Chapter 1
Introduction to GPS

The NAVSTAR Global Positioning System (GPS) is a satellite-based radio-positioning and time-transfer system designed, financed, deployed, and operated by the U.S. Department of Defense. GPS has also demonstrated a significant benefit to the civilian community who are applying GPS to a rapidly expanding number of applications. What attracts us to GPS is:

- The relatively high positioning accuracies, from tens of metres down to the millimetre level.
- The capability of determining velocity and time, to an accuracy commensurate with position.
- The signals are available to users anywhere on the globe: in the air, on the ground, or at sea.
- Its is a positioning system with no user charges, that simply requires the use of relatively low cost hardware.
- It is an all-weather system, available 24 hours a day.
- The position information is in three dimensions, that is, vertical as well as horizontal information is provided.

The number of civilian users is already significantly greater than that of the military users. However, for the time being the U.S. military still operates several "levers" with which they control the performance of GPS (Section 1.2.3). Nevertheless, despite the handicap of GPS being a military system there continues to be tremendous product innovation within the civilian sector, and it is ironic that this innovative drive is partly directed to developing technology and procedures to overcome some of the constraints to GPS performance which have been applied by the system's military operators.

1.1 Introduction to the System Components

1.1.1 *System Design Considerations*

Development work on GPS commenced within the U.S. Department of Defense in 1973, the motivation being to develop an all-weather, 24-hour, global positioning system to support the positioning requirements for the armed forces of the U.S. and its allies. (For a background to the development of the GPS system the reader is referred to [1].) The system was therefore designed to replace the large variety of navigational systems already in use, and great emphasis was placed on the system's reliability and survivability. In short, a number of stringent conditions had to be met:

- suitable for all classes of platform: aircraft (jet to helicopter), ship, land (vehicle-mounted to handheld) and space (missiles and satellites),
- able to handle a wide variety of dynamics,
- real-time positioning, velocity and time determination capability to an appropriate accuracy,
- the positioning results were to be available on a single global geodetic datum,
- highest accuracy to be restricted to a certain class of user,
- resistant to jamming (intentional and unintentional),
- redundancy provisions to ensure the survivability of the system,
- passive positioning system that does not require the transmission of signals from the user to the satellite(s),
- able to provide the service to an unlimited number of users,
- low cost, low power, therefore as much complexity as possible should be built into the satellite segment, and
- total replacement of the Transit[1] satellite and other terrestrial navaid systems.

This led to a design based on the following essential concepts:

- A one-way ranging system, in which the satellites transmit signals, but are unaware of who is using the signal (no receiving function). As a result the user (or listener) cannot easily be: (a) detected by the enemy (military context), or (b) charged for using the system (civilian context).
- Use of the latest atomic clock and microwave transmission technology, including spread-spectrum techniques.
- A system that makes range-like measurements with the aid of pseudo-random binary codes

modulated on carrier signals.
- Satellite signals that are unaffected by cloud and rain.
- A multiple satellite system which ensures there is always a sufficient number of satellites visible simultaneously anywhere on the globe, and at any time.
- Positioning accuracy degradation that is graceful.

What was perhaps unforeseen by the system designers was the power of product innovation, which has added significantly to the versatility of the GPS as a system for precise positioning and navigation. For example, GPS is able to support a number of positioning and measurement modes in order to satisfy simultaneously a variety of users, from those satisfied with general navigational accuracies (tens of metres) to those demanding very high (sub-centimetre) relative positioning accuracies. *It has now so penetrated certain applications areas that it is difficult for us to imagine life without GPS!*

Rarely have so many seemingly unrelated technological advances been required to make a complex system such as GPS work. Briefly they are:

Space System Reliability: The U.S. space program had by 1973 demonstrated the reliability of space hardware. In particular, the Transit system had offered important lessons. The Transit satellites were originally designed to last 2-3 years in orbit, yet some of the satellites have operated well beyond their design life. In fact Transit continued to perform reliably for over 25 years.

Atomic Clock Technology: With the development of atomic clocks a new era of precise time-keeping had commenced. However, before the GPS program was launched these precise clocks had never been tested in space. The development of reliable, stable, compact, space-qualified atomic frequency oscillators (rubidium, and then cesium) was therefore a significant technological breakthrough. The advanced clocks now being used on the GPS satellites routinely achieve long-term frequency stability in the range of a few parts in 10^{14} per day (about 1 sec in 3,000,000 years!). This long-term stability is one of the keys to GPS, as it allows for the autonomous, synchronised generation and transmission of accurate timing signals by each of the GPS satellites without continuous monitoring from the ground.

Quartz Crystal Oscillator Technology: In order to keep the cost of user equipment down, quartz crystal oscillators were proposed (similar to those used in modern digital watches), rather than using atomic clocks as in the GPS satellites. Besides their low cost, quartz oscillators have excellent short-term stability. However, their long-term drift must be accounted for as part of the user position determination process.

Precise Satellite Tracking and Orbit Determination: Successful operation of GPS, as well as the Transit system, depends on the precise knowledge and prediction of a satellite's position with respect to an earth-fixed reference system. Tracking data collected by ground monitor stations is analysed to determine the satellite orbit over the period of tracking (typically one week). This reference ephemeris is extrapolated into the future and the data is then up-loaded to the satellites. Prediction accuracies of the satellite coordinates, for one day, at the few metre level have been demonstrated.

Spread-Spectrum Technology: The ability to track and obtain any selected GPS satellite signal (a receiver will be required to track a number of satellites at the same time), in the presence of considerable ambient noise is a critical technology. This is now possible using spread-spectrum and pseudo-random-noise coding techniques.

Large-Scale Integrated Circuit Technology: To realise the desired low cost, low power and small size necessary for much of the user equipment, the GPS program relies heavily on the successful application of VLSI circuits, and powerful computing capabilities built onto them.

The GPS system consists of three segments (figure 1.1). (Good general references on the GPS system are [2,3].):
- The **Space Segment**: comprising the satellites and the transmitted signals.

- The **Control Segment**: the ground facilities carrying out the task of satellite tracking, orbit computations, telemetry and supervision necessary for the daily control of the space segment.
- The **User Segment**: the entire spectrum of applications equipment and computational techniques that are available to the users.

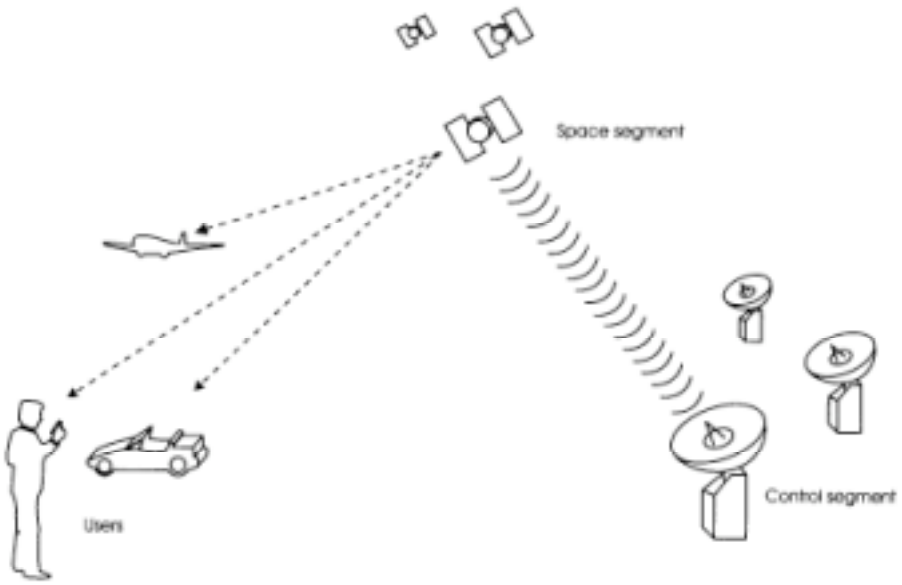

Figure 1.1: GPS System Elements.

1.1.2 *The Space Segment*

The Space Segment consists of the constellation of spacecraft and the signals broadcast by them which allow users to determine position, velocity and time. The basic functions of the satellites are to:

- Receive and store data transmitted by the Control Segment stations.
- Maintain accurate time by means of several onboard atomic clocks.
- Transmit information and signals to users on two L-band frequencies.
- Provide a stable platform and orbit for the L-band transmitters.

