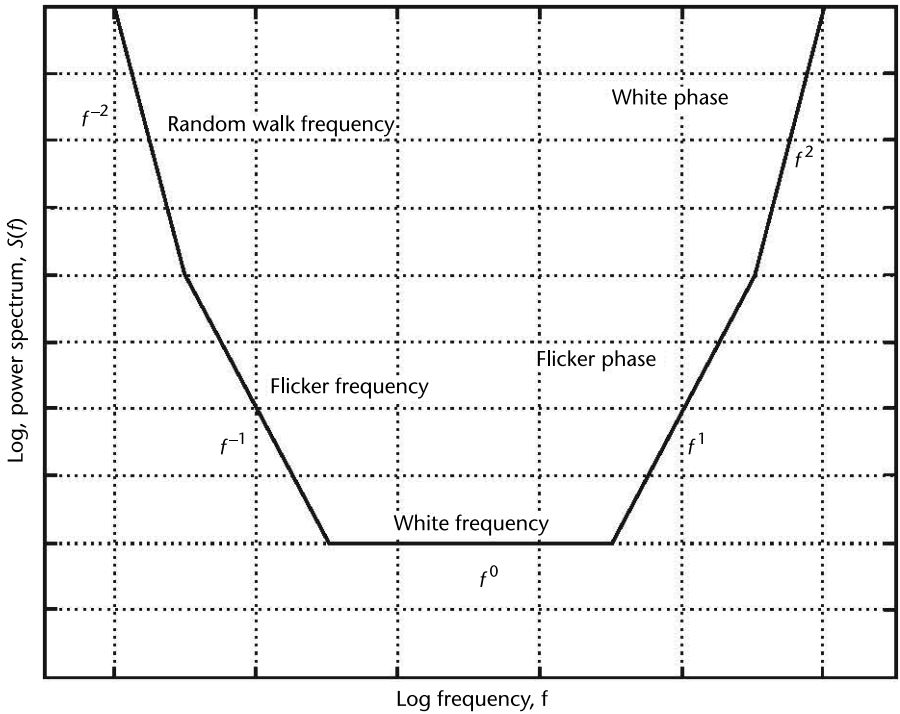


Figure B.1 Power spectral densities for five random frequency noise processes: random walk frequency, flicker frequency, white frequency, flicker phase, and white phase.

$$\sigma_y^2(\tau) = \frac{1}{2} E\left[(\bar{y}_{k+1} - \bar{y}_k)^2 \right]$$

where

$$\bar{y} = \frac{\phi(t_k + \tau) - \phi(t_k)}{2\pi \nu_0 \tau}$$

τ is the sampling interval and E is the expected value operator. In theory, E is an infinite sum of elements, but in practice, the sum is limited to a large but finite number.

The square root of the Allan variance is referred to as the Allan deviation.

B.3.2 Hadamard Variance

The Allan variance works well for cesium-based AFS with no linear drift effects. It is also often used to characterize the stability of quartz crystal oscillators. Rubidium-based AFS have a significant linear drift above the random noise, which degrades the fidelity of the Allan variance and thus does not provide an accurate measure of stability. The linear drift can be removed by a separate processing step, but an alternate measure of stability has been defined that overcomes this inherent limitation of the Allan variance. This measure is referred to as the Hadamard variance [4], which

removes any linear drift and is thus not effected by linear drift. Thus, the Hadamard variance is a good measure of stability for rubidium AFS.

The Hadamard variance, $H\sigma_y^2(\tau)$ is defined by

$$H\sigma_y^2(\tau) = \frac{1}{2}E\left[\left(\overline{y}_{k+2} - 2\overline{y}_{k+1} + \overline{y}_k\right)^2\right]$$

As in the Allan variance, E is the expected value operator. In theory, E is an infinite sum of elements, but in practice the sum is limited to a large but finite number.

Note that the Allan variance is a two-sample variance requiring two time sample values for each point, while the Hadamard variance is a three-sample variance requiring three time samples for each point. Thus, the Hadamard variance requires more computations.

For example, the GPS master control station uses the Hadamard variance and its variations to measure oscillator stability [5–7]. This is appropriate considering that, at the time of this writing, the constellation has a predominance of rubidium standards (Blocks IIR, IIR-M and IIF).

References

[1] "IEEE Standard Definitions of Physical Quantities for Fundamental Frequency and Time Metrology—Random Instabilities," IEEE Std. 1139–2008, IEEE Standards Coordinating Committee 27 on Time and Frequency, approved February 27, 2009.

[2] Allan, D., "Statistics of Atomic Frequency Standards," *Proceedings of the IEEE*, Vol. 54, pp. 221–230, February 1966.

[3] Walter, T., "Characterizing Frequency Stability—A Continuous Power-Law Model with Discrete Sampling," *IEEE Transactions on Instrumentation and Measurement*, 1994.

[4] Riley, W., "NIST Special Publication 1065, Handbook of Frequency Stability Analysis," National Institute of Standards and Technology, July 2008.

[5] Howe, D., et al., "Total Estimator of the Hadamard Function Used for GPS Operations," *Proceedings of the 32nd Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*, November 2000.

[6] Hutsell, S., "Relating the Hadamard Variance to MCS Kalman Filter Clock Estimation," *Proceedings of the 27th Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*, November 29–December 1, 1995, pp. 291–302.

[7] Hutsell, S., et al., "Operational Use of the Hadamard Variance in GPS," *Proceedings of the 28th Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*, December 1996, pp. 201–213.

APPENDIX C

Free-Space Propagation Loss

John W. Betz

C.1 Introduction

Calculating propagation loss is a fundamental tool in systems engineering for GNSS, as it is necessary to relate the power at a source (e.g., a satellite transmitter or an interferer) to the power at a destination (e.g., a GNSS receiver). The propagation loss typically depends on the distance between source and destination, as well other factors.

The simplest common expression for propagation loss is called *free-space propagation loss*, as it applies in free space (the source and receiver are located in a vacuum or equivalent, with no other objects in the vicinity). Although this expression is often employed, there are widespread misunderstandings of its applicability (under what conditions does it apply?) and its technical characteristics (e.g., in what sense is free-space propagation loss frequency-dependent?).

Entire texts, for example, [1], are devoted to radio wave propagation—predicting, measuring, and compensating for its effects. This appendix only touches on one simple and common model for radio wave propagation: free-space propagation loss. It also addresses a related topic—how to convert back and forth between power flux densities and power spectral densities.

C.2 Free-Space Propagation Loss

Propagation loss is defined as the ratio of the power transmitted in the direction of the receive antenna to the power at the terminals of a receive antenna, for a unity-gain receive antenna. If the receive antenna has gain other than unity, the received power is divided by the receive antenna gain in taking this ratio. The transmit

antenna actually radiates P_T watts, and has a gain of G_T (dimensionless), producing an effective isotropic radiated power (EIRP) of $P_T G_T$ watts. The receive antenna has a gain of G_R, the power at the receive antenna terminals is denoted P_R, so that the propagation loss is the dimensionless quantity

$$\Lambda = \frac{P_T G_T}{\left(P_R / G_R\right)} = \frac{P_T G_T G_R}{P_R}. \tag{C.1}$$

Because (C.1) is merely a definition, it could also be defined as the reciprocal of what is shown. The particular definition was selected so that the numerator is typically greater than the denominator, making the propagation loss usually a quantity greater than unity, or positive when expressed in decibels. This corresponds with common usage (e.g., "a 180-dB propagation loss").

It is often convenient to perform calculations for a receive antenna having unit gain ($G_R = 1$), calculating the received isotropic power (RIP).

The free-space propagation loss model described in this appendix applies when the transmitting antenna and receiving antenna are located in free space (ideally, a vacuum) where there are no other nearby conductive objects and no obstructions. In practice at L-band at least, it is sufficient that the line-of-sight path between transmitter and receiver is not obstructed, that there are no obstructions even near the line-of-sight path, and that the transmitter-to-receiver line-of-sight path is far from conducting surfaces, even the Earth's surface. If one of these conditions do occur, actual propagation loss may be much greater than predicted using the free-space model.

Further, the transmitting antenna and receiving antennas must be separated by many wavelengths so that they are not within each other's near fields. At L-band, wavelengths are less than half a meter, so several meters of separation is adequate for antennas having modest gain at L-band.

Detailed criteria for quantifying the conditions under which free-space propagation applies and ways to predict propagation losses under conditions other than free space can be found in [1] and are beyond the scope of this appendix. In many cases, free-space propagation is a good first-order model for L-band propagation from space to a terrestrial or airborne receiver, from an airborne transmitter to an airborne receiver, or from an airborne transmitter to the ground (or for same these paths with transmitter and receiver exchanged). These situations are clearly of interest to GNSS.

Consider a transmitter radiating an EIRP of $P_T G_T$. As the electromagnetic wave propagates, its power spreads out in a spherical pattern, so that the same amount of power remains in a given solid angle measured from the transmit antenna. However, the power flux density, which is the power per unit area in the shell of the sphere, diminishes as the radius of the sphere increases with increasing distance from the transmitter.

Now assume that the solid angle is small and the radius of the sphere is large enough that the solid angle can be approximated by a flat patch tangent to the sphere and thus normal to the line of sight between transmit antenna and receiver.

The effective area of an antenna, A, is given by

$$A = \frac{\lambda^2 G}{4\pi} \tag{C.2}$$

where $\lambda = c/f$ is the wavelength, with as the speed of propagation, f is the frequency, and G is the antenna gain. When the receive antenna gain is G_R, the effective area of the receive antenna is

$$A_R = \frac{G_R \lambda^2}{4\pi} = \frac{G_R c^2}{f^2 4\pi} \tag{C.3}$$

Observe that the effective area of an antenna having a given gain is inversely proportional to the square of the frequency. For the same antenna gain with increasing frequency, the antenna's area must become smaller.

Returning to the earlier discussion of an electromagnetic wave emanating outward from a transmitter, the power spatial density (having units of W/m^2) at a point on a sphere at radius d from the transmit antenna is

$$\Phi = \frac{P_T G_T}{4\pi d^2} \tag{C.4}$$

The power spatial density is also known as the *power flux density* (PFD). Observe that the PFD decreases with the square of the distance from the transmitter, so that the PFD (the received power per unit area) is independent of frequency and depends only on the distance from the transmitter.

The power at the receive antenna's terminals is given by the product of the PFD at the receive antenna and the effective area of the receive antenna

$$P_R = \Phi A_R \tag{C.5}$$

Substituting (C.3) and (C.4) into (C.5) yields

$$\begin{aligned} P_R &= \left(\frac{P_T G_T}{4\pi d^2}\right)\left(\frac{G_R \lambda^2}{4\pi}\right) \\ &= P_T G_T G_R \left(\frac{\lambda}{4\pi d}\right)^2 \end{aligned} \tag{C.6}$$

Expression (C.6), often called the Friis equation [2], allows calculation of the received power given the EIRP ($P_T G_T$) and the receive antenna gain (G_R). When (C.6) is calculated for an isotropic receive antenna, for which $G_R = 1$, the result is the RIP.

Sometimes the free-space propagation model is generalized to account for an excess propagation loss beyond the free-space loss. This excess propagation loss could be caused by attenuation due to the atmosphere, foliage penetration, building penetration, or polarization mismatch. The effect of this excess power loss is modeled by a dimensionless multiplicative factor L that takes on values between

unity and infinity, with unity indicating no excess loss, and infinity indicating complete blockage. As in the definition of propagation loss, L is defined to match common terminology (e.g., "an excess loss of 2 dB"). The resulting expression for received power is

$$P_R = \frac{P_T G_T G_R}{L}\left(\frac{\lambda}{4\pi d}\right)^2 \tag{C.7}$$

Computation of the received power is commonly performed in decibels. Denoting the quantities of units as superscripts allows (C.7) to be rewritten in decibels as

$$\begin{aligned}\left(P_R\right)_{dBW} &= \left(P_T\right)_{dBW} + \left(G_T\right)_{dB} + \left(G_R\right)_{dB} - L_{dB} + 20\log_{10}\left(\frac{\lambda}{4\pi d}\right) \\ &= \left(P_T\right)_{dBW} + \left(G_T\right)_{dB} + \left(G_R\right)_{dB} - L_{dB} - 21.98 - 20\log_{10}\left(\frac{d}{\lambda}\right)\end{aligned} \tag{C.8}$$

The latter expression is particularly simple, using a constant and the separation between transmitter and receiver expressed as the number of wavelengths.

Finally, the generalized free-space propagation loss (which includes excess loss) is found from (C.1) and (C.7) to be

$$\Lambda = L\left(\frac{4\pi d}{\lambda}\right)^2 \tag{C.9}$$

with $P_R = \frac{P_T G_T G_R}{\Lambda}$ and $(P_R)_{dBW} = (P_T)_{dBW} + (G_T)_{dB} + (G_R)_{dB} - \Lambda_{dB}$ where $\Lambda_{dB} = 10\log_{10}(\Lambda)$.

While (C.9) is a very compact expression for free-space propagation loss, simplistic interpretation of this expression leads to the faulty conclusion that, because free-space propagation loss increases with frequency, there is a frequency-dependent attenuation mechanism in free space. The correct interpretation is that the loss in PFD (in W/m^2) with distance from the transmitter does not depend on frequency, as seen in (C.4). However, free-space propagation loss is defined to include the effects of a receive antenna having a gain (often unity) that remains constant over frequency. Because an antenna of given gain has smaller effective area at higher frequencies, the fixed-gain antenna collects a smaller fraction of the power flux density at higher frequencies, resulting in lower received power at higher frequencies.

As the antenna area contributes to the free-space propagation loss as commonly defined, free-space propagation loss increases with frequency. If free-space propagation loss were instead defined for fixed effective area of the receive antenna rather than fixed gain of the receive antenna, (C.5) shows that the free-space propagation loss would then be independent of frequency (but the antenna would become increasingly directive at higher frequencies, since it would remain the same physical size).

C.3 Conversion Between Power Spectral Densities and Power Flux Densities

While PFDs arise often in documents involving spectrum protection and radio frequency interference, most signal theory is written in terms of PSDs. This section describes how to convert between the two quantities.

Recall that a PFD describes the power per unit area (often a square meter) in a propagating electromagnetic wave, while a PSD describes the power per bandwidth (often 1 Hz, but sometimes 1 kHz, 4 kHz, or 1 MHz) in a signal. These are very different concepts and quantities, and conversion between them requires an intermediate quantity—power—as well as definition of the receive antenna's effective area and of the normalized (unit power) power spectral density for the unit-power signal, in units of seconds (or reciprocal hertz).

To convert from PFD to PSD, first use (C.5) and the given effective area of the receive antenna. Often, a unity-gain antenna is assumed. [Note from (C.3) that at frequencies greater than $\frac{c}{\sqrt{4\pi}} \cong 84.3$ MHz, the effective area of a unity-gain antenna is less than unity, so for calculations involving GNSS, the effective area is typically negative when expressed in decibels.] The result is power, in units of watts. Multiply the power by the normalized power spectral density to obtain the actual power spectral density in units of W/Hz. To find the power spectral density in a given bandwidth centered at a given center frequency, merely integrate the actual power spectral density over that bandwidth at that frequency. In many cases, the latter step can be approximated by evaluating the PSD at the center frequency, and then multiplying it by the bandwidth. As long as the actual power spectral density is well approximated by a straight (not necessarily horizontal) line over the given bandwidth, the result is valid.

To convert from PSD to PFD, integrate the PSD over all frequencies to determine the total power. Then, using (C.5), divide the total power by the effective area of the receive antenna (for frequencies of typical interest in GNSS, this involves adding a positive quantity in decibels) to obtain the PFD.

References

[1] Parsons, J. D., *The Mobile Radio Propagation Channel*, 2nd ed., New York: John Wiley and Sons, 2000.

[2] Friis, H. T., "A Note on a Simple Transmission Formula," *Proceedings IRE*, Vol. 34, 1946, pp. 254–256.

Index

J

K

L

M

N

Q

R

S

T