Chapter 19
Applications: Time-Based Navigation

19.1 Introduction

Among the areas of application made practicable by the advent of atomic clocks we list the following:

Space Science: Long-distance tracking and data acquisition from "deep" space probes such as Voyager.

Radio Astronomy: Very long baseline interferometry (VLBI), made possible through a common phase reference.

Planetary Motion: The dynamics of the Earth as a planet and the variability of the length of the day.

Radio Navigation: Perhaps the most useful application. Land-based networks Loran-C and Omega, and the satellite-based systems the TRANSIT system, culminating in the NAVSTAR Global Positioning System (GPS) and GLONASS.

We will limit our discussion to the fundamental way in which precision timing is critical to the success of these applications—so fundamental, in fact, that they are unthinkable without atomic standards. We will devote only a brief discussion to the first three and reserve our attention to a more detailed description of GPS.

19.2 "Deep" Space Probes

In the tracking of interplanetary probes such as Mariner and Voyager, the great distances that must be bridged, reaching out over hundreds of millions of miles into space, and the limited electrical power available aboard such "deep" space probes clearly put a great burden on the system used to track and communicate with them. Overcoming the weakness of the return signals received from the probe at the ground stations is further compounded by the varying Doppler shift in frequency as the probe pursues its trajectory. Large, high-gain (directional) antennas must be used with tens of kilowatts of power transmitted in the uplink and every

effort made to enhance the signal-to-noise ratio of the return signal. This has led to the development of correlation techniques for the handling of digital codes imposed as phase modulation on a stable carrier wave and the use of phase-locked receivers. In such a receiver a servo loop forces the phase of a locally generated stable signal to track the phase of the incoming weak signal. The communication link with the spacecraft is made by way of a *transponder* on board the spacecraft; that is, after the microwave signal "beamed up" from the ground tracking station is received by the spacecraft and demodulated to retrieve the commands and data impressed on it, it is converted (coherently) to a *different* frequency, modulated with the desired data, and beamed back to the ground station. Since the propagation time for the microwave (S-band, $\nu \approx 3$ GHz) signal to complete the round-trip journey may be tens of minutes, it means that noise and phase instability of the ground station oscillator over this relatively long time will limit the ability to communicate with the spacecraft. Phase noise on the order of less than tenths of a radian are required.

Of the atomic standards, the incomparable phase stability of the hydrogen maser in the time intervals involved in this application was early recognized; in fact, the earliest application of the maser was as a reference in the Deep Space Network operated by the Jet Propulsion Laboratory for NASA.

19.3 Very Long Baseline Interferometry

The second natural application is in radio astronomy, namely what is called very long baseline interferometry (VLBI). As already mentioned in connection with the Ramsey separated fields in magnetic resonance detection, the use of two separated antennas effectively increases the aperture of a radio telescope (along one dimension) to equal the baseline they define. We recall that the ability to resolve the detailed features of distant radio-emitting objects, that is, the resolving power, is determined by λ/D, where D is the diameter of the effective antenna and λ the wavelength. Since ground-based radio astronomy deals with radiation having wavelengths in the range from about 2 cm to 30 m, even for the short wavelength limit a resolving power comparable to the human eye for visible radiation implies an antenna diameter in excess of the world's largest steerable dish, the 100 m diameter one of the Effelsberg radio telescope in Germany. The largest fixed antenna, with a diameter of 305 m, is the one near Arecibo in Puerto Rico; it is a huge bowl of wire netting set in a natural valley with the focal point receiving antenna mounted on a girder 130 m above the dish. Although capable of detecting the faintest radio sources, its resolving power is not better than the human eye at visible wavelengths.

To understand how to get around this practical limit on antenna size, we must look at the way a radio "image" is constructed in a radio telescope—or an optical telescope, for that matter. In an ideal model, the radiation from each point in the distant object arrives at the antenna as a plane wave traveling in a slightly different direction and reaches different points on the antenna at different times and therefore

with a different phase. From each of these points, then, the wave is reflected coherently to converge toward the focus, where it combines with the wave from other points on the antenna to produce a resultant with a certain amplitude and phase.

In an actual radio telescope using one dish antenna, only the intensity at the focus is detected, and a plot of the intensity distribution of the distant radio source is made by scanning the direction of the antenna. If instead of allowing the radio wave to converge to the focus, we imagine the amplitude and phase at the surface of the antenna to be measured at different points, then the resultant that *would* be obtained at the focus can be calculated theoretically. That is, an "image" can be theoretically synthesized using amplitude and phase information obtained over a finite area; it happens that it is not even necessary to have a very high density of points to gain in resolution. That is, a finite array of widely spaced phase-tracking receivers with smaller antennas can be used. This is not feasible at optical frequencies, but phase-lock techniques at radio and microwave frequencies allow phase information even for the weakest signals to be recoverable. Thanks to the existence of atomic standards to supply a constant, common phase reference for distant receivers, it is possible to have an effective antenna "the size of the Earth." Given such a phase reference, the relative phase of the radio waves reaching antennas that may be thousands of kilometers apart can be detected. Depending on the frequency range of the radio waves under observation, this implies synchronization of reference oscillators at the different antennas to within the order of nanoseconds. With present-day atomic frequency standards, it is possible to meet this requirement for antennas literally continents apart, with a proportionate increase in resolving power. The transfer of phase information, which is equivalent to time transfer, between such distant locations has been an important challenge for a long time because it is so critical to the establishment of a radio navigational system, about which more will be said later in this chapter.

19.4 The Motion of the Earth

The next application we shall briefly mention is in the detailed study of the Earth's motion. As we noted in an earlier chapter on time scales based on astronomical observations, the detailed motion of the Earth is complex. Superposed on the basic motions of spin about its axis and revolution around the sun we have the precession of the spin axis relative to the figure axis (movement of the poles), the precession of the axis of spin in space (precession of the equinoxes), and the slowing down of the spin rate, among other things. The precession of the spin axis relative to the axis of symmetry of the Earth is possible because of the slight oblateness of the Earth's shape, and it leads to a very small circular movement of the poles a few meters in radius with a period of 430 days (the *Chandler period*). The precession of the spin axis in space, we recall, is due to external torques exerted on the nonspherical Earth by the sun and moon, and it leads to the much slower precession of the equinoxes, requiring 26,000 years to complete one cycle.

This manifests itself as the slipping of the seasons with respect to the months of the year. Since aside from the Chandler period, these phenomena occur on a relatively long time scale, their precise measurement requires clocks that establish a precisely *uniform* time scale extending over these long intervals. By uniform we mean a scale in which unit intervals are identical no matter at what point along the scale they happen to be. Faith in the atomic time scale being more uniform than the astronomical one stems from the belief that to the present accuracy there are no subtle long-term systematic effects on a quantum system to cause a departure from uniformity. On the other hand, the dynamical behavior of the Earth is expected to show those kinds of complex behavior on the basis of well-established theory. The keeping of precise atomic time over long periods enables data to be obtained that are useful in checking computational models based on that theory.

19.5 Radio Navigation

We take up now the main subject of this chapter: radio navigation and the Global Positioning System. Almost from the beginning of radio it was recognized that communication was not its only application. First came radio direction finding and radio beacons, then came the rapid development of radar during the second world war, followed by Loran (from the first letters of *long range navigation*) and Omega, which are radio-navigational networks of fixed radio stations of known location. Finally, after Sputnik came space-borne stations using satellites, which provide global coverage: the U.S.A TRANSIT and NAVSTAR/GPS, the Russian GLONASS, and most recently the European GALILEO component of a global navigation satellite system (GNSS).

19.5.1 Radar

The detection of radio echoes from distant targets, which is the basis of what came to be called radar (the first letters of *radio detection and ranging*), was first achieved around 1937, two years before the beginning of the second world war. However, its development, a tightly held secret at the time, saw an enormous impetus during the war, since it proved to be of critical importance in providing early warning of the approach of enemy aircraft. Since then, of course, there have been developed many types of radars with varied characteristics to fit the many military and civilian applications. Of particular note is the Doppler radar, in which not only is the range determined, but also the relative velocity of the target.

In a typical radar system, a short burst of electromagnetic radiation is transmitted using a dish antenna, and the radiation back-scattered by illuminated objects is focused by the same antenna, detected, and suitably displayed. The range is determined from the time *delay* between the instants of transmission of the pulse and reception of its echo. Since the velocity of the radiation in free space is 3×10^8 m/s,

to each microsecond increment in the delay corresponds an increment of 150 meters in the range of the target. The accuracy of the range determination then clearly depends on the accuracy with which the relative time/phase delay between the transmitted and received signals can be measured. To focus the return wave and thereby achieve adequate angular resolution in locating the target requires that the diffraction of the reflected wave at the antenna be kept small. This dictates the use of relatively short wavelength radiation in order that a dish antenna of reasonable size can be used. For that reason microwave radiation is generally used: an early common choice was 3 cm microwaves at a frequency around 10 GHz (X-band). Since the same antenna is commonly used for both transmission and reception, a fast-acting electronic switch must protect the receiver during transmission. As with radio receivers, the radar receiver has a local oscillator to heterodyne with the incoming signal to produce an intermediate frequency signal, which is then amplified and possibly taken through further lower-frequency amplifying stages before the signal is finally displayed on the screen of an oscilloscope. In the PPI (*p*lan *p*osition *i*ndicator) display, the dot on the oscilloscope screen moves out from the center on a radial line, and the returning echo causes the intensity of the dot to brighten. The radial line rotates in synchronism with the antenna and therefore gives directly the relative bearing of the target.

Radars in which the phase of the transmitted wave is available to be used as reference, that is, coherent radars, are widely used to enable Doppler information to be extracted. An important application of Doppler radar is as a *moving target indicator* (MTI), which discriminates between a moving target that reflects a Doppler-shifted frequency, and the ground, sea, or cloud clutter that appear at the same range. The phase stability of the oscillator in such radars will clearly determine the velocity resolution; however, it is only very short term stability that counts, since even for a range of 1000 km the propagation delay is less than 10 milliseconds. This means that without a high-power maser or some exotic superconducting cavity oscillator with extraordinarily low phase noise, a high-quality quartz oscillator is an adequate choice for this application.

19.5.2 Loran-C

Unlike the operating principle of radar, in which the user must actively transmit radiation in order to receive echoes, Loran involves only the reception of coded time signals broadcast from a network of fixed stations of known location. A radio-frequency carrier in the 1,750 kHz–1,950 kHz range was used in the original Loran-A chain, now superseded by the more accurate multichain Loran-C system operating at the even lower frequency of 100 kHz. This lower frequency gives the Loran-C system a much greater useful range, since the lower the frequency, the smaller the attenuation rate of the propagation mode in which the wave travels along the surface of the Earth, that is, the *ground wave*. Other modes of propagation involve waves reflected from electrically conducting layers of the

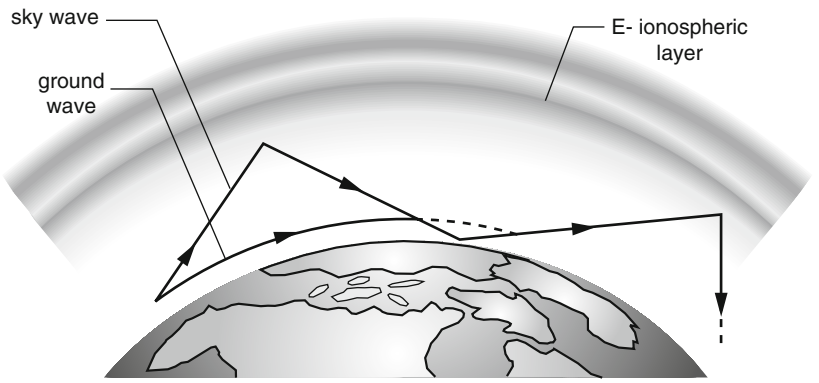

Figure 19.1 The ground wave and sky wave modes of radio propagation around Earth

atmosphere called the ionosphere; these are the *sky wave* responsible for global short-wave communication, shown in Figure 19.1. The system depends on the radio propagation time being an accurate measure of the distance between transmitter and receiver. Although the sky wave range is much greater than the ground wave, there is much greater uncertainty in the propagation time, since the altitude of the ionospheric layers is subject to change between daylight and nighttime hours, and depends on sunspot activity, etc. They arrive at the receiver later than the desired ground wave and must be carefully discriminated against in the timing signal circuitry. The propagation time of ground waves is to a first approximation proportional to distance; however, to achieve accuracies in the microsecond range requires making secondary corrections, depending mostly on the electrical conductivity of the surface and to a lesser extent the propagation properties of the atmosphere. These corrections can be computed accurately for propagation over seawater; however, transmission times over paths involving land with different types of terrain are much less predictable.

If a user ship or aircraft maintained precise time in synchronism with a transmitting station, then the one-way signal propagation time, assuming a known radio wave velocity, could be used to determine the range to that station. This, however, would require the user to carry a good atomic clock to maintain synchronism with sufficient accuracy to be useful, severely limiting the number of users who could avail themselves of the system. To overcome this limitation, the Loran system operates on the principle of the user determining the *differences* in the propagation delays of three or more precisely synchronized transmissions from widely separated stations, forming a network covering an extended geographical area. If we neglect at first the curvature of the Earth's surface, then surface navigation (no altitude information) requires a minimum of three stations. This can be seen graphically by plotting all the points that have a given constant difference of delay, and hence difference of range from two stations of known location. Figure 19.2 shows

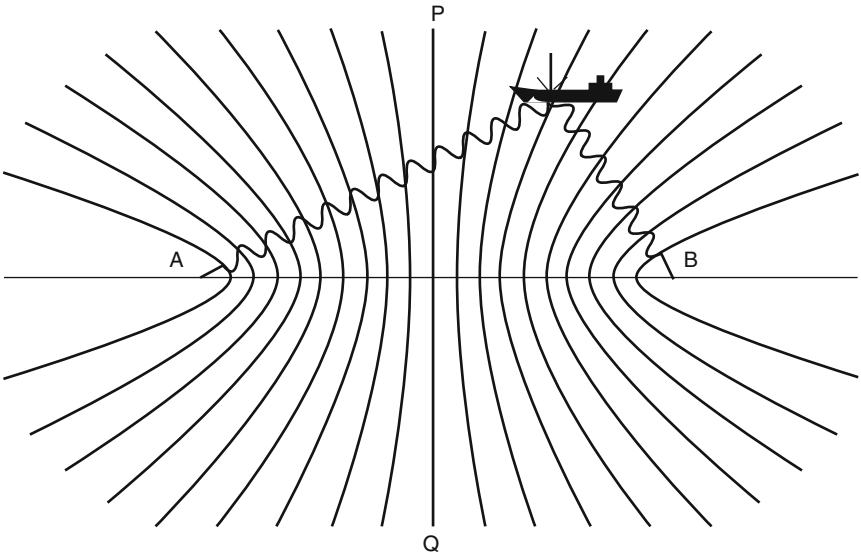

Figure 19.2 The hyperbolic lines of constant propagation delay between Loran stations at A and B

such a plot for two stations A, B; the locus of points having a constant *difference* in their distance from two fixed points is, in fact, a *hyperbola* with A and B as foci. For this reason this radio navigational system is sometimes called a hyperbolic system. The line AB is the *baseline*, and its perpendicular bisector, the *centerline* PQ, is the locus of points equidistant from A and B and therefore corresponds to zero difference. The *extensions* of the baseline to infinity away from the points A and B also correspond to a constant difference, namely the propagation time directly from A to B. To fix the position of the receiver, another set of hyperbolas is necessary, giving the lines of constant difference in time delay with respect to another pair of stations. From two observed delays between different pairs of stations, two hyperbolas are selected, one from each set associated with a pair of stations. These may theoretically intersect at two points; however, in that event independent information or the delay difference from another pair may be required to resolve the ambiguity. Since the hyperbolas in the vicinity of the baseline extensions have arms that tend to close up into a hairpin shape, the likelihood of ambiguity is greatest there, and navigators avoid using station pairs whose baseline extensions are near enough that the chance of ambiguity is a source of concern. In fact, there is another important reason to avoid the baseline extension region: It has to do with the size of the error in range incurred by a given error in time delay. From Figure 19.2, where lines of constant time difference are plotted for *equal* increments in that time difference, we see that in the neighborhood of the baseline extensions the lines are spread out much more than, for example, along the baseline itself. This means that for a given