Several constellations of GPS satellites have been deployed, and several more are planned. The *experimental* satellites, the so-called "Block I" satellites, were built by the Rockwell Corporation. The first was launched in February 1978, and the last of the eleven satellite series (one exploded on the launchpad) was launched in 1985. The *operational* series of GPS satellites, the "Block II" and "Block IIA" satellites, were also built by the Rockwell Corporation. The 20 *replacement* "Block IIR" series of satellites, first launched in 1997, are built by the General Electric Corporation (now the Lockheed Martin Corporation). The "Block IIF" series are still in the design phase and may, for example, incorporate an additional civilian transmission frequency. They are planned for launch from 2005 onwards. The operational satellite I.D.s are separated into three space vehicle numbering series: SVN 13 through 21 for the Block II, SVN 22 through 40 for Block IIA, and SVN 41 and above for the Block IIR satellites.

The current status of the GPS constellation and such details as the launch and official commissioning date, the orbital plane and position within the plane, the satellite I.D. number(s), etc., can be obtained from several electronic GPS information sources on the Internet. Section 1.2.1 describes the general satellite orbit characteristics.

Each GPS satellite transmits a unique navigational signal centred on two L-band frequencies of the electromagnetic spectrum, permitting the ionospheric propagation effect on the signals to be eliminated. At these frequencies the signals are highly directional and so are easily reflected or blocked by solid objects. Clouds are easily penetrated, but the signals may be blocked by foliage (the extent of blockage is dependent on the type and density of the leaves and branches, and whether they are wet or dry). The signal is transmitted with enough power to ensure a minimum signal power level of -

160dBw at the earth's surface (the maximum it is likely to reach is about -153dBw, see [3]). The satellite signal consists of the following components (figure 1.2):

- The two L-band carrier waves.
- The ranging codes modulated on the carrier waves.
- The so-called "navigation message".

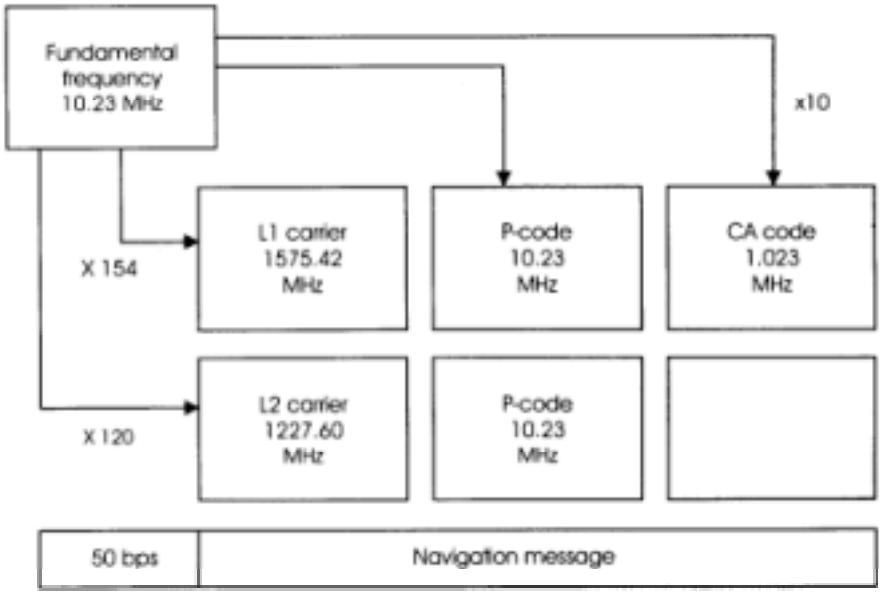

Figure 1.2: GPS Satellite Signal Components.

Modulated onto the carrier waves are the PRN ranging codes and navigation message for the user. The primary function of the ranging codes is to permit the *signal transit time* (from satellite to receiver) to be determined. The transit time when multiplied by the velocity of light then gives a measure of the receiver-satellite "range" (in reality the measurement process is considerably more complex). The navigation message contains the satellite orbit information, satellite clock parameters, and pertinent general system information necessary for real-time navigation to be performed. All signal components are derived from the output of a highly stable atomic clock. Each GPS satellite is equipped with several cesium and rubidium atomic clocks.

1.1.3 *The Control Segment*

The Control Segment consists of facilities necessary for satellite health monitoring, telemetry, tracking, command and control, satellite orbit and clock data computations, and data uplinking. There are five ground facility stations: Hawaii, Colorado Springs, Ascension Island, Diego Garcia and Kwajalein. All are owned and operated by the U.S. Department of Defense and perform the following functions:

- All five stations are *Monitor Stations*, equipped with GPS receivers to track the satellites. The resultant tracking data is sent to the Master Control Station.
- Colorado Springs is the *Master Control Station* (MCS), where the tracking data are processed in order to compute the satellite ephemerides and satellite clock corrections. It is also the station that initiates all operations of the space segment, such as spacecraft manoeuvring, signal encryption, satellite clock-keeping, etc.
- Three of the stations (Ascension Is., Diego Garcia, and Kwajalein) are *Upload Stations* allowing for the uplink of data to the satellites. The data includes the orbit and clock correction information transmitted within the navigation message, as well as command telemetry from the MCS.

Overall operation of the Control and Space Segments is the responsibility of the U.S. Air Force Space Command, Second Space Wing, Satellite Control Squadron at the Falcon Air Force Base, Colorado.

Each of the upload stations can view all the satellites once a day. All satellites are therefore in contact with an upload station three times a day, and new navigation messages as well as command telemetry can be transmitted to the GPS satellites every eight hours if necessary. The computation of: (a) the satellite orbits or "ephemerides", and (b) the determination of the satellite clock errors, are the most important function of the Control Segment. The first is necessary because the GPS satellites function as "orbiting control stations" and their coordinates must be known to a relatively high accuracy, while the latter permit a significant measurement bias to be reduced.

The GPS satellites travel at high velocity (of the order of 4 km/sec), but within a more or less regular orbit pattern. After a satellite has separated from its launch rocket and it begins orbiting the earth, it's orbit is defined by its initial position and velocity, and the various force fields acting on the satellite. In the case of the gravitational field for a spherically symmetric body (a reasonable approximation of the earth at the level of about 1 part in 10^3) this produces an elliptical orbit which is fixed in space -- the *Keplerian ellipse*. Due to the effects of the other, non-spherical gravitational components of the earth's gravity field, and non-gravitational forces, which perturb the orbit, the actual trajectory of the satellite departs from the ideal Keplerian ellipse (figure 1.3).

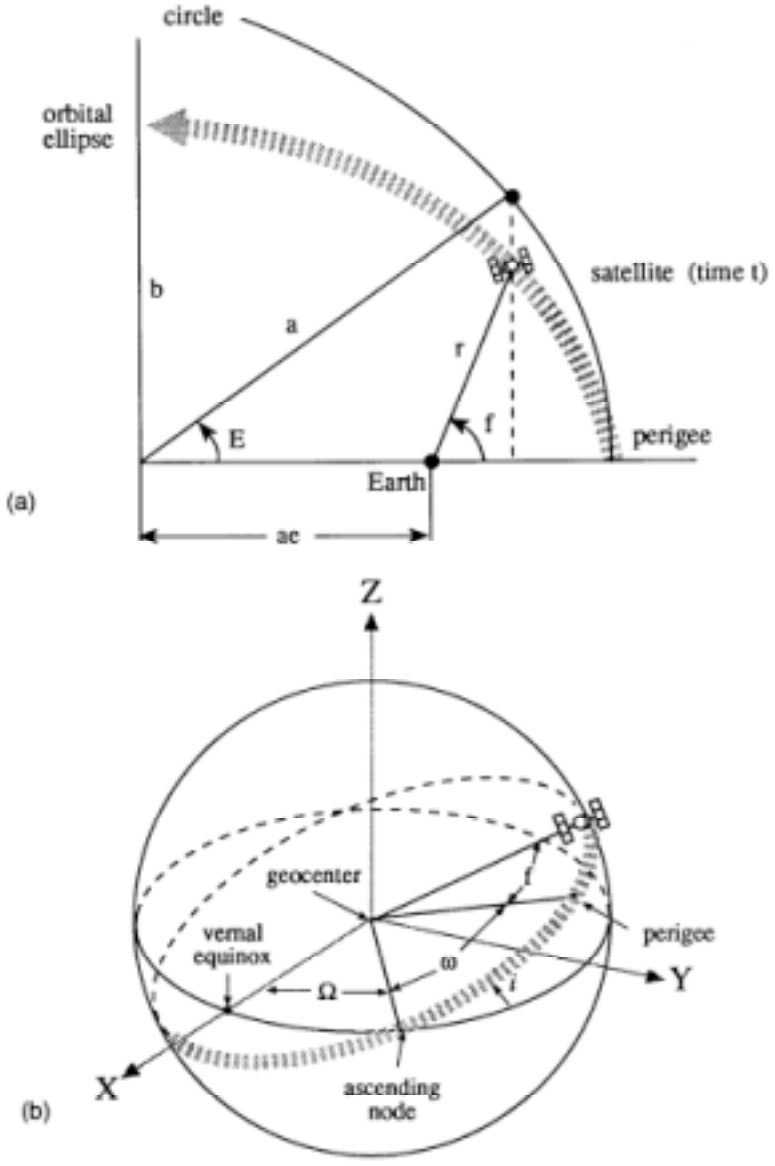

Figure 1.3: (a) The Keplerian Ellipse, and (b) Keplerian Orbital Elements.

The most significant forces that influence satellite motion are:

- the spherical and non-spherical gravitational attraction of the earth,
- the gravitational attractions of the sun, moon, and planets (the "third body" effects),
- atmospheric drag effects,
- solar radiation pressure (both direct and albedo effects), and
- the variable part of the earth's gravitational field arising from the solid earth and ocean tides.

To determine the motion of a satellite to a high precision these *perturbing* forces need to be modelled accurately. If these forces were known perfectly and the initial position and velocity of the satellite were given, then the integration of the Equations of Motion would give the satellite's position and velocity at any time in the future.

However, the perturbing forces are not known to sufficient precision. An "orbit computation" process is therefore performed, in which satellite observations obtained at tracking sites of known position (in the case of GPS, the monitor stations of the Control Segment) are analysed in order to produce an orbit that is a "best fit" to the available observations. This involves the adjustment of the appropriate parameters of the orbit, possibly together with several additional force model parameters (see [4]). Determining the orbit is a complex procedure, and in the case of GPS satellites this process occurs on a continuous, automatic basis.

The product of the orbit computation process at the MCS is the *satellite ephemeris* (or trajectory). A satellite ephemeris may be expressed in a number of forms:

- a list of 3-D coordinates, and velocities, at regular intervals of time,
- the Keplerian elements at some reference epoch, plus their rate-of-change with time,
- a polynomial representation of the trajectory in a suitable reference system, such as along-track, cross-track, or radial components, or
- satellite position and velocity at some reference time epoch, and requiring these values to be derived for subsequent times by integrating the Equations of Motion.

The GPS broadcast ephemeris, as represented in the navigation message, is actually a combination of all of the above orbit representations ([12]). The orbital ephemerides are expressed in the reference system most appropriate for positioning, which is an *earth-fixed* reference system such as WGS84. Hence the Control Segment has the function of propagating the satellite datum (Section 1.2.5), which users connect to via the transmitted satellite ephemerides.

The behaviour of each GPS satellite clock is monitored against GPS Time, as maintained by an ensemble of atomic clocks at the GPS Master Control Station. The satellite clock *bias, drift* and *drift-rate* relative to GPS Time are explicitly determined in the same procedure as the estimation of the satellite ephemeris. The clock behaviour so determined is made available to all GPS users via clock error coefficients (defining the mis-synchronization with GPS Time) in a polynomial form broadcast in the navigation message. However, what is available to users is really a *prediction* of the clock behaviour for some future time interval. Due to random deviations -- even cesium and rubidium oscillators are not entirely predictable -- the deterministic models of satellite clock error are only accurate to about 20 nanoseconds. This is not precise enough for accurate range measurement (see Section 1.3.6).

As the GPS system matures we can expect that the satellites will operate with greater independence from the ground-based Control Segment, without significant degradation in performance. For example the "Block IIR" and "Block IIF" satellites will have a crosslink capability enabling between-satellite communication and ranging. *The satellites will talk to each other!* This data will be processed to produce the ephemeris information within the space segment, with relatively little operator control having to be exercised.

1.1.4 *The User Segment -- The Applications*

This is the part of the GPS system with which we are most concerned -- the space and control segments being largely transparent to the operations of the navigation function. Of interest is the range of GPS:

- Applications,

- Equipment, and
- Positioning strategies.

The "engine" of commercial GPS product development is, without doubt, the *user applications*. Each day new applications are being identified, each with its unique requirements with regards to: accuracy of the results, reliability, operational constraints, user hardware, data processing algorithms, latency of the GPS results, etc. To make sense of the bewildering range of GPS applications it may be useful to classify them according to the following:

(1) Land, Sea and Air Navigation and Tracking, including enroute as well as precise navigation, collision avoidance, cargo monitoring, vehicle tracking, search and rescue operations, etc. *While the accuracy requirement may be modest and the user hardware is generally comparatively low cost, the reliability, integrity and speed with which the results are needed is generally high.*
(2) Surveying and Mapping, on land, at sea and from the air. Includes geophysical and resource surveys, GIS data capture surveys, etc. *The applications are of relatively high accuracy, for positioning in both the static and moving receiver mode, and generally require specialised hardware and data processing software.*
(3) Military Applications. *Although these are largely mirrored by civilian applications, the military GPS systems are generally developed to "military specifications" and a greater emphasis is placed on system reliability.*
(4) Recreational Uses, on land, at sea and in the air. *The primary requirement is for low cost instruments which are very easy to use.*
(5) Other specialised uses, such as time transfer, attitude determination, spacecraft operations, atmospheric studies, etc. *Obviously such applications require specially developed, high cost systems, often with additional demanding requirements such as real-time operation, etc.*

GPS user equipment has undergone an extensive program of development, both in the military and civilian area. In this context, GPS "equipment" refers to the combination of:
- hardware,
- software, and
- operational procedures or requirements.

While the military R&D programs have concentrated on achieving a high degree of miniaturisation, modularisation and reliability, the civilian user equipment manufacturers have, in addition, sought to bring down costs and to develop features that *enhance* the capabilities of the positioning system. The following general remarks can be made. Civilian users have, from the earliest days of GPS availability, demanded ever increasing levels of performance, in particular higher accuracy and improved reliability. This is particularly true of the survey user seeking levels of accuracy several orders of magnitude higher than that of the navigation user. In some respects the GPS user technology is being driven by the precise positioning market -- in much the same way that automotive technology often benefits from car racing. Yet another major influence on the development of GPS equipment has been the increasing variety of civilian applications. For although there may exist a similar positioning accuracy requirement across many user applications, to address a particular application in the most satisfactory manner, a specific combination of hardware and software features is often required.

It is expected that the worldwide market for GPS receiver equipment will grow from about US$1 billion at the present time, to over US$8 billion by the year 2000! Market surveys suggest that the greatest growth is expected to be in the commercial and consumer markets such as ITS applications, integration of GPS and cellular phones, and portable GPS for outdoor recreation and similar activities. These could account for more than 60% of the GPS market by the turn of the century. There are at present over 100 manufacturers of GPS instruments of varying kinds -- GPS instrumentation is discussed further in Section 1.4.

1.1.5 *The User Segment -- Positioning Principles*

The basic concept of GPS positioning is that of *positioning-by-ranges*. The geometrical principles of positioning can be demonstrated in terms of the intersection of locii. In the two-dimensional case, a measured range to a known point constrains the position to lie on circle with the measured range as radius. In three dimensions a measured range to a known point constrains the position in 3-D space to

lie on the surface of a sphere centred at the known point, with radius being the measured distance (figure 1.4). In the case of GPS, the distance measurement is made to a satellite with known position (coordinates are obtained from the satellite ephemeris data transmitted within the navigation message), however the principle applies to any range measuring positioning system, terrestrial or satellite-based.

In two dimensions, position can be defined as the intersection of two circles, involving distances d_1 and d_2 to two known points, as shown in figure 1.5. Note that there are two possible solutions, only one of which is correct. In general one solution can be discarded rather easily through apriori knowledge of approximate position and velocity. Another possibility is to measure another range to a third point and if all ranges are measured without error the intersection of three LOPs is a single uniquely defined point.

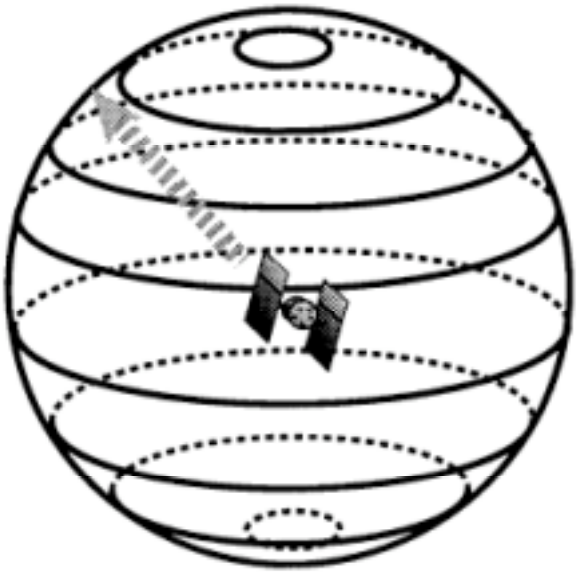

Figure 1.4: Surfaces of Position for Range Measurements.

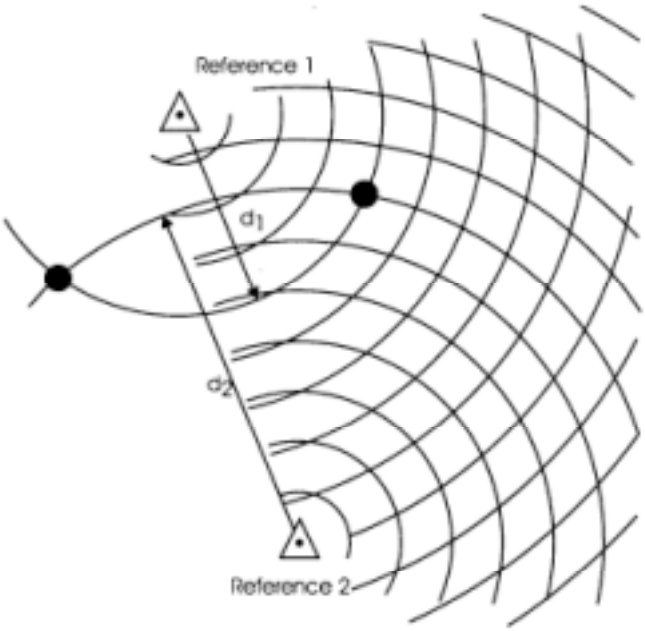

Figure 1.5: The Intersection of Circular Lines of Position for 2-D Positioning.

In the three-dimensional case, the intersection of three spheres describes two points in space, only one of which is correct (figure 1.6). Hence, a minimum of three ranges are required, to three separated known points, in order to solve the 3-D position problem. The quality of the positioning solution is dependent, amongst other things, on the accuracy with which the ranges can be measured and the geometry of the intersection.

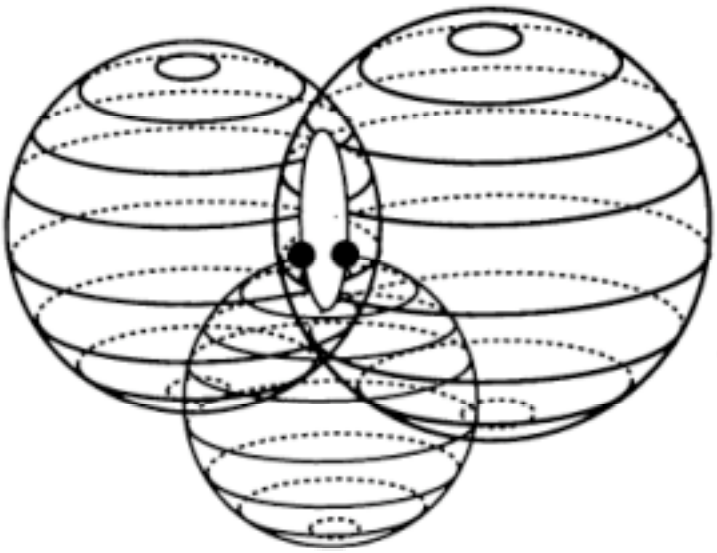

Figure 1.6: Intersection of Surfaces of Position Based on Range Measurements.

If the point being positioned is stationary, the two (or three) ranges do not need to be measured simultaneously. If the point is moving however, all ranges must be measured simultaneously (or over an interval of time during which the point has not moved by an amount greater than the uncertainty of the "fix"). Because the GPS constellation was designed to ensure at least four satellites are always visible anywhere on the earth, satellite positioning using simultaneously measured ranges is the basic positioning strategy for most navigation applications.

However, there is still the issue of how to account for *measurement biases*, as the technology used for making GPS range measurements does not give calibrated distance from the receiver to the satellite. Disturbing influences and errors in fact contaminate the range measurements to an unacceptable degree (in effect the radii of the spheres are incorrect -- Section 1.3.5), and hence the basic positioning principle is modified in several ways to satisfy the varying levels of accuracies required by different applications. (Chapter 2 describes some of the GPS enhancements that have been developed, or are under development.)

1.2 GPS Satellite Constellation and Signals

1.2.1 *GPS Constellation Design*

The operational Block II/IIA satellite constellation was to be fully deployed by the late 1980's. However, a number of factors, the main one being the Space Shuttle Challenger disaster (28 January 1986), has meant that the GPS system only became operational in the 1990's as far as most users were concerned. *Full Operational Capability* was declared on 17 July 1995 -- 24 Block II/IIA satellites operating satisfactorily. At an altitude of approximately 20,200km, a constellation of 24 functioning GPS satellites is sufficient to ensure that there will always be *at least four satellites visible*, at all unobstructed sites on the globe. Typically there are 6 to 10 satellites visible most of the day. The U.S. Department of Defense has undertaken to guarantee 24 satellite coverage 70% of the time, and 21 satellite coverage 98% of the time.

As the GPS satellites are in nearly circular orbits, at an altitude of approximately 20,200km above the earth (figure 1.7), this has a number of consequences:

- Their orbital period is approximately 11hrs 58mins, so that each satellite makes two revolutions in one sidereal day (the period taken for the earth to complete one rotation about its axis with respect to the stars).
- At the end of a sidereal day (23hrs 56mins in length) the satellites are again over the same position on earth.
- Reckoned in terms of a solar day (24hrs in length), the satellites are in the same position in the sky about four minutes earlier each day.
- The orbit groundtrack approximately repeats each day, except that there is a very small drift of the orbital plane to the west which is arrested by periodic manoeuvres.

Figure 1.7: The GPS Constellation "Birdcage".

The following general remarks can be made with regard to satellite constellation design for navigation purposes:

- The higher a satellite, the longer it is visible above the horizon (the extreme case is the geostationary satellites).
- The higher a satellite, the better the coverage due to longer fly-over passes and extended visibility of the satellite across large areas of the earth.
- The higher a satellite, the less the rate-of-change of distance, and the lower the Doppler frequency of a transmitted signal.
- The greater the angle of inclination, the more northerly the track of the sub-satellite point across the surface of the earth.
- No satellite can be seen simultaneously from all locations on the earth.
- Depending on the positioning principles being employed, there may be a requirement for observations to be made to more than one satellite simultaneously from more than one ground station.

The Block II satellites are deployed in six orbital planes at 60^{o} intervals about the equator, with each containing four satellites. The satellites can be moved round their orbits if it becomes necessary to "cover" for a failed satellite. The orbital planes are inclined at an angle of 55^{o} to the equatorial plane (figure 1.3).

As the satellites are at an altitude of more than three times the earth's radius, a particular satellite may

be above an observer's horizon for many hours, perhaps 6-7 hours or more in the one pass. At various times of the day, and at various locations on the surface of the earth, the number of satellites and the length of time they are above an observer's horizon will vary. Although at certain times of the day there may be up to 12 satellites visible simultaneously, there are nevertheless occasional periods of degraded satellite coverage (though naturally their frequency and duration will increase if satellites fail). "Degraded satellite coverage" is generally defined in terms of the magnitude of the Dilution of Precision (DOP) factor, a measure of the quality of satellite geometry. The highest the DOP value, the poorer the satellite geometry. For example, if all the visible satellites are located in the same part of the sky, the intersections of the SOPs will be very obtuse. (GPS DOPs are discussed further in Section 2.1.3.)

1.2.2 *GPS Signal Components*

The basis of the GPS signal are the two L-band carrier signals. These are generated by multiplying the fundamental frequency f_o (10.23MHz) by 154 and 120, yielding the two microwave L-band carrier waves L1 and L2 respectively (figure 1.2). The frequency of the two waves are: $f_{L1} = f_o \times 154 =$ 1575.42 MHz, and $f_{L2} = f_o \times 120 =$ 1227.6 MHz. These are radio frequency waves capable of transmission through the atmosphere over great distances, but which cannot penetrate solid objects. Note that all GPS satellites transmit carrier waves at the same two L-band frequencies (unlike the GLONASS system, where a different frequency is assigned to each satellite -- see Section 2.3.2). However, the L-band carrier waves themselves carry no information, and must be modified (or modulated) in some way. In the Global Positioning System the L-band carrier waves are modulated by two *ranging codes*, and the *navigation message*. The two distinct GPS ranging codes are:

- The **C/A code** (sometimes referred to as the "clear/access" or "coarse/acquisition" code), sometimes also referred to as the "S code".
- The **P code** (the "private" or "precise" code) was designed for use only by the military, and other authorised users.

The L1 carrier was designed to be modulated with both the P and C/A codes, whereas the L2 carrier would be modulated only with the P code. Under the policy of Anti-Spoofing (Section 1.2.3) the P code is encrypted through modulation by a further secret code (the "W code") to produce a new "Y code". Both carrier signals contain the navigation message.

The C/A and P (or Y) codes provide the means by which a GPS receiver can measure one-way ranges to the satellites. These codes have the characteristics of random noise, but are in fact binary codes generated by mathematical algorithms and are therefore referred to as "pseudo-random-noise" or PRN codes. Both the C/A and P code generating algorithms are known, and are based on a simple Tapped Feedback Shift Register scheme (see, for example [2,3]). One C/A code is assigned to each GPS satellite (the PRN code number is often used as the satellite I.D.). Each C/A code is a 1023 "chip" long binary sequence, generated at a rate of 1.023 million chips per second (that is, 1.023 MHz), thus the entire C/A code sequence repeats every millisecond.

The P code is a far more complex binary sequence of 0's and 1's, being some 267 days long with a chipping rate at the fundamental frequency f_o (10.23 MHz). The resolution of this code (length of the P code chip) is ten times the resolution of the C/A code. Instead of assigning each satellite a unique code, as is the case with the C/A code, the P code is allocated such that each satellite transmits a one week portion of the 267 day long PRN sequence, restarting the code sequence at the end of each week. Further details on how PRN codes are generated are given in, for example, [5,6].

To measure one-way range (from satellite to receiver), a knowledge of the ranging code(s) is required by the user's receiver. Knowing which PRN code is being transmitted by a satellite means that a receiver can generate a local replica of the same code sequence. These PRN codes possess a very important attribute: a given C/A (or P or Y) code will fully correlate with an exact replica of itself only when the two codes are aligned, and has a low degree of correlation with other alignments.

As the same P (or Y) code is synchronously modulated on both carrier waves, any difference in signal transit time of the same PRN sequence is due to the *retardation of the two L-band signals by a different amount* as they travel through the ionosphere. The effect of the ionosphere on signal propagation is essentially a function of signal frequency (Section 2.1.1), hence measurement on both

frequencies is a very effective way of overcoming the ionospheric signal delay.

1.2.3 *The Civilian - Military Relationship*

Although GPS is a military navigation system the civilian sector represents an important (and rapidly growing) user group that has increasingly lobbied the U.S. Government in order to influence: (a) the direction of GPS system development; (b) official GPS policy concerning enhancement and control; and (c) the design of follow-on systems to GPS for the 21st century. Several policy decisions have already been made which impact on GPS performance. Some of these actions were agreed to during the early system design phase, while others were made only after much of the present system had been deployed:

- Two PRN ranging codes are implemented. The C/A code is intended for general, civilian use, while the P code was reserved for military and other authorised users. Because of the P code's higher measurement resolution it was expected that the accuracy of positioning using the P code would be much better than that possible using the C/A code. However, it was found that the performance of C/A code positioning was often no worse than that of P code positioning by a factor of two. (The latest C/A code tracking technology permits ranging quality almost as good as P code ranging and hence, all other things being equal, positioning accuracy close to that expected from the processing of P code ranges should be possible.)

- These two levels of positioning performance were designed into the GPS system from the very beginning. The positioning service based on using C/A code ranging data is referred to as the "Standard Positioning Service" (SPS), while the service based on P code ranging data is known as the "Precise Positioning Service" (PPS).

- As a result of the demonstration of a surprisingly good level of SPS accuracy, the policy of "Selective Availability" (SA) was endorsed in order to artificially widen the gap between the two levels of positioning ([7]). SA is an intentional degradation of the accuracy of GPS horizontal positioning to 100m and vertical positioning at the 160m (at the 95% confidence level) for SPS users. SA is implemented through an encryption of the navigation message whereby a part of the transmitted ephemeris and satellite clock data is falsified (the so-called "epsilon" effect) and the satellite clock is "dithered" (the so-called "delta" effect). SA therefore affects the precision of all measurements, code and carrier phase. SA does not affect PPS users who have user equipment able to decipher the correct ephemeris and clock error data.

- Any GPS hardware manufacturer is able to construct a P code ranging receiver (the P code PRN generation algorithm is published). The non-military market for P code receivers was always assumed to be very small, and one that could be controlled by the U.S. Government through the issuance of "export licences", etc. However, a significant demand by the land surveying market for dual-frequency phase tracking GPS receivers (code-correlating phase tracking receivers need a knowledge of the P code PRN in order to make a carrier phase measurement) led to an expansion in the production of P code capable positioning equipment.

- Under another policy known as "Anti-Spoofing" (AS), access is denied to the P code modulated on both L-band frequencies. AS was implemented on 31 January 1994, through the encryption of a secret "W code" onto the P code. The rationale behind this decision was that by keeping the military PRN code secret, an enemy of the U.S. could not jam the signal using a ground-based transmitter, nor "spoof" military GPS receivers by transmitting a false P code signal from a satellite. However, several GPS receiver manufacturers have developed proprietary techniques for making dual-frequency measurements even in the presence of AS.

- Dual-frequency observations will lead to more accurate positioning results than single frequency observations, because the ionospheric bias can be eliminated from the code range measurements. The fact that the C/A code is only modulated on the L1 carrier is therefore an intentional design decision to ensure that the SPS service cannot deliver the accuracy that the PPS service can, even with SA off.

It must be emphasised that the situation as far as the policies on SA and AS is under almost

continuous review, particularly since the Presidential Decision Directive on the "U.S. National GPS Policy" was released in March 1996. The reader is referred to [8,9,10], which describe both the "official" policies and options on GPS, and report the debate and outcome of studies for alternative models of joint civilian-military GPS operation. The Vice President in early 1999 also announced "GPS Modernization Plans", which include the transmission of a third frequency.

1.2.4 *Why is the GPS Signal is so Complicated?*

GPS was designed as a complex military navigation system, which would also be used by civilians, hence we can summarise the reasons why the signal is so complicated as:

(a) The requirement for GPS to be *multi-user system* is most easily satisfied if it designed as *a listen-only, or self-positioning system.* An unlimited number of users could be supported because no return signal was necessary, however, the measurements and positioning procedures must then be based on *one-way (satellite-to-receiver) signals.*

(b) The very important requirement for *real-time positioning* influenced a number of design criteria related to the satellite signals:
- Simultaneous measurements from many satellites, but all signals are at the same frequency (though Doppler-shifted) -- *need to identify signals by use of different codes*
- Unambiguous range measurements were required -- *need to determine signal delay*
- Satellite positions are needed -- *broadcast ephemerides*

Hence the PRN codes were used to distinguish the different satellite signals. The measurement of a "range" to a particular satellite required that a comparison be made of the PRN code generated within the receiver (corresponding to the satellite in question) to the PRN code sequence contained within the incoming satellite signal (see Section 1.3.2).

(c) The requirement for GPS to support *high accuracy positioning* also had implications for satellite signal design:
- High frequency modulation -- *P code at 10 MHz*
- Dual-frequency signal -- *permits ionospheric delay estimation*
- Microwave carrier frequency -- *1.2 to 1.6 GHz*

(d) The *anti-jamming* requirement could be partly met through the use of *spread-spectrum codes*. This also fulfilled the requirement for a low power signal that could still be detected even though it was below the ambient signal noise level. By ensuring that some of the PRN codes are secret (known only to military and a few authorised users), jamming or "spoofing" of signals was made much more difficult. (However, it is possible to jam a GPS receiver if the jamming source is within several metres of the receiver.)

(e) As the system was intended as a "dual-use" technology, satisfying both *military and civilian users*, this had several implications the most important of which was that the civilians were not to have the same level of accuracy as the military users:
- The use of two PRN codes -- *P and C/A code*
- Restriction on the dual-frequency signal for civilians -- *C/A code only on the L1 frequency*
- Microwave carrier frequency -- *1.2 to 1.6 GHz*

1.2.5 *GPS Satellite Ephemerides*

How are the GPS satellite ephemerides computed? As the forces of gravitational and non-gravitational origin perturb the motion of the GPS satellites, the coordinates of the satellites in relation to the WGS84 reference system (Section 3.1) must be continually determined through the analysis of tracking data. In the case of the GPS broadcast ephemeris, this procedure is a three-step process ([6]):
- An off-line orbit determination is performed through the analysis of tracking to generate a "reference orbit". This is an initial estimate of the satellite trajectories computed from about one week's of tracking from the five Control Segment monitor stations.
- An on-line daily updating of the reference orbit using a Kalman filter as new data are added.

This provides the current estimates of the satellite orbit which is used to predict the future orbit.

- The ephemeris is estimated for 1 to 14 days into the future. To obtain the necessary broadcast information, curve fits are made to 4 to 6 hour portions of the extrapolated ephemeris, and hourly orbit parameters determined.

Note that these parameters are not true Keplerian elements as they only describe the ephemeris over the interval of applicability and not for the whole orbit. (Although only intended for use during the transmission period, they do, however, adequately describe the orbit over intervals of 1.5 to 5 or more hours, with a graceful degradation in accuracy.). The user can derive the earth-centred, earth-fixed WGS84 Cartesian coordinates of the GPS satellite from the broadcast orbital parameters, using an algorithm described in [11], and implemented in every GPS receiver.

1.3 GPS MEASUREMENTS

1.3.1 *The Transmitted Signal*

The signal that actually leaves a GPS satellite antenna is a combination of the three components: carrier wave, ranging codes and navigation message. The generation of the signal to be transmitted is carried out in a number of steps, and relies on the fact that all the components are derived by multiplying or dividing the fundamental frequency (figure 1.2). There are two distinct procedures for the combination of signal components: (a) BPSK modulation of binary sequences onto the carrier waves, and (b) modulo-2 addition of binary sequences.

Biphase Shift Key Modulation (BPSK) is the technique used to added a binary signal to a sine wave carrier. This amounts to causing a 180° phase shift in the carrier each time the binary sequence undergoes a transition from "0" to "1", or "1" to "0". The P code plus navigation message is modulated on both the L1 and L2 carriers, while the C/A code plus navigation message is only modulated on the L1 carrier.

An example of binary-to-binary modification of codes is the modulo-2 addition of the navigation message data to the C/A code. Because of the difference in frequency between the C/A code and the message stream this has the effect of inverting 20 C/A code binary "states" whenever the data bit of the navigation message is equal to "1". Conversely, when the data bit is "0" the C/A code sequence remains unaffected. The same satellite message is also modulated onto the P code sequence using this modulo-2 addition procedure.

There are two main types of measurements that can be made on the GPS signals ([13]):

- range observations based on the PRN codes, sometimes referred to as "code range" or "code phase", and
- carrier phase observations, which are more precise range-type measurements, but which have a much higher degree of "ambiguity" than the code ranges (see Section 2.2.2).

1.3.2 *The GPS Range Measurements*

The PRN codes are accurate time marks that permit the receiver's navigation computer to determine the time of transmission for any portion of the satellite signal. Before considering how ranging is carried out we need to describe, in general terms, how the incoming satellite signal is processed within a GPS receiver. See references [2,3] for more details than are presented here!

Neglecting any discussion on extraneous noise (in relation to which PRN codes have special properties), and assuming that the only satellite signals received at the antenna originate from one GPS satellite, the following simplistic procedure is carried out within a receiver "channel" (Section 1.4.2). The L1 carrier modulated by the C/A code is converted to a signal of lower frequency and then mixed with a locally generated matching C/A code. The local C/A code is generated on a different time scale (due to non-synchronization of receiver clock to GPS Time and the travel time of the signal from the satellite to antenna) to that of the incoming C/A code. As soon as the incoming signal and the receiver C/A code sequences are aligned within the receiver the ones and zeros of the two codes cancel, leaving

the received carrier modulated only by the binary navigation message.

The extraction of the code range, or more precisely the determination of the amount by which the receiver generated PRN code must be shifted to align it with the incoming signal, is carried out with the aid of a PRN code correlator in some form "delay-lock loop" scheme (see, for example [13]). *How accurately is this carried out?* The C/A code has a sequence rate of 1.023 Mbps, corresponding to a resolution of about 300m (speed of light divided by the frequency). The P (or Y) code, on the other hand, has a so-called "chip" rate of 10.23 Mbps, and hence an effective resolution of about 30m. As a *rule-of-thumb*, alignment of the incoming and receiver generated codes is generally possible to within about 1-2% of the chipping rate, hence the measurement precision of C/A code ranging is of the order of 3-5m, and for P code ranging it is of the order of 0.3-0.5m. *However, under the policy of "Anti-Spoofing", the Y code is encrypted and is therefore not available to civilian applications.*

It should be noted, however, that modern "narrow correlator" C/A code measurement technology as implemented in several "top-end" GPS receivers has demonstrated ten times better correlation performance for the C/A code correlation than that quoted above.

1.3.3 *The GPS Carrier Phase Measurements*

The wavelengths of the carrier waves are very short (approximately 19cm for L1 and 24cm for L2) compared to the C/A and P code wavelengths. Assuming a measurement resolution of 1-2% of the wavelength, this means that carrier phase can be measured to millimetre precision compared with a few metres for C/A code measurements. Unfortunately a phase measurement is *ambiguous* as it cannot discriminate one (L1 or L2) wavelength from another. In other words, time of transmission information for the L-band signal cannot be imprinted onto the carrier wave as is done using PRN codes (this would be possible only if the PRN code frequency was the same as the carrier wave, rather than 154 or 120 times lower in the case of the P code, and 1540 or 1200 times lower for the C/A code). The basic phase measurement is therefore an angle in the range 0^o to 360^o. *It is nevertheless the basis for high precision GPS positioning (Section 2.2).*

1.3.4 *Ranging Using PRN Codes*

Consider for a moment a perfect system, where all satellite clocks are synchronized to the same time system: the GPS Time (GPST). Furthermore, the ground receiver's clock also maintains the same synchronization, and none of the clocks drift from this GPST scale. Now suppose the satellite starts transmitting its L1 carrier (modulated with the combined C/A PRN code and navigation data). At the same instant, the receiver begins generating the C/A PRN code corresponding to that particular satellite (see figure 1.8). Under these circumstances the satellite and receiver generated C/A codes would be output in unison. When the satellite signal is received, however, it will be lagging the receiver generated code due to the *signal transit time*. Multiplying the time offset required to align the two codes (in effect determining the signal transit time) by the speed of light yields the satellite to receiver distance.

Measuring ranges simultaneously in this fashion to three satellites would fix one's position at the intersection of three spheres of known radii (the satellite ranges), centred at each satellite whose coordinates can be calculated from the navigation message. In reality the situation is more complex:

- GPS receivers are equipped with crystal clocks that do not keep the same time as the more stable satellite clocks (the satellite clocks can be nearly synchronised to GPST using the clock correction model transmitted in the navigation message). *Consequently each range is contaminated by the receiver clock error*. This range quantity is therefore referred to as **pseudo-range**, and in order for the user to derive position from pseudo-range data, the receiver equipment is required to track (a minimum of) four satellites, and solve for four unknown quantities: the three-dimensional position components and the receiver-clock offset (from GPST) -- see Section 1.3.5. *This is the basis of GPS real-time navigation*, and why GPS could be considered an example of a time-difference-of-arrival system.

- There is in fact a 300 km "ambiguity" in the C/A code pseudo-range measurements (300 km is the approximate length of the C/A code sequence). That is, all measured "distances" appear to

have a range of 0 to 300 km. This ambiguity is resolved in a number of ways, but the easiest to assume that if the approximate receiver position is known to within say 100 km, the "missing" component of the distance can be determined, and hence the raw pseudo-range measurement can be corrected for this ambiguity to obtain the full satellite-receiver distance.

- Ranging (and hence receiver position determination) can be carried out using the C/A code or the P code. P code ranging can be done on the combination of the two frequencies, hence eliminating the bias due to ionospheric refraction. Furthermore, the C/A code is "coarser", and hence the C/A derived ranges are subject to greater measurement "noise". The absence of a C/A code on L2 is intentional, as one of the accuracy limitations of the GPS system for the general class of civilian users.

- As previously mentioned, this differentiation between ranging codes, and the formulation of policies for their use (in peacetime and in times of global emergencies), is responsible for the provision of two GPS services: The Precise Positioning Service based on P or Y code (dual-frequency) ranging, and the Standard Positioning Service based on single frequency C/A code ranging (Section 1.2.3).

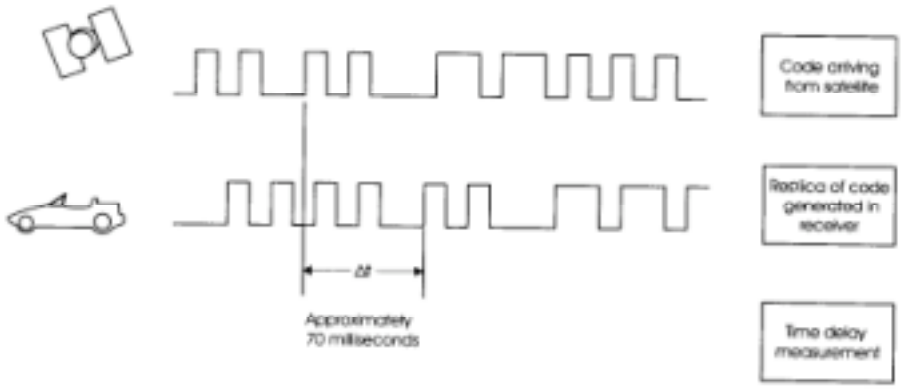

Figure 1.8: One-Way Ranging Using PRN Codes.

1.3.5 *An Observation Model of the Pseudo-Range*

The observation equation for a receiver-clock-biased range is (see, for example [14]):

$$p = \rho + \delta_{rc}(t_r).c \qquad (1.1)$$

where c is the velocity of electromagnetic radiation in a vacuum (or simply the "velocity of light"), δ_{rc} is the receiver clock error caused by the receiver oscillator (assume satellite clock time is "true" time) at time of reception t_r, p is the measured range and ρ is the true "geometric" range. Each observation made by the receiver can be parameterised as follows:

$$(x^s - x)^2 + (y^s - y)^2 + (z^s - z)^2 = (p - \delta_{rc}.c)^2 \qquad (1.2)$$

where x^s, y^s, z^s is the coordinate of the satellite and x, y, z is the coordinate of the receiver. Note that the time argument has been discarded (satellite coordinates are time-dependent, and so are the receiver coordinates if the receiver is moving).

As the 3-D coordinate of the satellite is known, then each measurement p contains four parameters which may be considered unknown: the 3-D coordinate of the receiver (x, y, z) and the receiver

clock error (δ_{rc}). By making four measurements, to four different satellites, the following system of equations is obtained:

$$\begin{aligned}
(x^{s1} - x)^2 + (y^{s1} - y)^2 + (z^{s1} - z)^2 &= (p^1 - \delta_{rc}.c)^2 \\
(x^{s2} - x)^2 + (y^{s2} - y)^2 + (z^{s2} - z)^2 &= (p^2 - \delta_{rc}.c)^2 \\
(x^{s3} - x)^2 + (y^{s3} - y)^2 + (z^{s3} - z)^2 &= (p^3 - \delta_{rc}.c)^2 \\
(x^{s4} - x)^2 + (y^{s4} - y)^2 + (z^{s4} - z)^2 &= (p^4 - \delta_{rc}.c)^2
\end{aligned} \qquad (1.3)$$

which has a unique solution (see, for example [15]). If more than four measurements are made the method of Least Squares, or other estimation techniques, can be used to determine the optimum solution. *Does the receiver clock error have to be estimated at each epoch?* That depends upon such factors as:

- How well the clock error is estimated.
- How often the position solution is carried out.
- The stability of the clock.

Under the assumptions that (a) there is a range measurement uncertainty of about 1 metre attributable to receiver "noise", and (b) the receiver is equipped with a quartz crystal clock (stability of 0.1 nanoseconds/second), then the uncertainty of the clock time after 30 seconds is as great as the range measurement error. Hence, once the clock error were determined, it would have to be independently re-estimated at least every 30 seconds, otherwise it would dominate the range error budget. This is one of the reasons why this procedure could not be used if pseudo-range measurements to different satellites were not made "simultaneously" (here, within 30 seconds of each other), so that the clock error could be assumed to be a constant.

This is the conventional strategy used for pseudo-range-based GPS positioning: the receiver clock error is simply treated as an additional unknown and the *navigation problem can be considered as 4-D estimation*. However, the GPS satellite clock error is assumed to be a known quantity, and parameters defining this clock error are transmitted in the navigation message. The ionospheric delay is also modelled using transmitted parameters. An estimate of the tropospheric delay may be obtained from a simple tropospheric model such as Hopfield. All other biases are assumed to be insignificant compared to the measurement noise (that is, their impact on the position solution quality is negligible -- see Section 2.1.1 for a discussion).

1.3.6 *Coping with the Satellite Clock Bias*

As the satellite clock error is the largest source of GPS measurement bias it deserves closer study. Under the assumption that the satellite clock error is an *unknown* quantity, the observation equation for such a satellite-biased range:

$$p = \rho + \delta^{sc}(T^s).c \qquad (1.4)$$

δ^{sc} is the satellite clock error caused by the satellite oscillator not being synchronized to "true" time (GPST). p is the measured range, ρ is the true range and T^s is the time of transmission. Each observation made by the receiver can be parameterised as in equation (1.1), except for the replacement of the term δ^{sc} for δ_{rc}:

$$(x^s - x)^2 + (y^s - y)^2 + (z^s - z)^2 = (p - \delta^{sc}.c)^2 \qquad (1.5)$$

The 3-D coordinate of the satellite signal transmitter (x^s, y^s, z^s) is known, hence in the case of three range observations there are six unknowns in the system: the 3-D coordinate of the receiver (x_{r1}, y_{r1}, z_{r1}) and the three satellite clock error terms (δ^{sc1}, δ^{sc2}, δ^{sc3}). Hence, at first glance, six satellite-biased range observations are required to solve this positioning problem. It is not, however,

possible simply to make observations to more satellites as each new observation introduces a new satellite clock parameter. *There are two options for overcoming this dilemma.*

It is possible to take advantage of the fact that all observations made to a particular satellite are biased by the same amount (if made at the same time, or close enough together so that the satellite clock error can be assumed to have not changed by a significant amount). If three range observations are made from another station, whose coordinate is known (x_{r2}, y_{r2}, z_{r2}), then it is possible to obtain a system of six equations in six unknowns:

$$\begin{aligned}
(x^{s1} - x_{r1})^2 + (y^{s1} - y_{r1})^2 + (z^{s1} - z_{r1})^2 &= (p_{r1}^{s1} - \delta^{sc1}.c)^2 \\
(x^{s2} - x_{r1})^2 + (y^{s2} - y_{r1})^2 + (z^{s2} - z_{r1})^2 &= (p_{r1}^{s2} - \delta^{sc2}.c)^2 \\
(x^{s3} - x_{r1})^2 + (y^{s3} - y_{r1})^2 + (z^{s3} - z_{r1})^2 &= (p_{r1}^{s3} - \delta^{sc3}.c)^2 \\
(x^{s1} - x_{r2})^2 + (y^{s1} - y_{r2})^2 + (z^{s1} - z_{r2})^2 &= (p_{r2}^{s1} - \delta^{sc1}.c)^2 \\
(x^{s2} - x_{r2})^2 + (y^{s2} - y_{r2})^2 + (z^{s2} - z_{r2})^2 &= (p_{r2}^{s2} - \delta^{sc2}.c)^2 \\
(x^{s3} - x_{r2})^2 + (y^{s3} - y_{r2})^2 + (z^{s3} - z_{r2})^2 &= (p_{r2}^{s3} - \delta^{sc3}.c)^2
\end{aligned} \qquad (1.6)$$

for which a unique solution can be obtained. In a conceptual sense this is the basis of *differential GPS positioning*, although in reality there are several possible implementations (Section 2.1.2 and 2.2).

The other strategy for accounting for satellite clock error is for the GPS operators to periodically determine the clock error. As the satellite clocks have significantly better long-term drift characteristics than the receiver clocks, a suitable clock error model could be a time polynomial:

$$\delta^{sc} = a_0 + a_1 (t - t_{oc}) + a_2 (t - t_{oc})^2 \qquad (1.7)$$

where:
- a_0 is the clock bias term
- a_1 is the clock drift term
- a_2 is the clock drift-rate
- t is satellite clock time
- t_{oc} is some reference epoch for the definition of the coefficients

What is actually available to users via the navigation message is a *prediction* of the satellite clock error behaviour for some time into the future (24 hours or more). Such deterministic models of satellite clock error are accurate to about 20 nanoseconds, or about six metres in equivalent range, depending upon the time since last navigation message update. Hence the measured pseudo-ranges may be corrected for satellite clock error, and then the observation model in equation (1.1) can be used.

Selective Availability complicates matters because it is a further artificial "dithering" of the satellite clock causing an additional several tens of metres error in the pseudo-range measurement (Section 1.2.3).

1.4 GPS Instrumentation

The following components of a generic GPS receiver can be identified (figure 1.9):

- *Antenna* and *Preamplifier*: Antennas used for GPS receivers have broadbeam characteristics, thus they do not have to be pointed to the signal source like satellite TV receiving dishes. The antennas are compact and a variety of designs are possible. There is a trend to integrating the antenna assembly with the receiver electronics.

- *Radio Frequency Section* and *Computer Processor*: The RF section contains the signal processing electronics. Different receiver types use somewhat different techniques to process the signal. There is a powerful processor onboard not only to carry out computations such as

extracting the ephemerides and determining the elevation/azimuth of the satellites, etc., but also to control the tracking and measurement function within modern digital circuits, and in some cases to carry out digital signal processing.

- *Control Unit Interface*: The control unit enables the operator to interact with the microprocessor. Its size and type varies greatly for different receivers, ranging from a handheld unit to soft keys surrounding an LCD screen fixed to the receiver "box".

- *Recording Device*: in the case of GPS receivers intended for specialised uses such as the surveying the measured data must be stored in some way for later data processing. In the case of ITS applications such as the logging of vehicle movement, only the GPS-derived coordinates and velocity may be recorded. A variety of storage devices were utilised in the past, including cassette and tape recorders, floppy disks and computer tapes, etc., but these days almost all receivers utilise solid state (RAM) memory or removable memory "cards".

- *Power Supply*: Transportable GPS receivers these days need low voltage DC power. The trend towards more energy efficient instrumentation is a strong one and most GPS receivers operate from a number of power sources, including internal NiCad or Lithium batteries, external batteries such as wet cell car batteries, or from mains power.

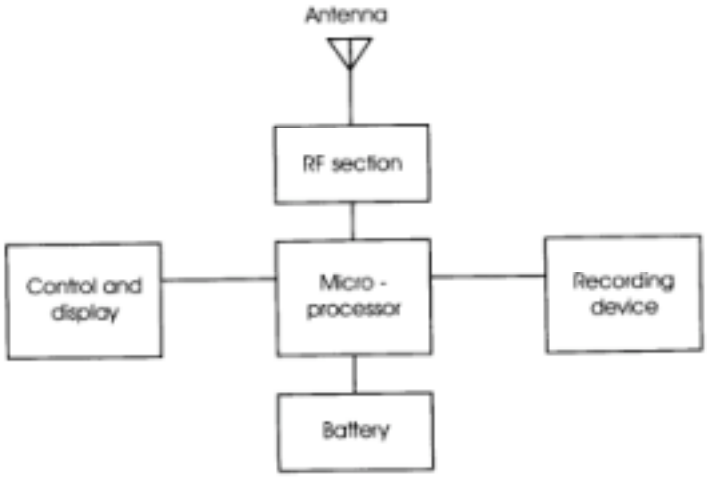

Figure 1.9: The Generic GPS Receiver.

The antenna and RF technology components are briefly discussed below. For further details the reader is referred to [6,16].

1.4.1 *Antennas*

The task of the antenna is to convert the energy of the arriving electromagnetic waves into an electric current that can be processed by the receiver electronics. There are a number of special considerations as far as antenna design is concerned:

- the antenna must be able to pick and discriminate very weak signals,
- the antenna may need to operate at just the L1 frequency, or at both the L1 and L2 frequencies,
- as the signals are right-hand circularly polarised, the GPS antenna must also be right-hand circularly polarised,
- antenna gain pattern that enhances the ability of the RF section to discriminate against multipath signals (such as left-hand circularly polarised signals),
- a stable electrical centre,
- low cost, and
- reliable.

There have been several types of GPS antennas used:

- monopole or dipole configurations,

- quadrifilar helices,
- spiral helices,
- microstrip,
- choke ring, and other multipath resistant designs

In short, the antennas are required to be rugged, simple in construction, have stable phase centres, be resistant to multipath, have good gain and pattern coverage characteristics. The microstrip antenna is almost universally used for navigation applications.

1.4.2 *Signal Processing Principles for Code-Correlating Receivers*

The RF section of a GPS receiver converts the incoming (preamplified) signal to a signal of a more manageable frequency. This intermediate frequency is obtained by mixing the incoming signal with a pure sinusoidal signal generated by the local oscillator (the quartz "clock"). The frequency of this beat frequency is the difference between the original (Doppler-shifted) received carrier frequency and the local oscillator. The intermediate or beat frequency is then processed by the signal tracking circuitry.

There are several classes of signal processing techniques that can be employed to make observations, as well as several proprietary implementations of tracking technologies. This is particularly the case for instruments which track carrier phase. It is beyond the scope of this chapter to discuss in detail the electronic circuitry and the reader is referred to [2,3]. As our interest is the pseudo-range measuring instruments, we need only consider the issues of the "code-correlating" signal processing technique, and some brief details of the processing *architecture* within the receiver, and in particular the form of the tracking "channel(s)".

Code-correlating receivers employ tracking loops to extract the useful measurements from the beat signal. A typical GPS receiver contains two types of tracking loops:

- the *delay-lock*, or code-tracking, loop, and
- the *phase-lock*, or carrier tracking, loop.

The delay-lock loop is used to make the alignment of the PRN code sequence (C/A or P code) that is contained in the signal from a satellite with an identical PRN code generated within the receiver. A correlator in the delay-lock loop continuously cross-correlates the two code streams, time shifting the receiver generated stream until alignment is obtained. The time shift is then converted to a pseudo-range measurement. Once the code-tracking loop is aligned, the PRN code can be stripped from the satellite signal. The resulting signal then passes to the phase-lock loop where the satellite message is extracted. Once the local oscillator is locked onto the satellite signal it will continue to follow the variations in the phase of the carrier as the satellite-receiver distance changes.

A GPS receiver may be described as *continuous* or *switching* depending on the type of channel(s) it has. A continuous-tracking receiver has dedicated hardware channels, and each channel tracks a single satellite, maintaining continuous code and/or phase lock on the signal. Each channel is controlled and sampled by the receiver's microprocessor with input/output operations being performed fast enough so that tracking is not disturbed. Continuous-tracking receivers may enjoy a signal-to-noise advantage over switching receivers in that the satellite signal is continuously available and may be more frequently sampled. A further advantage is a potential redundancy capability. Should one of the hardware channels fail, it may still be possible to obtain sufficient data to determine a position. One disadvantage of a multi-channel receiver is that the differences in signal path delay in the channels, the so-called inter-channel biases, must be well calibrated. This is the most common architecture used in GPS receivers today..

A *switching receiver* has hardware channels which sequentially sample the incoming signal from more than one satellite. There may be a single channel for all satellites, or each channel may track, say, two satellites. Code and/or carrier phase tracking for the individual signals is controlled by software (or, more usually, the "firmware") within the microprocessor. As a result, greater demands are placed on the microprocessor in a switching receiver and its programming is necessarily more complex -- in effect hardware complexity is exchanged for software complexity. If the cost of hardware components is a significant factor in determining the selling price of a receiver then in principle, the cost of a switching receiver should be cheaper than a continuous receiver (this is certainly the case for the low

cost GPS navigation units).

There are three basic kinds of switching receivers, distinguished by the time required to sequence through the signals tracked by a particular channel. A *multiplexing channel* is one for which the sequencing time to sample all satellites assigned to the channel is equal to 20 milliseconds, the period of one bit in the satellite navigation message. The sampling can be arranged so that no message bit boundary is spanned by any tracking interval. In this way, the messages from all satellites phase tracked by the channel can be read simultaneously. A multiplexing channel can be used to obtain both L1 and L2 data by alternating between the frequencies every 20 milliseconds. If a channel switches between signals at a rate which is asynchronous with the message bit rate, the channel is referred to as a *sequencing channel*. A fast sequencing channel is one which takes the order of a second or so to sequence through the signals. A slow sequencing channel may take several seconds or even minutes. A single sequencing channel would lose bits in a particular satellite message during those intervals spent sampling the signals from other satellites. Consequently, sequencing receivers may have an extra hardware channel just for message decoding. Alternatively, the navigation message must be decoded before the receiver starts the tracking cycle for real-time positioning (note that the message only changes once an hour).

The configuration of channels is selected so that, for example in the case of a navigation receiver, a minimum of four satellites can be tracked at the same time. This may call for a single switching channel, or two switching (between two satellite signals) channels, or even five channels (one used for calibrating the other four channels, and decoding the navigation message).

1.4.3 *Trends In GPS Instrumentation*

It is impossible to precisely predict trends in GPS instrumentation. The task is no easier if we focus our attention only to a specific market segment, such as positioning hardware for ITS applications. