CHAPTER 11

SATELLITE NAVIGATION

INTRODUCTION

1100. Early Developments In Satellite Navigation

The idea that led to development of the satellite navigation systems dates back to 1957 and the first launch of an artificial satellite into orbit, Russia's Sputnik I. Dr. William H. Guier and Dr. George C. Wieffenbach at the Applied Physics Laboratory of the Johns Hopkins University were monitoring the famous "beeps" transmitted by the passing satellite. They plotted the received signals at precise intervals, and noticed that a characteristic Doppler curve emerged. Since celestial bodies followed fixed orbits, they reasoned that this curve could be used to describe the satellite orbit. Later, they demonstrated that they could determine all of the orbital parameters for a passing satellite by doppler observation of a single pass from a single fixed station. The doppler shift apparent while receiving a transmission from a passing satellite proved to be an effective measuring device for establishing the satellite orbit.

Dr. Frank T. McClure, also of the Applied Physics Laboratory, reasoned that if the satellite orbit was known, doppler shift measurements could be used to determine one's position on earth. His studies in support of this hypothesis earned him the first National Aeronautics and Space Administration award for important contributions to space development.

In 1958, the Applied Physics Laboratory proposed exploring the possibility of an operational satellite doppler navigation system. The Chief of Naval Operations then set forth requirements for such a system. The first successful launching of a prototype system satellite in April 1960 demonstrated the doppler system's operational feasibility.

1101. NAVSAT, The First Satellite Navigation System

The **Navy Navigation Satellite System** (**NAVSAT**, also known as **TRANSIT**) was the first operational satellite navigation system. The system's accuracy was better than 0.1 nautical mile anywhere in the world. It was used primarily for the navigation of surface ships and submarines; but it also had some applications in air navigation. It was also used in hydrographic surveying and geodetic position determination.

NAVSAT uses the doppler shift of radio signals transmitted from a satellite to measure the relative velocity between the satellite and the navigator. Knowing the satellite orbit precisely, the navigator's absolute position can be accurately determined from the time rate of change of range to the satellite.

The Johns Hopkins University Applied Physics Laboratory developed NAVSAT for the U. S. Navy. The operation of the system is under the control of the U. S. Navy Astronautics Group with headquarters at Point Mugu, California.

1102. System Configuration, Operation, And Termination

The NAVSAT consists of 10 orbiting satellites and 3 orbiting spares; a network of tracking stations continuously monitoring the satellites and updating the information they transmit; and the receivers and computers for processing signals.

Each satellite is in a nominally circular polar orbit at an approximate altitude of 600 nautical miles. There are usually five satellites operating in the system. Five satellites in orbit provide redundancy; the minimum constellation for system operation is four. This redundancy allows for an unexpected failure of a satellite and the relatively long period of time required to schedule, prepare, and launch a replacement satellite. This redundancy also provides for turning off a satellite when (on rare occasions) its orbital plane precesses near another satellite's plane, or when the timing (phasing) of several satellites in their orbits are temporarily such that many satellites pass nearly simultaneously near one of the poles.

Each satellite contains: (1) receiver equipment to accept injection data and operational commands from the ground, (2) a decoder for digitizing the data, (3) switching logic and memory banks for sorting and storing the digital data, (4) control circuits to cause the data to be read out at specific times in the proper format, (5) an encoder to translate the digital data to phase modulation, (6) ultra stable 5 MHz oscillators, and (7) 1.5-watt transmitters to broadcast the 150- and 400-MHz oscillator-regulated frequencies that carry the data to earth.

The transit launch program ended in 1988. According to the Federal Radionavigation Plan, the Navy will cease operation of NAVSAT by the end of 1996, as the new Global Positioning System (GPS) comes into operation.

THE GLOBAL POSITIONING SYSTEM

1103. Basic System Description

The Federal Radionavigation Plan has designated the Navigation System using Timing and Ranging (NAVSTAR) Global Positioning System (GPS) as the primary navigation system of the U.S. government. GPS is a spaced-based radio positioning system which provides suitably equipped users with highly accurate position, velocity, and time data. It consists of three major segments: a **space segment**, a **control segment**, and a **user segment**.

The space segment contains 24 satellites. Precise spacing of the satellites in orbit is arranged such that a minimum of four satellites are in view to a user at any time on a worldwide basis. Each satellite transmits signals on two radio frequencies, superimposed on which are navigation and system data. Included in this data is predicted satellite ephemeris, atmospheric propagation correction data, satellite clock error information, and satellite health data. This segment consists of 21 operational satellites with three satellites orbiting as active spares. The satellites orbit in six separate orbital planes. The orbital planes have an inclination relative to the equator of 55° and an orbital height of 20,200 km. The satellites complete an orbit approximately once every 12 hours.

GPS satellites transmit **pseudorandom noise (PRN)** sequence-modulated radio frequencies, designated L1 (1575.42 MHz) and L2 (1227.60 MHz). The satellite transmits both a **Coarse Acquisition Code** (C/A code) and a **Precision Code** (P code). Both the P and C/A codes are transmitted on the L1 carrier; only the P code is transmitted on the L2 carrier. Superimposed on both the C/A and P codes is the Navigation message. This message contains satellite ephemeris data, atmospheric propagation correction data, and satellite clock bias.