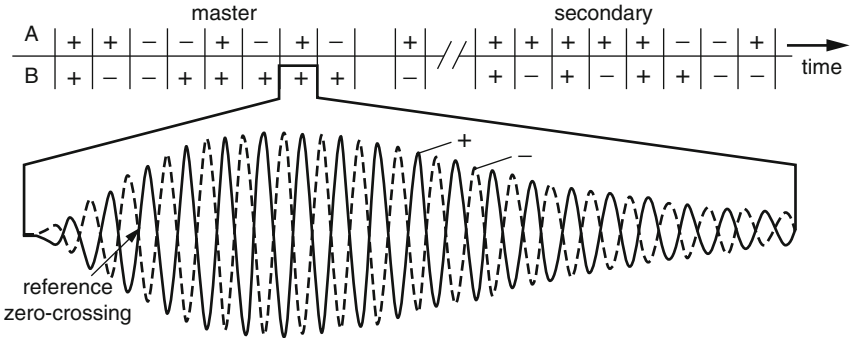

Figure 19.3 Waveform of a pulse in the groups broadcast by Loran-C stations (Maxim, 1992)

error in the time delay measurement, the error in position is larger in this area than elsewhere. While on the subject of errors, we note that an important factor is the angle between the two hyperbolas at the point of intersection. If we admit to a certain error in the timing from each transmitter, then the hyperbolic lines should be replaced by hyperbolic bands, whose widths reflect the possible statistical spread in the time delay. The area of overlap of the two bands now defines the error in the position, an area that is clearly bigger if the bands are nearly parallel than if they are nearly perpendicular.

The Loran-C stations are grouped in *chains*, each covering a certain geographical region, with each chain having one station designated as *master* (M) and two or more other *secondary* stations (W, X, Y, ...), any one of which may be paired with the master to form a master–slave pair, or *rate*.

Each station broadcasts the same 100 kHz carrier frequency, pulse-modulated in groups of pulses that are repeated at a rate unique to the chain, which helps to identify it. Pulse modulation is used with precisely defined pulse envelopes because of its advantages from the point of view of power utilization in communicating time with the best signal-to-noise ratio. The phase of the carrier is switched between successive pulses between 0° and 180° according to a binary code, the two phases corresponding to 1 and 0 in the binary system. These phase codes are chosen to distinguish between the master stations and the secondaries, and they alternate in time between codes designated as A and B, as shown in Figure 19.3. The phase modulation has the further benefit of helping to discriminate against the undesirable sky wave signal. The signal format for the secondary stations consists of groups of eight equally spaced phase-coded pulses 1 millisecond apart, while the master broadcasts groups contain an additional pulse 2 milliseconds after the first eight. The time interval between successive transmissions of the master's pulse groups is called the *group repetition interval* (GRI). Each chain is identified by a GRI *designator*, which is the group repetition intervals in units of 10 microseconds; for example, the northeast U.S. chain has the GRI designator 9960, and therefore the interval between groups of pulses is 99.6 milliseconds. This interval is chosen

so that the signals have time to propagate throughout the chain well before the beginning of the next group of pulses. Possible confusion is avoided by having a well-ordered, precisely timed sequence of transmissions from members of the chain. The transmissions occur in a sequence started by the master followed usually in the order of the letter designations of the secondary stations: W followed by X, then Y, then Z, etc. First the master begins the cycle; the first secondary does not transmit until it has received the master's signal and added a further delay called the *coding delay*; then the next secondary transmits after similar delays; and so on. Since the period between pulse groups is nearly 0.1 second, and the propagation times are typically less than 10 milliseconds, there is no ambiguity as to the identity of pulses from master and slave—the slave signal is known to arrive later, except along the extension of the base-line away from the slave station, where the signals would coincide were it not for the coding delay. This delay not only facilitates reading the time difference in this case, but it also may be changed at will, and it can be used for security in times of war. The timing information is recovered from the phase-modulated pulse signals using phase-sensitive detection and fixing a fiduciary point in a pulse to mark the time, independently of the pulse amplitude. The waveform of each pulse in a group is shown in Figure 19.4. The timing of the pulse is defined by the third zero crossing of the signal, as indicated in the figure, to take advantage of the sharp rise in amplitude early in the pulse, where little sky wave "contamination" is expected. The instrumentation used in the system has a resolution of a few hundredths of 1 microsecond. By radio communication, the oscillator at the master station provides corrections to the other station oscillators in the network as slaves, keeping them within microsecond synchronism. To maintain this degree of synchronism independently, without corrections over a period (say) of one week, implies a long-term stability on the order of 1 part in 10^{12}, which is assured by using a cesium standard. In order to provide a precise time distribution service in addition to navigation, and to tie networks covering diverse regions of the globe, the master oscillator is tracked relative to the time standards at the U.S. Naval Observatory, and adjustments are made to maintain accuracy within tolerance.

19.5.3 The Omega Network

The main drawback to Loran is the limited range of the ground wave, on the order of 1,500 km, necessitating a proliferation of chains and complex receivers to reach global coverage. An alternative radio navigational network called Omega achieves long-range coverage with a network of only eight stations by transmitting at much lower frequencies, in the *very low frequency* (VLF) band allocated for navigation, between 10 kHz and 14 kHz. The stations are scattered throughout the world, from Norway to Hawaii, and they transmit according to a precisely timed schedule; for example, station A (Norway) begins its transmission format on 10.2 kHz for 0.9 s, then is silent for 0.2 s, comes back on 13.6 kHz for 1 s, is silent again for 0.2 s,

and finally transmits on 11.33 kHz for 1.1 s and falls silent for the balance of 10 s before repeating the sequence. Like Loran, Omega is a hyperbolic system; however, unlike Loran, where the leading edges of pulses received from pairs of stations are matched to obtain the *time* difference, in the Omega system it is the *phase* difference that is observed. Since the phase repeats periodically along the wave, there will be an ambiguity as to the whole number of cycles difference that may be present. To resolve this ambiguity, the Omega receiver must be "initialized" by setting the initial coordinates, after which the receiver will automatically track the phase relationship between the received signals.

As already pointed out, by choosing very low frequency transmissions, long-range propagation is assured—even some penetration into seawater is possible. However, the waveguide-like modes of propagation between the Earth's surface and the lower D-region of the ionosphere have a phase velocity that is sensitive to the behavior of the ionosphere, giving rise to a number of phenomena that must be taken into account when assessing the reliability of a fix. Thus there is a change in the effective height of the ionosphere from around 70 km during daylight hours and 90 km during the night. Furthermore, disturbances occur in the ionosphere during and after "sunspot activity," which has been observed to occur with some regularity, repeating on the average every 11.4 years. Other potential sources of error include interference between different waveguide modes of propagation, and *wrong way* propagation, where the actual signal has traveled the long way around the Earth, rather than directly from the transmitter.

19.6 Navigation by Satellite

The use of satellites for navigation, surveying, and time dissemination has the major advantage of line-of-sight radio communication to cover very large geographical regions. It avoids the problem in surface communication from ground stations of uncertainties in the propagation velocity of radio waves over varying surface conditions around the Earth's curvature, as well as sky wave contamination. It is interesting that it was not many years after the launching of the first satellite, Sputnik I, in 1957, that I. Smith filed a patent describing a satellite system from which time codes could be emitted that would be received on Earth delayed by the propagation times, setting up hyperbolic *lines of position*, a straightforward extrapolation from the Loran concept.

In the U.S., support for a space-based navigational system came through military funding first by the Navy, at the Johns Hopkins Applied Physics Laboratory, and then by the Air Force, at the Aerospace Corporation. This typifies the channeling of research funds in the U.S. through the military services. There are numerous examples of technological advances made possible by government funding, usually through the military, which otherwise would not have been realized. A prime example is the H-maser. Private industry would not have developed it because the market for it was too small to defray developmental costs and turn a

profit in a reasonable time frame. As it was, the initial development was undertaken at NASA and under NASA contracts, first to Varian Associates, and then to Hewlett Packard Co. Fortunately, in instances such as atomic time standards and satellite technology the interests of the military happen also to accrue benefits to the public at large. This is proving to be particularly true of the Global Positioning System.

The Navy-sponsored effort led to the satellite navigational system TRANSIT, designed as an all-weather *surface* navigational system for Navy vessels, including strategic submarines. Its limited application to surface navigation and intermittent coverage made it unsuitable for high-speed aircraft and missiles, where continuous three-dimensional navigation is necessary. For this reason the Air Force opted in 1963 for a global navigational system first studied at the Aerospace Corp., a system that evolved into what came to be called the Global Positioning System (GPS). By 1965 the Air Force had decided to let out contracts for the development of user receivers. Other systems, such as the Army SECOR, were also being evaluated at that time, but by 1974 it was determined that a joint military project would be undertaken based on the Air Force GPS concept.

Such a system would have the advantage of *continuous* and total geographical coverage with line-of-sight communication. Its realization, however, clearly depends on the ability to sustain satisfactory operation of an accurate time standard on board the spacecraft, and to broadcast time signals and orbital data with sufficient power and accuracy from an orbiting spacecraft. The use of satellites also relies on the ability to compute accurate *ephemerides* (plural of ephemeris: orbital position as a function of time); this is essentially different from fixed land-based stations—a satellite occupies many different known positions (albeit at different times), almost like having several transmitters in a network stretched out along the orbit. However, to measure propagation delays from the same satellite at widely separated points along its orbit presumes that the receiver is able to maintain precise time over the relevant part of the orbit. In the absence of that, to obtain a fix in three dimensions, that is altitude in addition to latitude and longitude, requires signals to be received simultaneously from four satellites with synchronized clocks.

In the TRANSIT navigational satellite system developed by the U.S. Navy, first declared operational in 1964, there were four, and later six, satellites orbiting at an altitude of around 1100 km in nearly circular *polar* orbits. They completed their orbits in a little over 100 minutes, and as the Earth rotated under them, they provided good global coverage. The system was originally developed to determine coordinates of Navy vessels and aircraft, but eventually civilian use was authorized, and the system was used for surveying as well as navigation. Early experiments by the U.S. Defense Mapping Agency and the U.S. Coast and Geodetic Survey showed accuracies on the order of 1 m at a fixed point after several day's observation, using postprocessed precise ephemerides.

The desire to have the system accessible to users equipped only with a quartz oscillator with good short-term stability, rather than an atomic standard, is met, in effect, by measuring a *difference* in signal delay, but this time not necessarily between different transmitters as in Loran, but between successive incremental

positions of the *same* satellite. But the continuously changing signal delay due to the motion of the source is nothing more than the Doppler effect. One of the important observables is the point at which the Doppler shift in the signal passes through zero and reverses sign: this occurs when a satellite passes through the point closest to the receiver. Noting the precise times when this occurs and having an accurately computed ephemeris, so that the positions of a satellite are accurately known for those times, ultimately fixes the position of the receiver.

Needless to say, the success of such a system depends on the stability of the clocks on board the spacecraft and the ability to communicate accurate orbital information. The system's reliance on Doppler suggests that it is the short-term phase/frequency noise that will limit precision, while long-term drifts should be small to avoid the need for frequent corrections to stay within the tolerance limits. The Transit satellites were equipped with ultrastable quartz oscillators with a drift rate of a few parts in 10^{11} per day—atomic standards for on-board spacecraft applications were still under development. The long-term drift of these quartz oscillators could be modeled mathematically, allowing time corrections to be extrapolated. They controlled the frequency of transmission at 150 MHz and 400 MHz for Doppler tracking and navigation. The satellites were tracked by widely separated fixed ground stations of known location (TRANET) using the same basic Doppler technique used in tracking Sputnik. While the Doppler frequency shift itself gives information on the relative velocity (range rate), the accumulated *phase* shift it causes (computed mathematically by integrating the Doppler frequency with respect to time) gives the *change* in satellite–receiver distance. To derive the actual distances as the satellite continues in its orbit requires independent knowledge of its distance at least at one point (mathematically, to determine the integration constant).

The TRANSIT system had two important shortcomings: First, although the six orbiting satellites were able to provide global coverage, it was not continuous. A satellite passed overhead (at the equator) every 100 minutes or so, and users had to interpolate their position using dead reckoning between passes; under worst-case conditions a user might require several hours between fixes. Second, the navigational accuracy was only slightly better than Loran-C, relying as it did on on-board quartz clocks rather than atomic clocks.

19.7 The Global Positioning System (GPS)

In view of these deficiencies, and with the intervening developments in portable atomic clock and satellite technology, the U.S. Department of Defense in 1973 directed its Joint Program Office to oversee the development, evaluation, and deployment of an accurate space-based *global positioning system* (GPS). The product of that effort is the present *Nav*igation System with Timing and Ranging (NAVSTAR). Specifically, the functions of the system are to enable military, and now civilian, users to determine accurately, under all-weather conditions, their

Figure 19.4 A typical GPS satellite

position and velocity, and to transfer precise time on a continuous basis anywhere on or near Earth. This was to be achieved by having coded time and ephemeris signals broadcast from a number of satellites, each carrying an atomic clock, in such orbits as to ensure that a sufficient number of them are in view at all times, anywhere on Earth. The ranging method is again based on the propagation time of radio waves from the satellites to the user: either using the propagation delay in receiving the coded time signal, or the accumulated phase difference between the broadcast carrier wave and the user's reference oscillator/clock. As with the other time-based navigation systems, this system is designed to require receivers equipped only with a relatively inexpensive quartz clock. Since in general the receiver clock will not remain in exact synchronism with the satellite clocks, the range computed using the uncorrected propagation time observed is called the *pseudo-range*; the true range is obtained by correcting for the clock error. In order to fix the position of the receiver in *three* dimensions—longitude, latitude, and altitude—three *true* ranges are necessary. This can readily be seen if we imagine spherical surfaces drawn around the satellites as centers with radii equal to the true ranges to the receiver. If only two ranges are known, then the receiver may be at any point on the circle of intersection of the two corresponding spherical shells, whereas if the range to a third satellite is known, the position of the receiver is uniquely determined as being at the intersection of the third spherical shell with that circle. If the ranges to the satellites are only pseudo-ranges because of the clock error, then of course, the position so determined will be in error, and the spherical shell drawn around a *fourth* satellite with the pseudo-range as radius will

not pass through the same position as the other three. However, if we assume that all the satellite clocks are in perfect synchronism, so that a single clock error exists because of a drift in the receiver clock, then it will be possible to compute the correction to the receiver clock which will convert the pseudo-ranges to true ones, and make the *four* spherical surfaces with corrected radii pass through a unique point, the true position of the receiver. This system then requires that the signals from at least *four* satellites be in full view globally at all times.

A description of the system separates naturally into three *segments*: the satellite system, the ground monitor and control stations, and the users, including the many types of receivers. We will give a brief description of these in that order.

19.7.1 The Satellite Constellation

From what has been said, the number of satellites and their orbits must be chosen so that signals from at least four of them can be received simultaneously anywhere on Earth, 24 hours a day. This assumes that because of the unpredictable motion of the receiver, signals must be received from four different satellites at the same *epoch*; however, in the event that the receiver is stationary or moving slowly, then signals received sequentially from satellites at several different epochs would suffice to fix the position of the receiver, since the satellites will occupy different known positions at these epochs. Nevertheless, even in this case, to be assured of an immediate and accurate fix, four satellites must simultaneously be in full view at all times.

The ultimate choice of the number of satellites in the *constellation* and their orbits evolved from a number of proposals early in the 1980s, ranging from a 24-satellite constellation in 3 orbital planes inclined 63° to the equator to 18 active satellites with three in each of six orbital planes. The present policy calls for a constellation consisting of four active satellites orbiting in each of six planes with an inclination of 55°, making a total of 24 satellites, plus four more spare satellites to replace mal-functioning operating satellites.

There are five categories of GPS satellites, designated as Block I, II, IIA, IIR, and IIF satellites. The typical appearance of a GPS satellite is illustrated in Figure 19.4. The eleven satellites making up Block I were launched in the period between 1978 and 1985. With the exception of one booster failure on the seventh in the series, all the launches were successful. The design life of these satellites was only 4.5 years, yet two of them were still operating satisfactorily after twice that period. In common with all the subsequent satellites, these carried atomic clocks in addition to sophisticated radio communication equipment, as well as a propulsion system for orbital corrections. They were powered by two 7 sq. meter solar panels, weighed 845 kg, and were placed in orbits inclined at 63°. The Block I satellite signals were not secured to prevent general accessibility to civilian users; this is in contrast to the Block II satellites, some of which broadcast inaccessible coded signals. By 1994 the remaining Block I satellites, still functioning at reduced power, had been made redundant and have been boosted out of orbit and left for

scientific tests. The first of the Block II satellites, weighing over 1,500 kg, was launched in 1989 using a Delta II rocket; subsequently, over the period 1989–90 eight more satellites in this series were launched and placed in four different orbital planes inclined at 55° to the equator, with an altitude of about 20,000 km. From 1990 to 1994 fifteen more satellites were launched, which are classed as Block IIA, the "A" denoting advanced, which have the capability of communicating with each other, and some of which carry *optical* corner cube reflectors, which reverse the direction of a tracking laser beam from a ground station, independently of the orientation of the satellite. This facilitates the tracking of a satellite using laser ranging, the optical analogue to radar, with an accuracy approaching the centimeter range. This capability is matched by the accuracy of frequency and time made possible by *four* on-board atomic standards: two rubidium and two cesium, with long-term frequency stability of a few parts in 10^{13} and 10^{14} per day, respectively. This corresponds to an average drift of about 3 nanoseconds per day for the cesium standard, which for a radio wave traveling at 3×10^8 m/s implies an error of about 1 meter in ranging.

The Block IIR (the "R" for "replenishment") satellites were scheduled for launch using the Space Shuttle in 1996, however, due to a number of technical problems, including the production of the sophisticated on-board atomic clocks, the delivery and launch were delayed. The first Block IIR satellite was delivered to Cape Canaveral in September, 1996, and after extensive tests to verify a smooth integration with the existing GPS, was ready for launch in 1997. Unfortunately, that launch was unsuccessful due to failure in the Delta II launch vehicle. The launch was rescheduled for a later date. This new generation of satellites has a design operational lifetime of ten years, two and a half years longer than the Block II/IIA satellites. They are considerably heavier, at over 2,000 kg, and would be able to accommodate space-adapted hydrogen masers, as onboard frequency/time standards; however satellites in this series so far have carried three rubidium frequency standards. As a class, the H-masers have long demonstrated superior short-term stability, and the thought of flying them in satellites has long been cherished by some who have devoted their careers to that end. However, the massive ion vacuum pumps and large magnetic shields that characterize laboratory installations presented a formidable obstacle in meeting the size and weight constraints of a spacecraft. The frequency stability expected of the hydrogen space masers is better than one part in 10^{14}, a tenfold improvement over the Block IIA standards, with, it is hoped, a corresponding upgrade in system performance. Of course now, with laser-cooled Cs standards and optical ion standards, the choice of on board clocks is much more competitive.

A fourth generation improved Block IIF satellite series has been under development since 1996, first under a U.S. Air Force contract with Rockwell, and more recently under a U.S.$1.5 billion development contract with the Boeing company. These satellites are 50% heavier than the Block IIR satellites will accommodate more equipment and expanded missions with a design lifetime of 15 years. They will provide civilian users with significantly more accurate signals. Their launching

into orbit requires the use of larger launch vehicles, such as ones developed under the Evolved Expendable Launch Vehicle (EELV) program. This program involves two families of launch vehicles: Atlas V and Delta IV.

19.7.2 The Orbital Parameters

A major premise in the use of satellites as platforms for radio transmitters forming a navigational network is that their precise positions are predictable at all times, and that this information can be communicated to the user. That this is the case hinges on the fact that the satellites follow orbits that to a very good approximation are Keplerian ellipses with Earth's center at one of the foci. To completely specify the motion of a particle in 3-dimensional space, acted on by known forces, requires in general six numbers, which may, for example, be the three coordinates and the three components of velocity at some point in time (epoch). It follows that the most general elliptical orbit in space requires five parameters to specify it completely, and one parameter to specify the position of the particle in the orbit. The number of orbital parameters results from the two angles required for the orientation of the plane of the orbit, another angle to specify the orientation of the ellipse in that plane, and two more to specify the semi-major axis and ellipticity of the ellipse. These parameters are illustrated in Figure 19.5, where the position of the particle in the orbit, historically known as the *anomaly* (sic), is shown as the angle θ at the focus of the ellipse where the center of mass of the system is located. For GPS satellites the semi-major axis is nominally 26,560 km, and the orbital period is *half* a sidereal day, that is, half the time for a complete rotation of the Earth with respect

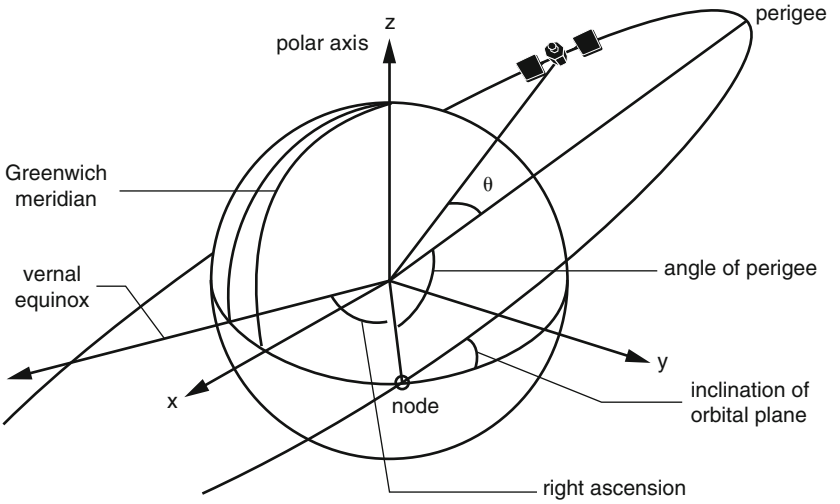

Figure 19.5 The definition of the satellite orbital parameters (Hofmann–Wellenhof, 1994)

to the stars (actually with respect to the vernal equinox, which is very nearly the same thing). Having the satellite complete its orbit a whole number of times each sidereal day ensures that its ground track repeats every sidereal day; that is, ideally it will pass overhead at a given point on Earth at the same time every sidereal day.

The simple reentrant (closed) elliptical Keplerian orbit is predicted theoretically according to Newton for a satellite attracted only to a rigid, homogeneous, spherical Earth, which acts on the satellite as would a point mass at the center of the Earth. An actual Earth satellite is subject to conditions that differ slightly from that, differences that are called *perturbations.* Fortunately, these are all small compared to the forces that give rise to the Keplerian orbits, a fact that is exploited theoretically in arriving at corrections by a process of successive approximation. The result is that the orbital parameters are subject to change in time and must therefore be corrected either by actually activating thrusters, or updating actual tracking data and computing the changes in the parameters they imply.

19.7.3 Perturbations Affecting the Orbit

There are two types of perturbations: those that have a gravitational origin, such as those arising from the presence of the moon and sun, and the oblateness of the Earth; and those that are nongravitational, such as solar radiation pressure, solar wind, and air drag. Since there is potentially a bewildering number of different factors that may perturb the motion of the satellite, we need to stipulate what would be a tolerable error in the satellite position. If we set that tolerance at 1 meter deviation over one orbital period, we would find that a *constant* perturbing force must not cause an acceleration greater than 10^{-9} ms^{-2}. To appreciate the relative size of such a force, we note that the primary gravitational pull on a GPS satellite due to the Earth, which keeps it in orbit, is GM_{E}/r^2, where G is Newton's gravitational constant, M_{E} is the mass of the Earth, and $r = 26{,}560$ km is the semi-major axis of the orbit. A quick calculation yields ≈ 0.57 ms^{-2}, almost a billion times greater than the tolerable perturbation!

The largest perturbation comes from the nonsphericity of the Earth. Because of the Earth's rotation about its axis, there is a centrifugal force that varies from a maximum at the equator, diminishing with latitude, becoming zero at the poles. In consequence of this, the net inward force at the surface, the observed weight, is greatest at the poles and diminishes toward the equator. The equilibrium figure of a plastic body is an oblate spheroid, a slightly flattened sphere with an elliptical cross section through its axis. The actual oblateness of the Earth is small—the diameter from pole to pole is only about 43 km shorter than through the equator, or about one part in 298. Of course, the detailed shape and structure of the Earth is far from a smooth homogeneous oblate spheroid; the degree of detail that is significant clearly depends on how far the satellite is from the surface. Indeed, if the object had been to study the topography of the Earth through satellites, the orbits would have been chosen to bring out the very perturbations that the GPS system must

avoid. At an altitude of about 20,000 km, the GPS satellites are sufficiently far from the Earth's surface that the oblate spheroid model, which introduces less than one part in 10^4 correction, is considered adequate, and higher-order approximations are expected to yield negligible improvement in orbital accuracy. An analysis of the effect of the Earth's oblateness on a satellite orbit shows that there is a slow precession (a rotation of the perigee) at a rate proportional to $(5\cos^2\theta - 1)$, where θ is the angle the orbital plane makes with the Earth's equatorial plane. This rate is nearly zero if $\theta \approx 63°$, which provides a rationale for choosing that angle for the early satellites. Moreover, the analysis shows that the mean time to complete an orbit is also a function of θ, and the variation in this case is proportional to $(3\cos^2\theta - 1)$, which is nearly zero for $\theta \approx 55°$, hence the choice of that angle for the later satellites.

The other gravitational type of perturbation comes from the presence of the sun and moon, and is referred to as *tidal effects*, since the attractions of the same two bodies account for the tidal action on the Earth. A rough estimate of the variation of the sun's gravitational pull over the orbit of the satellite using Newton's inverse square law of gravitation yields for the sun's perturbation about 2×10^{-6} ms^{-2}, while for the moon the figure is about 5×10^{-6} ms^{-2}. Related, indirect perturbations due to the tidal deformation of the "solid" Earth, as well as the oceanic tides, are very much smaller, in the range of 10^{-9} ms^{-2}.

Of the non-gravitational perturbations, the most important is the solar radiation pressure. We recall that radiation, whether a laser beam or radio wave, carries linear momentum, and therefore the absorption and scattering of sunlight by a satellite results in forces being exerted on it. Since the scattering in general is not the same in all directions, it follows that the force experienced by a satellite is not necessarily in the direction of the incoming rays of the sun—there will be a smaller transverse component. The actual perturbation produced obviously depends on the *solar constant*, a measure of the intensity of solar radiation falling on the satellite ($S = 1.4$ kW/m^2), the cross section presented by the satellite to the sun's rays, the reflectivity of the surfaces, etc. There is the further complication that the satellite may pass through the shadow of the Earth; that is, it may experience periods of solar eclipse. The computed size of the perturbation is on the order of SA/cM_s, where A/M_s is the ratio of cross section to mass of the satellite, which yields $\approx 10^{-7}$ ms^{-2}. This shows that radiation pressure produces a very significant perturbation, one that must be well modeled and taken into account if the desired accuracy is to be achieved.

Finally, in addition to solar radiation there is the *solar wind:* the sun continuously emits particles, mostly high-speed electrons and protons. Near the Earth's orbit the average speed of the protons is about 400 kms^{-1}, and its number density ranges from 2×10^6 to 10^7 particles m^3. Assuming that the particles are completely stopped on collision with the spacecraft, the resulting acceleration, for example on a spacecraft having $A/M_s = 0.03$, is less than 10^{-10} ms^{-2}, and therefore negligible.

19.7.4 The Control Segment of GPS

The crucial functions of monitoring the orbits of the satellites and the frequency and phase of their on-board atomic clocks, and updating orbital parameters for ephemeris prediction, are the responsibility of the control segment of GPS. This comprises ground-based stations of known geodetic position, including a master control station and three other control stations, as well as a worldwide network consisting of five monitoring stations. The master control station is at the Consolidated Space Operations Center in Colorado Springs, Colorado. It collects the satellite tracking data from the worldwide monitoring stations, from which it computes the updated orbital and atomic clock parameters. This information, along with other operational commands, is sent to the three ground-based control stations to *upload* to the satellites. The monitor/tracking network stations are located at Colorado Springs, Ascension Island (South Atlantic), Diego Garcia (Indian Ocean), Kwajalein (North Pacific), and Hawaii. These stations are equipped with precise cesium clocks and receivers that continuously track all satellites in view. The signal propagation times, which yield the *pseudo-ranges*, are obtained every 1.5 seconds, are "smoothed" to allow for ionospheric and meteorological variables, and transmitted as 15 minute interval data to the master control station. The three ground control stations are positioned also at the three monitor station sites on Ascension, Diego Garcia, and Kwajalein.

The implementation of a global time-based navigational system clearly requires the definition of a suitable frame of reference, including time. Position and time of any element of the system must be referred to a common, invariant set of coordinate axes and clocks. To specify the position of a point over a finite region of the Earth's surface, as with Loran-C (or ordinary surveying, for that matter), requires only certain fiduciary reference points to establish a baseline and the measurement of appropriate angles, which are then in a sense the coordinates of any point in that surface region. Obviously, this will not work for a global system, where positions in *space* surrounding the Earth is included: For that we need ideally an *inertial* frame of reference fixed in space and not tied to the Earth and partaking of its complicated gyroscopic motion, etc. One such system that neglects only the residual variation in the gravitational field over the Earth–satellite system, takes the direction at a specified epoch of the Earth's angular momentum axis (which is constant apart from the slow precession of the equinoxes) as the coordinate z-axis, and the direction of the vernal equinox, which is perpendicular to the Earth's axis and lies in the orbital plane (the *ecliptic*), as the origin of the longitude coordinate. The angle of latitude with respect to the z-axis and the radial distance from the Earth's center complete the system. It is in terms of the coordinates in this *quasi-inertial* (nonaccelerating) geocentric system that the computation of the satellite orbits is carried out. However, for the practical purposes of the navigator, what is required are his coordinates and altitude with respect to an Earth-fixed system using the prime meridian through Greenwich as the origin of longitude. Such a system is the (Conventional)

*T*errestrial *R*eference *F*rame (TRF), in which the z-axis is taken to be the *mean* position of the Earth's rotational axis during the arbitrary period from A.D. 1900 to 1905. For a true prediction of altitude it is not sufficient to assume a spherical Earth—the oblateness must be taken into account—and therefore, since 1987 GPS has used the *W*orld *G*eodetic *S*ystem WGS-84 (these things are regularly updated and therefore their designation includes the last two digits of the year), in which the Earth's figure is an ellipsoid with semi-major axis 6,378,137 meters, and the geometric flattening is 1 in 298.2572. This system is operationally established by a set of ground-based control stations serving as reference points, with particularly accurate position-fixing facilities including laser ranging and very long baseline interferometry (VLBI). Once the orbits of the satellites are computed using coordinates in the free quasi-inertial system, they must then be *transformed* into those in the practical Earth-fixed system, using the known motions of the one system relative to the other.

The operation of the GPS system presumes that all elements, including the navigator, maintain close time synchronism, and that when clocks drift apart, as they naturally will, it is possible to model their behavior mathematically to predict clock errors. The satellites and ground-based stations are all equipped with precise atomic clocks, which of course keep what is defined as atomic time. Navigators, on the other hand, are closely tied to what is called *U*niversal *T*ime (UT), which is defined in terms of the *mean* solar day, the basis of civil time. This is the average of the *apparent* solar (24 hour) day, which varies throughout the year, taken over that period. The time scale used by GPS is based on what is called *U*niversal *T*ime *C*oordinated (UTC); the unit in this system is the atomic second. However, because of the possibility of long-term drift of the universal time with respect to atomic time, they are kept in step to within one second by inserting a "leap" second as required. This results in a piecewise uniform time scale that tracks universal time; any fractional difference between the different time scales is monitored and published by national observatories charged with this service.

We should recall at this point the *Sagnac effect*, a relativistic effect on time measurement associated with the rotation of the Earth, which we introduced in Chapter 7. A coordinate system fixed in the Earth is noninertial, since its rotation with respect to "the fixed stars" constitutes an accelerated motion (not of speed, but changing direction). Consequently, as stated earlier, it is Einstein's theory of *general relativity* that is involved. According to the theory, if we imagine we have two identical, precise clocks at some point on the Earth's equator, and one remains *fixed* while the other is taken *slowly* (with respect to the Earth) along the equator all the way around until it reaches its starting point, then the time indicated on the two clocks will not agree. The difference $\Delta\tau$ was quoted as being given by the following:

$$\Delta\tau = \pm\frac{2\Omega}{c^2}S, \tag{19.1}$$

where $\boldsymbol{\Omega}$ is the angular velocity of the Earth (7.3×10^{-5} rad/sec) and S is the area ($\pi R_E^2 = 1.3 \times 10^{14}$ m^2) enclosed by the path of the moving clock. The formula yields a significant time difference of about $\pm 1/5$ microsecond. This is not insignificant in the present context and must be taken into account.

The determination of satellite orbital position as a function of time, from tracking data obtained by monitoring stations of known location, is the reciprocal problem to that of navigation using signals received from satellites having known orbital positions. We have seen that to completely predict a satellite position requires six numbers: These may be the three coordinates of position in 3-dimensional space and the three components of velocity at a given epoch, or the three coordinates at two distinct epochs. The first instance casts the problem as one of using the equations of motion to predict the motion subsequent to given *initial values*, and the second as fitting the general solution to given *boundary values.* The satellite orbital motion is solved using the quasi-inertial space coordinate system, whereas the positions of the tracking stations (as well as coordinates required by the users) are naturally with respect to the Earth-fixed conventional terrestrial system. The observational nexus between the two systems is through the precise tracking of GPS satellites afforded by laser ranging and VLBI from points coincident with some of the GPS monitoring stations. The need for mathematical transformations between the two systems makes the task of determining and updating the satellite ephemerides a complicated one, but nevertheless a manageable one thanks to integrated circuits and on-board computers.

19.7.5 Coding of GPS Satellite Signals

From the beginning, the idea of a precise global system of satellite navigation was inspired and brought into being by the U.S. Department of Defense, and its military importance makes the need for security pretty obvious. On the other hand, there was a desire to accommodate civilian users of the system, which has led to an elaborate coding scheme for controlling full accessibility to the system; unauthorized users were to have only a purposely degraded accuracy of positioning. Two carrier frequencies are actually broadcast in order to provide information on the dispersion of the radio waves as they traverse the ionosphere, that is, the frequency dependence of the propagation velocity of the radio waves through the ionized layers surrounding the Earth. This enables the ionospheric delay to be mathematically modeled in correcting the observed satellite pseudo-range. The two frequencies are the 154th and 120th harmonic of the fundamental frequency at 10.23 MHz based on the on-board atomic frequency standards, with stability on the order of parts in 10^{13}. The two frequencies, which fall in the so-called microwave L-band, are at $L_1 = 1.57542$ GHz and $L_2 = 1.22760$ GHz, with wavelengths about 19 cm and 24 cm respectively. In addition to the time-ranging codes already mentioned, these carrier frequencies are also modulated with data including the updated satellite orbital parameters, GPS time and satellite clock reading and drift, etc.

The type of code used is the so-called *pseudorandom number (PRN)* code, a particularly appropriate choice both for security and precision time comparison. It is a binary code generated by a shift register in such a way that within a certain group that repeats periodically, the binary bits are more or less randomly distributed. As an example, consider a 5-bit shift register in which at every clock pulse the bits advance to the right one place, the extreme right-hand bit becoming part of the output. The criterion for choosing the bit that replaces the one on the far left is the key to the code: For example, it may be chosen to be 0 if the bits in the third and fourth places are the same, otherwise a 1. A possible PRN code generated this way would be (011010111100010). A different choice of the criterion clearly would lead to a different output sequence. This binary code is impressed on the carrier wave as a biphase modulation; that is, the phase for a binary 1 is shifted by 180° with respect to a binary 0, as shown in Figure 19.6. The extent of "randomness" of the PRN code can now be seen by computing the *autocorrelation* function of the signal biphase modulated according to it.

To do this we simply form the product of the signal with its duplicate *shifted* a whole number of clock intervals, and then sum (integrate) over the whole period of the sequence. Since a phase shift of 180° is equivalent to a reversal of sign, we find that the autocorrelation has a maximum when there is *no* shift, since products of signals with the same sign are summed in the entire code. On the other hand, if we compute the correlation of the signal with its duplicate shifted even one space to the right or left, the result is zero, as can be verified for the PRN code we gave as an example. Moreover, the correlation will be small between the PRN coded signal and *any* other binary sequence that does not match it identically, since some negative products of signals of opposite signs are included in the sum. It follows that not only does the code provide a sharp time-matching function between a coded signal and its duplicate, but it also enables this match to be available only to those who can generate a duplicate code. It therefore fulfills the added function of controlling access to the information carried on the satellite signals. By assigning different

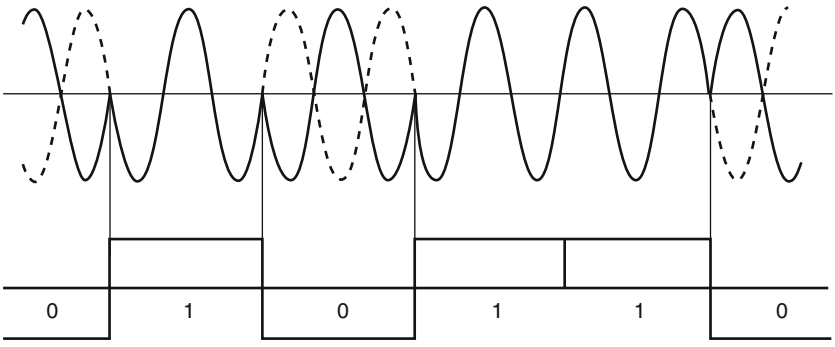

Figure 19.6 Binary phase modulation of satellite pseudorandom number signals

individual PRN codes to the many satellites that make up the GPS constellation, the need to be able to identify each satellite is fulfilled without any ambiguity.

The original concern of the U.S. Department of Defense that the precise positioning capability of GPS be secure against compromise or use by an enemy in times of war, while allowing a degraded capability to the general public, led to a rather complicated coding scheme. The dual precision capability is achieved principally by having two distinct PNR codes for the satellite clock readings impressed on the carrier waves broadcast from the satellites: the so-called C/A (*C*oarse/*A*cquisition) code, and the P (*P*recision) code. A further security provision is A-S (*A*nti-*S*poofing), by which is meant a countermeasure against anyone sending out false signals with the GPS signature, thereby confounding the system, or worse. The anti-spoofing feature consists in using a code (the W-code) to encrypt the P-code, yielding what is designated as the Y-code. The C/A-code is available to the general public, as is the P-code; but of course not the Y-code, although there is strong support to the idea of making the full capability of the system available to everyone.

The C/A-code is generated by combining the outputs of *two* 10-bit fed-back shift registers using binary addition modulo-2, that is, without "carrying." The clock rate for generating the C/A code is one-tenth the 10.23 MHz atomic-based reference, that is, 1.023 MHz, and it is repeated every millisecond. The interval between two bits in the code of just under 1 microsecond corresponds to the propagation delay for an increment in the range of 300 meters. It is impressed only on the L_1 carrier wave in phase quadrature (displaced 90°) to the Y-code, which is on both carrier waves L_1 and L_2.

To generate the P-code is somewhat more complicated: It is a certain combination of *two* PRN sequences, each generated by two registers, repeating about every 1.5 seconds, one containing over 15 million bits, while the other contains an *additional* 37 bits. Because of the difference in the number of bits in the two PRN sequences, the combined sequence will repeat only when a whole number of repetitions of the one sequence has an equal number of bits as another whole number of repetitions of the second sequence. Expressed symbolically, if n_1 and n_2 are the numbers of bits in the two PNR sequences, then the combined sequence will repeat after p repetitions of the first sequence, or q repetitions of the second, provided that p and q are the *smallest* whole numbers for which $pn_1 = qn_2$. For example, if $n_1 = 6$ and $n_2 = 4$, then $p = 2$ and $q = 3$, and the combined sequence repeats every $pn_1 = qn_2 = 12$ bits. Of course, in the language of elementary arithmetic, 12 is simply the least common multiple of 6 and 4. Getting back to the P-code, we find that the sequence that results from combining one having about 15 million bits with another of slightly greater number will not repeat until after about 200 trillion bits! At the clock rate of 10.23 MHz, the interval between successive bits is less than one-tenth of a microsecond (corresponding to a range interval of 30 meters), and the code repeats every 266.4 days. The total code length is divided into one-week segments, which are assigned to satellites defining their PRN identification

number. To show explicitly the broadcast signal as a function of time, let $S_{\mathrm{C/A}}(t)$, $S_{\mathrm{Y}}(t)$, $S_{\mathrm{D}}(t)$ be the sequence of $+1/-1$ constituting the binary codes C/A, Y, and the navigational and other data. Then we have

$$\begin{aligned} L_1 &= a_1 S_{\mathrm{Y}}(t) S_{\mathrm{D}}(t) \cos(2\pi\nu_1 t) + b_1 S_{\mathrm{C/A}}(t) S_{\mathrm{D}}(t) \sin(2\pi\nu_1 t), \\ L_2 &= a_2 S_{\mathrm{Y}}(t) S_{\mathrm{D}}(t) \cos(2\pi\nu_2 t), \end{aligned} \tag{19.2}$$

where ν_1 and ν_2 are the frequencies of the carrier waves L_1 and L_2. The total navigational data message consists of 1500 bits subdivided into 5 subframes, generated at a clock frequency of 50 Hz, so that it takes 30 seconds for the whole message. It contains the satellite orbital position (ephemerides) update, GPS and satellite clock time including various numbers to model the satellite clock correction, and other data of a "housekeeping" nature. The first subframe contains among other things the GPS week number, numerical coefficients to model the satellite clock correction, predictions of range accuracy, and age of data. The second and third subframes contain the satellite ephemerides. The contents of the fourth and fifth subframes change from one message to the next, repeating after 25 "pages"; the total information contained in all the different pages therefore takes 25×30 seconds, or 12.5 minutes, to broadcast. The pages of the fourth and fifth subframes are broadcast by all satellites; in addition to those reserved for military use, these pages contain data relating to the ionosphere, UTC, satellite health status, and low-precision orbital data on all GPS satellites.

19.7.6 Corrections to Signal Propagation Velocity

To reach ground-based or airborne receivers, the satellite transmission must penetrate the ionospheric layers of the atmosphere, as well as the troposphere. Although the velocity of propagation of the signal is little different from that in free space, nevertheless, the great distances involved lead to an accumulated effect on the arrival time of the signal, which because of *dispersion* differs for the two carrier frequencies L_1 and L_2. It can be shown on the basis of a simple model, in which the electric field component of the radio wave drives otherwise free electrons into oscillation, that the frequency dependence of the *phase* is approximately as follows:

$$V_{\mathrm{phase}} = \frac{c}{\sqrt{1 - \frac{80.6 N_{\mathrm{e}}}{\nu^2}}}. \tag{19.3}$$

The electron concentration N_e is stratified horizontally, increasing stepwise with altitude from the E-layer with around 10^{11} electrons/m^3 at 100 km to the F_2-layer above 300 km with around 10^{12} electrons/m^3. At the GPS signal carrier frequencies, the effect on the velocity is on the order of 3 parts in 10^5; this corresponds to a correction on the order of 10 m. We should note that for a dispersive medium, such as the ionosphere, it is necessary to specify exactly what the velocity refers to: Is it the crests of an infinite wave train, or the leading edge

of a pulse, or what? We recall that a complex waveform can be analyzed into its Fourier frequency components, and these will travel with different velocities, so that the waveform will in general change while it advances, leaving no feature whose rate of advance would be useful to define the velocity. If the waveform is such that its spectrum is contained in a narrow range centered on one frequency, such as we have in each of the GPS broadcasts, then the modulation pattern on the carrier is preserved and travels with the so-called *group velocity*. In the case of propagation through the ionosphere the group velocity is smaller than the velocity of light in free space, while the phase velocity is greater. Hence pseudo-ranges derived from time-code matching are based on a different velocity from those based on the relative phase between the carrier wave and local reference oscillator. It is interesting to note in passing that there are instances where the computed group velocity is actually *greater* than the velocity of light in free space: The dispersion in such media is described as *anomalous*. In the days before Einstein's theory of relativity this would not have been considered particularly unsettling; as it was, in the early days of that theory, there was general relief when Sommerfeld and Brillouin showed that in fact the *beginning* of a radio transmission always travels with just the velocity of light, and that *signal* velocity is not the same as the group velocity in cases where the dispersion is anomalous. The signal velocity is always less than or equal to the velocity of light.

As already pointed out, the concentration of electrons in the ionosphere N_e and its distribution with respect to altitude vary according to exposure to the sun and sunspot activity. Furthermore, the radio waves must pass through the layers of the ionosphere along a path that obviously varies with the position of the satellite in its orbit. The ionospheric correction to the signal delay must therefore be continuously monitored. This is the justification for using the two broadcast frequencies, L_1 and L_2; the signal delays provide two numbers to solve for the two unknown quantities: the pseudo-range and the effective electron concentration.

Unlike the ionosphere, the neutral troposphere is nondispersive; that is, the velocity of a wave does not depend on its frequency, and there is no difference between the velocity of propagation of the phase of the carrier and the modulation impressed on it. The refractive index is a function of the atmospheric temperature, pressure, and water vapor concentration. Several semiempirical models of the refractive index as a function of altitude have been developed; based on one of these the tropospheric delay is estimated along the slant path from the satellite to the receiver. The correction for this amounts to only a few meters in the pseudo-range.

19.7.7 The User Segment

Finally, we come to the user segment of the system. This serves in addition to the military services a large and expanding body of civilian users: navigators ranging

from those of high-speed aircraft, to pleasure boat operators, to hikers in the woods. With the ongoing drive to make the full capabilities of the system accessible to the general public it may not be long before GPS is incorporated into a multitude of technologies affecting the lifestyle of ordinary people.

The types of GPS receivers currently available are many, and they vary widely in sophistication and cost. They may have special features to enhance their performance in specific applications, such as high-speed navigation, or large-scale surveying, or precise synchronization of remote clocks. But basically, they are special radio receivers with precise phase/time tracking and navigational data processing capabilities.

As radio receivers, their first essential component is the antenna. For a GPS receiver, the design of the antenna and it physical environment are particularly important. Ideally, it should convert the oscillating electric (or magnetic) field component of an otherwise freely propagating wave into an oscillation of current in the receiver circuitry. The phase of that current oscillation must track exactly that of the wave, allowing at most a fixed phase offset irrespective of the orientation of the antenna. This is obviously crucial in applications where the receiver is subject to rapid movement. To receive signals from several different satellites, whether simultaneously or sequentially, requires an antenna whose response is not strongly dependent on the direction of the incoming wave; that is, it should be *omnidirectional*, although it is desirable to discriminate against low-elevation signals, which are likely to be contaminated by spurious reflected waves. The directional properties of a simple dipole antenna (a straight conductor) or loop antenna rule them out; most antennas in general use are microstrip antennas. The antenna section of the receiver may include a pre-amplification stage and a down-conversion of the frequency before transmission to the radio-frequency section. Some units are designed to receive only the primary L_1 frequency, whereas others receive both L_1 and L_2.

In the radio frequency section, the phase of each carrier frequency is tracked using a phase-lock loop, in which the phase of a controlled oscillator is locked to that of the received carrier using the output of a phase-comparator in a feedback loop. The separation of signals from the different satellites is achieved using correlation techniques on the C/A pseudorandom number codes. The locally generated code is automatically shifted in time to produce a maximum correlation with the incoming signal; the time shift gives, aside from clock errors etc., the wave propagation delay. An important indicator of the degree of sophistication and cost of a receiver is the number of satellites it can track simultaneously. Either the signal from each of four satellites is directed along separate parallel circuits, constituting four channels, or the same channel may be used to sequentially process the signal from different satellites. Lower-cost units are of the latter type.

In addition to the radio frequency section, a microprocessor is incorporated in a GPS unit with memory, keyboard and display. This is necessary, of course, to make all the necessary corrections to the observed time delays, to use the broadcast ephemerides of the satellites, and to solve the equations to obtain the coordinates

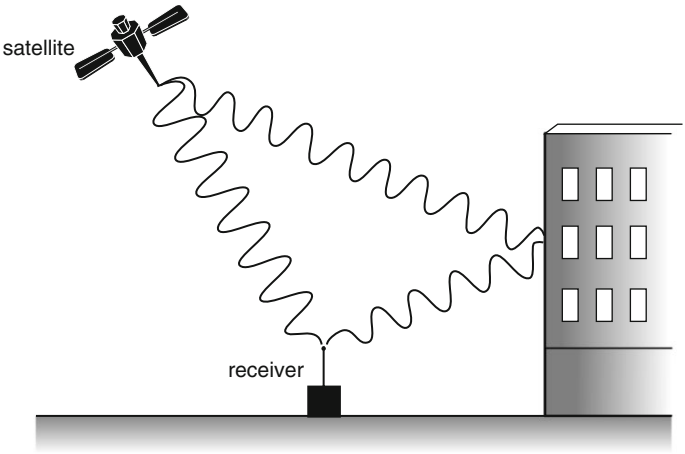

Figure 19.7 Multipath signals produced by reflections from the environment of the antenna

and time of the user. Care must be exercised in the choice of location of the antenna; even the best geometric and electrical design of the antenna will be to no avail if the physical surroundings can reflect portions of the wavefront, causing them to arrive at the antenna along different paths, as shown in Figure 19.7. The differing delays thus produced in the arrival time of the signal are called *multipath* errors.