GPS assigns a unique C/A code and a unique P code to each satellite. This practice, known as **code division multiple access (CDMA)**, allows all satellites the use of a common carrier frequency while still allowing the receiver to determine which satellite is transmitting. CDMA also allows for easy user identification of each GPS satellite. Since each satellite broadcasts using its own unique C/A and P code combination, it can be assigned a unique **PRN sequence number**. This number is how a satellite is identified when the GPS control system communicates with users about a particular GPS satellite.

The control segment includes a **master control station** (MCS), a number of monitor stations, and ground antennas located throughout the world. The master control station, located in Colorado Springs, Colorado, consists of equipment and facilities required for satellite monitoring, telemetry, tracking, commanding, control, uploading, and navigation message generation. The monitor stations, located in Hawaii, Colorado Springs, Kwajalein, Diego Garcia, and Ascension Island, passively track the satellites, accumulating ranging data from the satellites' signals and relaying them to the MCS. The MCS processes this information to determine satellite position and signal data accuracy, updates the navigation message of each satellite and relays this information to the ground antennas. The ground antennas then transmit this information to the satellites. The ground antennas, located at Ascension Island, Diego Garcia, and Kwajalein, are also used for transmitting and receiving satellite control information.

The user segment is designed for different requirements of various users. These receivers can be used in high, medium, and low dynamic applications. An example of a low dynamic application would be a fixed antenna or slowly drifting marine craft. An example of a medium dynamic application would be a marine or land vehicle traveling at a constant controlled speed. Finally, an example of a high dynamic application would be a high performance aircraft or a spacecraft. The user equipment is designed to receive and process signals from four or more orbiting satellites either simultaneously or sequentially. The processor in the receiver then converts these signals to three-dimensional navigation information based on the World Geodetic System 1984 reference ellipsoid. The user segment can consist of stand-alone receivers or equipment that is integrated into another navigation system. Since GPS is used in a wide variety of applications, from marine navigation to land surveying, these receivers can vary greatly in function and design.

1104. System Capabilities

GPS provides multiple users with accurate, continuous, worldwide, all-weather, common-grid, three-dimensional positioning and navigation information.

To obtain a navigation solution of position (latitude, longitude, and altitude) and time (four unknowns), four satellites must be selected. The GPS user measures pseudorange and pseudorange rate by synchronizing and tracking the navigation signal from each of the four selected satellites. Pseudorange is the true distance between the satellite and the user plus an offset due to the user's clock bias. Pseudorange rate is the true slant range rate plus an offset due to the frequency error of the user's clock. By decoding the ephemeris data and system timing information on each satellite's signal, the user's receiver/processor can convert the pseudorange and pseudorange rate to three-dimensional position and velocity. Four measurements are necessary to solve for the three unknown components of position (or velocity) and the unknown user time (or frequency) bias.

The navigation accuracy that can be achieved by any user depends primarily on the variability of the errors in making pseudorange measurements, the instantaneous geometry of the satellites as seen from the user's location on Earth, and the presence of **Selective Avaliability** (SA). Selective Availability is discussed further below.

1105. Global Positioning System Basic Concepts

As discussed above, GPS measures distances between satellites in orbit and a receiver on or above the earth and computes spheres of position from those distances. The intersections of those spheres of position then determine the receiver's position.

The distance measurements described above are done by comparing timing signals generated simultaneously by the satellites' and receiver's internal clocks. These signals, characterized by a special wave form known as the pseudo-random code, are generated in phase with each other. The signal from the satellite arrives at the receiver following a time delay proportional to its distance traveled. This time delay is detected by the phase shift between the received pseudo-random code and the code generated by the receiver. Knowing the time required for the signal to reach the receiver from the satellite allows the receiver to calculate the distance from the satellite. The receiver, therefore, must be located on a sphere centered at the satellite with a radius equal to this distance measurement. The intersection of three spheres of position yields two possible points of receiver position. One of these points can be disregarded since it is hundreds of miles from the surface of the earth. Theoretically, then, only three time measurements are required to obtain a fix from GPS.

In practice, however, a fourth measurement is required to obtain an accurate position from GPS. This is due to receiver clock error. Timing signals travel from the satellite to the receiver at the speed of light; even extremely slight timing errors between the clocks on the satellite and in the receiver will lead to tremendous range errors. The satellite's atomic clock is accurate to 10^{-9} seconds; installing a clock that accurate on a receiver would make the receiver prohibitively expensive. Therefore, receiver clock accuracy is sacrificed, and an additional satellite timing measurement is made. The fix error caused by the inaccuracies in the receiver clock is reduced by simultaneously subtracting a constant timing error from four satellite timing measurements until a pinpoint fix is reached. This process is analogous to the navigator's plotting of a visual fix when bearing transmission error is present in his bearing repeater system. With that bearing error present, two visual LOP's will not intersect at a vessel's true position; there will be an error introduced due to the fixed, constant error in the bearing transmission process. There are two ways to overcome such an error. The navigator can buy extremely accurate (and expensive) bearing transmission and display equipment, or he can simply take a bearing to a third visual navigation aid. The resulting fix will not plot as a pinpoint (as it would were there no transmission error present); rather, it will plot as a triangle. The navigator can then apply a constant bearing correction to each LOP until the correction applied equals the bearing transmission error. When the correction applied equals the original transmission error, the resultant fix should plot as a pinpoint. The situation with GPS receiver timing inaccuracies is analogous; time measurement error simply replaces bearing measurement error in the analysis. Assuming that the satellite clocks are perfectly synchronized and the receiver clock's error is constant, the subtraction of that constant error from the resulting distance determinations will reduce the fix error until a "pinpoint" position is obtained. It is important to note here that the number of lines of position required to employ this technique is a function of the number of lines of position required to obtain a fix. In the two dimensional visual plotting scenario described above, only two LOP's were required to constitute a fix. The bearing error introduced another unknown into the process, resulting in three total unknowns (the x coordinate of position, the y coordinate of position, and the bearing error). Because of the three unknowns, three LOP's were required to employ this correction technique. GPS determines position in three dimensions; the presence of receiver clock error adds an additional unknown. Therefore, four timing measurements are required to solve for the resulting four unknowns.

1106. GPS Signal Coding

Two separate carrier frequencies carry the signal transmitted by a GPS satellite. The first carrier frequency (L1) transmits on 1575.42 MHz; the second (L2) transmits on 1227.60 MHz. The GPS signal consists of three separate messages: the P-code, transmitted on both L1 and L2; the C/A code, transmitted on L1 only; and a navigation data message. The P code and C/A code messages are divided into individual bits known as **chips**. The frequency at which bits are sent for each type of signal is known as the **chipping rate**. The chipping rate for the P-code is 10.23 MHz (10.23×10^6 bits per second); for the C/A code, 1.023 MHz (1.023×10^6 bits per second); and for the data message, 50 Hz (50 bits per second). The P and C/A codes **phase modulate** the carriers; the C/A code is transmitted at a phase angle of 90° from the P code. The periods of repetition for the C/A and P codes differ. The C/A code repeats once every millisecond; the P-code sequence repeats every seven days.

As stated above the GPS carrier frequencies are phase modulated. This is simply another way of saying that the digital "1's" and "0's" contained in the P and C/A codes are indicated along the carrier by a shift in the carrier phase. This is analogous to sending the same data along a carrier by varying its amplitude (amplitude modulation, or AM) or its frequency (frequency modulation, or FM). See Figure 1106a. In phase modulation, the frequency and the amplitude of the carrier are unchanged by the "information signal," and the digital information is transmitted by shift-

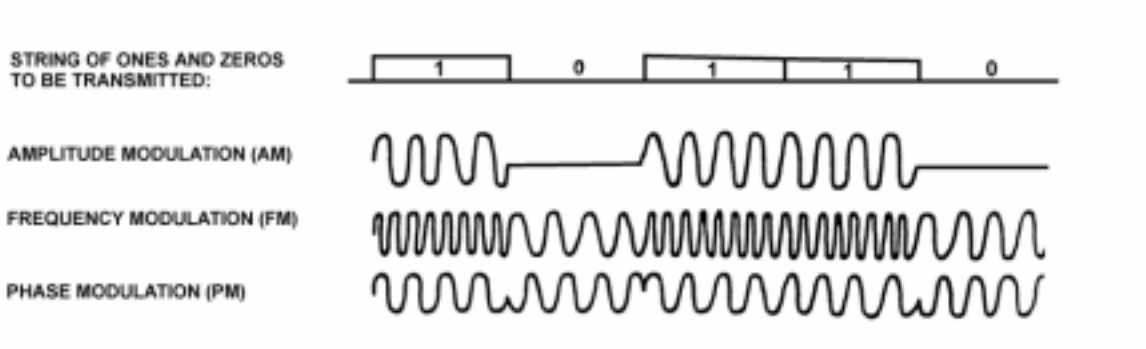

Figure 1106a. Digital data transmission with amplitude, frequency and phase modulation.

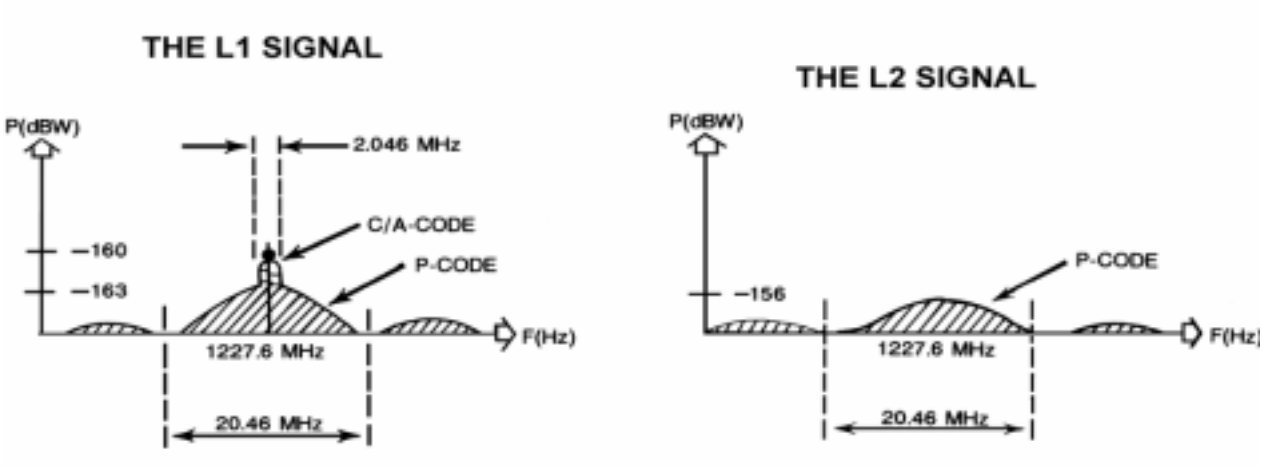

Figure 1106b. Modulation of the L1 and L2 carrier frequencies with the C/A and P code signals.

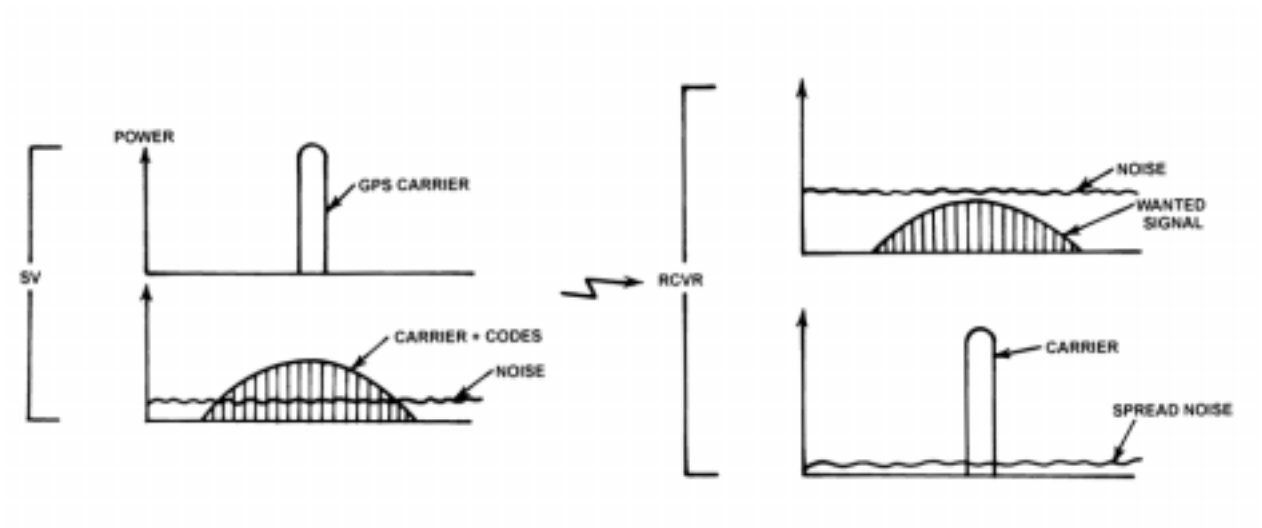

Figure 1106c. GPS signal spreading and recovery from satellite to receiver.

ing the carrier's phase. The phase modulation employed by GPS is known as bi-phase shift keying (BPSK).

Due to this BPSK, the carrier frequency is "spread" about its center frequency by an amount equal to twice the "chipping rate" of the modulating signal. In the case of the P code, this spreading is equal to (2×10.23 MHz) = 20.46 MHz. For the C/A code, the spreading is equal to (2×1.023 MHz) = 2.046 MHz. See Figure 1106b. Note that the L1 carrier signal, modulated with both the P code and C/A code, is shaped differently from the L2 carrier, modulated with only the P code. This spreading of the carrier signal lowers the total signal strength below the thermal noise threshold present at the receiver. This effect is demonstrated in Figure 1106c. When the satellite signal is multiplied with the C/A and P codes generated by the receiver, the satellite signal will be collapsed into the original carrier frequency band. The signal power is then raised above the thermal noise level.

The navigation message is superimposed on both the P code and C/A code with a data rate of 50 bits per second (50 Hz.) The navigation message consists of 25 data frames, each frame consisting of 1500 bits. Each frame is divided into five subframes of 300 bits each. It will, therefore, take 30 seconds to receive one data frame and 12.5 minutes to receive all 25 frames. The navigation message contains GPS system time of transmission; a **hand over word (HOW)**, allowing the transition between tracking the C/A code to the P code; ephemeris and clock data for the satellite being tracked; and almanac data for the satellites in orbit. It also contains coefficients for ionospheric delay models used by C/A receivers and coefficients used to calculate Universal Coordinated Time (UTC).

1107. The Correlation Process

The correlation process compares the signal received with the signal generated internal to the receiver. It does this by comparing the square wave function of the received signal with the square wave function generated by the receiver. The computer logic of the receiver recognizes the square wave signals as either a +1 or a 0 depending on whether the signal is "on" or "off." The signals are processed and matched by using an **autocorrelation function**.

This process defines the necessity for a "pseudo-random code." The code must be repeatable (i.e., non-random) because it is in comparing the two signals that the receiver makes its distance calculations. At the same time, the code must be random for the correlation process to work; the randomness of the signals must be such that the matching process excludes all possible combinations except the combination that occurs when the generated signal is shifted a distance proportional to the received signal's time delay. These simultaneous requirements to be both repeatable (non-random) and random give rise to the description of "pseudo-random"; the signal has enough repeatability to enable the receiver to make the required measurement while simultaneously retaining enough randomness to ensure incorrect calculations are excluded.

1108. Precise Positioning Service And Standard Positioning Service

Two levels of navigational accuracy are provided by the GPS: the **Precise Positioning Service (PPS)** and the **Standard Positioning Service (SPS)**. GPS was designed, first and foremost, by the U.S. Department of Defense as a United States military asset; its extremely accurate positioning capability is an asset access to which the U.S. military would like to limit during time of war. Therefore, the PPS is available only to authorized users, mainly the U.S. military and authorized allies. SPS, on the other hand, is available worldwide to anyone possessing a GPS receiver. PPS, therefore, provides a more accurate position than does SPS.

Two cryptographic methods are employed to deny the PPS accuracy to civilian users: **selective availability (SA)**

SA/A-S Configuration	SIS Interface Conditions	PPS Users	SPS Users
SA Set to Zero A-S Off	P-Code, no errors C/A-Code, no errors	Full accuracy, spoofable	Full accuracy,* spoofable
SA at Non-Zero Value A-S Off	P-Code, errors C/A-Code, errors	Full accuracy, spoofable	Limited accuracy, spoofable
SA Set to Zero A-S On	Y-Code, no errors C/A-Code, no errors	Full accuracy, Not spoofable**	Full accuracy,*** spoofable
SA at Non-Zero Value A-S On	Y-Code, errors C/A-Code, errors	Full accuracy, Not spoofable**	Limited accuracy, spoofable

* "Full accuracy" defined as equivelent to a PPS-capable UE operated in a similar manner.
** Certain PPS-capable UE do not have P- or Y-code tracking abilities and remain spoofable despite A-S protection being applied
*** Assuming negligable accuracy degradation due to C/A-code operation (but more susceptible to jamming).

Figure 1108. Effect of SA and A-S on GPS accuracy.

and **anti-spoofing (A-S)**. SA operates by introducing controlled errors into both the C/A and P code signals. SA can be programmed to degrade the signals' accuracy even further during time of war, denying a potential adversary the ability to use GPS to nominal SPS accuracy. SA introduces two errors into the satellite signal: (1) The **epsilon error**: an error in satellite ephemeris data in the navigation message; and (2) **clock dither**: error introduced in the satellite atomic clocks' timing. The presence of SA is the largest source of error present in an SPS GPS position measurement.

Anti-spoofing is designed to negate any hostile imitation of GPS signals. The technique alters the P code into another code, designated the Y code. The C/A code remains unaffected. The U.S. employs this technique to the satellite signals at random times and without warning; therefore, civilian users are unaware when this P code transformation takes place. Since anti-spoofing is applied only to the P code, the C/A code is not protected and can be spoofed.

Only users employing the proper cryptographic devices can defeat both SA and anti-spoofing. Without these devices, the user will be subject to the accuracy degradation of SA and will be unable to track the Y code.

GPS PPS receivers can use either the P code or the C/A code, or both, in determining position. Maximum accuracy is obtained by using the P code on both L1 and L2. The difference in propagation delay is then used to calculate ionospheric corrections. The C/A code is normally used to acquire the satellite signal and determine the approximate P code phase. Then, the receiver locks on the P code for precise positioning (subject to SA if not cryptographically equipped). Some PPS receivers possess a clock accurate enough to track and lock on the P code signal without initially tracking the C/A code. Some PPS receivers can track only the C/A code and disregard the P code entirely. Since the C/A code is transmitted on only one frequency, the dual frequency ionosphere correction methodology is unavailable and a ionospheric modeling procedure is required to calculate the required corrections.

SPS receivers, as mentioned above, provide positions with a degraded accuracy. The A-S feature denies SPS users access to the P code when transformed to the Y code. Therefore, the SPS user cannot rely on access to the P code to measure propagation delays between L1 and L2 and compute ionospheric delay corrections. Consequently, the typical SPS receiver uses only the C/A code because it is unaffected by A-S. Since C/A is transmitted only on L1, the dual frequency method of calculating ionospheric corrections is unavailable; an ionospheric modeling technique must be used. This is less accurate than the dual frequency method; this degradation in accuracy is accounted for in the 100 meter accuracy calculation. Figure 1108 presents the effect on SA and A-S on different types of GPS measurements.

1109. GPS Receiver Operations

In order for the GPS receiver to navigate, it has to track satellite signals, make pseudorange measurements, and collect navigation data.

A typical satellite tracking sequence begins with the receiver determining which satellites are available for it to track. Satellite visibility is determined by user-entered predictions of position, velocity, and time, and by almanac information stored internal to the receiver. If no stored almanac information exists, then the receiver must attempt to locate and lock onto the signal from any satellite in view. When the receiver is locked onto a satellite, it can demodulate the navigation message and read the almanac information about all the other satellites in the constellation. A carrier tracking loop tracks the carrier frequency while a code tracking loop tracks the C/A and P code signals. The two tracking loops operate together in an iterative process to acquire and track satellite signals.

The receiver's carrier tracking loop will locally generate an L1 carrier frequency which differs from the satellite produced L1 frequency due to a doppler shift in the received frequency. This doppler offset is proportional to the relative velocity along the line of sight between the satellite and the receiver, subject to a receiver frequency bias. The carrier tracking loop adjusts the frequency of the receiver-generated frequency until it matches the incoming frequency. This determines the relative velocity between the satellite and the receiver. The GPS receiver uses this relative velocity to calculate the velocity of the receiver. This velocity is then used to aid the code tracking loop.

The code tracking loop is used to make pseudorange measurements between the GPS receiver and the satellites. The receiver's tracking loop will generate a replica of the targeted satellite's C/A code with estimated ranging delay. In order to match the received signal with the internally generated replica, two things must be done: 1) The center frequency of the replica must be adjusted to be the same as the center frequency of the received signal; and 2) the phase of the replica code must be lined up with the phase of the received code. The center frequency of the replica is set by using the doppler-estimated output of the carrier tracking loop. The receiver will then slew the code loop generated C/A code though a millisecond search window to correlate with the received C/A code and obtain C/A tracking.

Once the carrier tracking loop and the code tracking loop have locked onto the received signal and the C/A code has been stripped from the carrier, the navigation message is demodulated and read. This gives the receiver other information crucial to a pseudorange measurement. The navigation message also gives the receiver the handover word, the code that allows a GPS receiver to shift from C/A code tracking to P code tracking.

The handover word is required due to the long phase (seven days) of the P code signal. The C/A code repeats every millisecond, allowing for a relatively small search window. The seven day repeat period of the P code requires that the receiver be given the approximate P code phase to narrow its search window to a manageable time. The handover word pro-

vides this P code phase information. The handover word is repeated every subframe in a 30 bit long block of data in the navigation message. It is repeated in the second 30 second data block of each subframe. For some receivers, this handover word is unnecessary; they can acquire the P code directly. This normally requires the receiver to have a clock whose accuracy approaches that of an atomic clock. Since this greatly increases the cost of the receiver, most receivers for non-military marine use do not have this capability.

Once the receiver has acquired the satellite signals from four GPS satellites, achieved carrier and code tracking, and has read the navigation message, the receiver is ready to begin making pseudorange measurements. Recall that these measurements are termed *pseudorange* because a receiver clock offset makes them inaccurate; that is, they do not represent the true range from the satellite, only a range biased by a receiver clock error. This clock bias introduces a fourth unknown into the system of equations for which the GPS receiver must solve (the other three being the x coordinate, y coordinate, and z coordinate of the receiver position). Recall from the discussion in section 1103 that the receiver solves this clock bias problem by making a fourth pseudorange measurement, resulting in a fourth equation to allow solving for the fourth unknown. Once the four equations are solved, the receiver has an estimate of the receiver's position in three dimensions and of GPS time. The receiver then converts this position into coordinates referenced to an earth model based on the World Geodetic System (1984).

1110. User Range Errors And Geometric Dilution Of Precision

There are two formal position accuracy requirements for GPS:

1) The PPS spherical position accuracy shall be 16 meters SEP (spherical error probable) or better.

2) The SPS user two dimensional position accuracy shall be 100 meters 2 drms or better.

Assume that a universal set of GPS pseudorange measurements results in a set of GPS position measurements. The accuracy of these measurements will conform to a normal (i.e. values symmetrically distributed around a mean of zero) probability function because the two most important factors affecting accuracy, the **geometric dilution of precision (GDOP)** and the **user equivalent range error (UERE)**, are continuously variable.

The UERE is the error in the measurement of the pseudoranges from each satellite to the user. The UERE is the product of several factors, including the clock stability, the predictability of the satellite's orbit, errors in the 50 Hz navigation message, the precision of the receiver's correlation process, errors due to atmospheric distortion and the calculations to compensate for it, and the quality of the satellite's signal. The UERE, therefore, is a random error which is the function of errors in both the satellites and the user's receiver.

The GDOP depends on the geometry of the satellites in relation to the user's receiver. It is independent of the quality of the broadcast signals and the user's receiver. Generally speaking, the GDOP measures the "spread" of the satellites around the receiver. The optimum case would be to have one satellite directly overhead and the other three spaced 120° around the receiver on the horizon. The worst GDOP would occur if the satellites were spaced closely together or in a line overhead.

There are special types of DOP's for each of the posi-

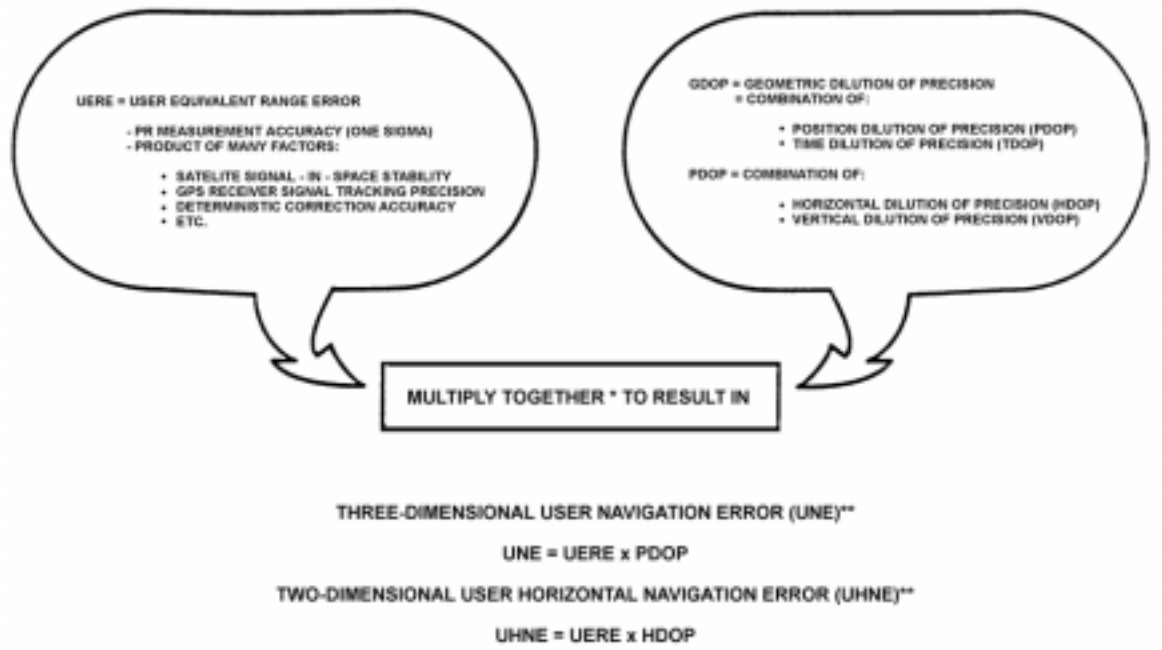

Figure 1110. Position and time error computations.

tion and time solution dimensions; these particular DOP's combine to determine the GDOP. For the vertical dimension, the **vertical dilution of precision (VDOP)** describes the effect of satellite geometry on altitude calculations. The **horizontal dilution of precision (HDOP)** describes satellite geometry's effect on position (latitude and longitude) errors. These two DOP's combine to determine the **position dilution of precision (PDOP)**. The PDOP combined with the **time dilution of precision (TDOP)** results in the GDOP. See Figure 1110.

1111. Ionospheric Delay Errors

Section 1107 covered errors in GPS positions due to errors inherent in the satellite signal (UERE) and the geometry of the satellite constellation (GDOP). Another major cause of accuracy degradation is the effect of the ionosphere on the radio frequency signals that comprise the GPS signal.

A discussion of a model of the earth's atmosphere will be useful in understanding this concept. Consider the earth as surrounded by three layers of atmosphere. The first layer, extending from the surface of the earth to an altitude of approximately 10 km, is known as the troposphere. Above the troposphere and extending to an altitude of approximately 50 km is the stratosphere. Finally, above the stratosphere and extending to an altitude that varies as a function of the time of day is the **ionosphere**. Though radio signals are subjected to effects which degrade its accuracy in all three layers of this atmospheric model, the effects of the ionosphere are the most significant; therefore, they will be discussed here.

The ionosphere, as the name implies, is that region of the atmosphere which contains a large number of ionized molecules and a correspondingly high number of free electrons. These charged molecules are those which have lost one or more electrons. No atom will loose an electron without an input of energy; the energy input that causes the ions to be formed in the ionosphere comes from the ultraviolet (U-V) radiation of the sun. Therefore, the more intense the sun's rays, the larger the number of free electrons which will exist in this region of the atmosphere.

The largest effect that this ionospheric effect has on GPS accuracy is a phenomenon known as **group time delay**. As the name implies, group time delay results in a delay in the time a signal takes to travel through a given distance. Obviously, since GPS relies on extremely accurate timing measurement of these signals between satellites and ground receivers, this group time delay can have a noticeable effect on the magnitude of GPS position error.

The group time delay is a function of several elements. It is inversely proportional to the square of the frequency at which the satellite transmits, and it is directly proportional to the atmosphere's **total electron content (TEC)**, a measure of the degree of the atmosphere's ionization. The general form of the equation describing the delay effect is:

$$\Delta t = \frac{(K \times TEC)}{f^2}$$

where

Δt = group time delay
f = operating frequency
K = constant

Since the sun's U-V radiation ionizes the molecules in the upper atmosphere, it stands to reason that the time delay value will be highest when the sun is shining and lowest at night. Experimental evidence has borne this out, showing that the value for TEC is highest around 1500 local time and lowest around 0500 local time. Therefore, the magnitude of the accuracy degradation caused by this effect will be highest during daylight operations. In addition to these daily variations, the magnitude of this time delay error also varies with the seasons; it is highest at the vernal equinox. Finally, this effect shows a solar cycle dependence. The greater the number of sunspots, the higher the TEC value and the greater the group time delay effect. The solar cycle typically follows an eleven year pattern. Solar cycle 22 began in 1986, peaked in 1991, and is now in decline. It should reach a minimum in 1997, at which time the effect on the group time delay from this phenomenon will also reach a minimum.

Given that this ionospheric delay introduces a serious accuracy degradation into the system, how does GPS account for it? There are two methods used: (1) the dual frequency technique, and (2) the ionospheric delay method.

1112. Dual Frequency Correction Technique

As the term implies, the dual frequency technique requires the ability to acquire and track both the L1 and L2 frequency signals. Recall from the discussion in section 1105 that the C/A and P codes are transmitted on carrier frequency L1, but only the P code is transmitted on L2. Recall also from section 1105 that only authorized operators with access to DOD cryptographic material are able to copy the P code. It follows, then, that only those authorized users are able to copy the L2 carrier frequency. Therefore, only those authorized users are able to use the dual frequency correction method. The dual frequency method measures the distance between the satellite and the user based on both the L1 and L2 carrier signal. These ranges will be different because the group time delay for each signal will be different. This is because of the frequency dependence of the time delay error. The range from the satellite to the user will be the true range combined with the range error caused by the time delay, as shown by the following equation:

$$R(f) = R_{actual} + \text{error term}$$

where R(f) is the range which differs from the actual range as a function of the carrier frequency. The dual frequency correction method takes two such range measurements, R(L1) and R(L2). Recall that the error term is a function of a constant divided by the square of the frequency. By combining the two range equations derived from the two frequency measurements, the constant term can be eliminated and one is left with an equation in which the true range is simply a function of the two carrier frequencies and the measured ranges R(L1) and R(L2). This method has two major advantages over the ionospheric model method. (1) It calculates corrections from real-time measured data; therefore, it is more accurate. (2) It alleviates the need to include ionospheric data on the navigation message. A significant portion of the data message is devoted to ionospheric correction data. If the receiver is dual frequency capable, then it does not need any of this data.

The vast majority of maritime users cannot copy dual frequency signals. For them, the ionospheric delay model provides the correction for the group time delay.

1113. The Ionospheric Delay Model

The ionospheric delay model mathematically models the diurnal ionospheric variation. The value for this time delay is determined from a cosinusoidal function into which coefficients representing the maximum value of the time delay (i.e., the amplitude of the cosine wave representing the delay function); the time of day; the period of the variation; and a minimum value of delay are introduced. This model is designed to be most accurate at the diurnal maximum. This is obviously a reasonable design consideration because it is at the time of day when the maximum diurnal time delay occurs that the largest magnitude of error appears. The coefficients for use in this delay model are transmitted to the receiver in the navigation data message. As stated in section 1112, this method of correction is not as accurate as the dual frequency method; however, for the non-military user, it is the only method of correction available.

1114. Multipath Reflection Errors

Multipath reflection errors occur when the receiver detects parts of the same signal at two different times. The first reception is the direct path reception, the signal that is received directly from the satellite. The second reception is from a reflection of that same signal from the ground or any other reflective surface. The direct path signal arrives first, the reflected signal, having had to travel a longer distance to the receiver, arrives later. The GPS signal is designed to minimize this multipath error. The L1 and L2 frequencies used demonstrate a diffuse reflection pattern, lowering the signal strength or any reflection that arrives at the receiver. In addition, the receiver's antenna can be designed to reject a signal that it recognizes as a reflection. In addition to the properties of the carrier frequencies, the high data frequency of both the P and C/A codes and their resulting good correlation properties minimize the effect of multipath propagation.

The design features mentioned above combine to reduce the maximum error expected from multipath propagation to less than 20 feet.

DIFFERENTIAL GPS

1115. Differential GPS Concept

The discussions above make it clear that the Global Positioning System provides the most accurate positions available to navigators today. They should also make clear that the most accurate positioning information is available to only a small fraction of the using population: U.S. and allied military. For most open ocean navigation applications, the degraded accuracy inherent in selective availability and the inability to copy the precision code presents no serious hazard to navigation. A mariner seldom if ever needs greater than 100 meter accuracy in the middle of the ocean.

It is a different situation as the mariner approaches shore. Typically for harbor approaches and piloting, the mariner will shift to visual piloting. The increase in accuracy provided by this navigational method is required to ensure ship's safety. The 100 meter accuracy of GPS in this situation is not sufficient. Any mariner who has groped his way through a restricted channel, in a fog obscuring all visual navigation aids will certainly appreciate the fact that even a degraded GPS position is available for them to plot. However, 100 meter accuracy is not sufficient to ensure ship's safety in most piloting situations. In this situation, the mariner needs P code accuracy. The problem then becomes how to obtain the accuracy of the Precise Positioning Service with due regard to the legitimate security concerns of the U.S. military. The answer to this seeming dilemma lies in the concept of **Differential GPS (DGPS)**.

Differential GPS is a system in which a receiver at an accurately surveyed position utilizes GPS signals to calculate timing errors and then broadcasts a correction signal to account for these errors. This is an extremely powerful concept. The errors which contribute to GPS accuracy degradation, ionospheric time delay and selective availability, are experienced simultaneously by both the DGPS receiver and a relatively close user's receiver. The extremely high altitude of the GPS satellites means that, as long as the DGPS receiver is within 100-200 km of the user's receiver, the user's receiver is close enough to take advantage of any DGPS correction signal.

The theory behind a DGPS system is straightforward. Located on an accurately surveyed site, the DGPS receiver

already knows its location. It receives data which tell it where the satellite is. Knowing the two locations, it then calculates the time it should take for a satellite's signal to reach it. It compares the time that it *actually* takes for the signal to arrive. This difference in time between the theoretical and the actual is the basis for the DGPS receiver's computation of a timing error signal; this difference in time is caused by all the errors to which the GPS signal is subjected; errors, except for receiver error and multipath error, to which both the DGPS and the user's receivers are simultaneously subject. The DGPS system then broadcasts a timing correction signal, the effect of which is to correct for selective availability, ionospheric delay, and all the other error sources the two receivers share in common.

For suitably equipped users, DGPS results in positions as accurate as if not more accurate than those obtainable by the Precise Positioning Service. For the mariner approaching a harbor or piloting in restricted waters near a site with a DGPS transmitter, the accuracy required for ship's safety is now available from a system other than plotting visual bearings. This capability is not limited to simply displaying the correct position for the navigator to plot. The DGPS position can be used as the prime input to an electronic chart system, providing an electronic readout of position accurate enough to pilot safely in the most restricted channel. The U.S. Coast Guard presently plans to install DGPS systems to provide 100% coverage along the eastern seaboard, the Gulf Coast, and the Pacific coast. Alaska and Hawaii will also be covered with a DGPS network. The DGPS signal will be broadcast using existing radiobeacons.

DGPS accuracy will revolutionize marine navigation. It is important to note, however, that, even with the development of the electronic chart and the proliferation of accurate, real-time electronic navigation systems, the mariner should not let his skills in the more traditional areas of navigation, such as celestial navigation and piloting, wane. They will become important secondary methods; any mariner who has put his faith in electronic navigation only to see the system suffer an electronic failure at sea can attest to the importance of maintaining proficiency in the more traditional methods of navigation. However, there is no doubt that the ease, convenience, and accuracy of DGPS will revolutionize the practice of marine navigation.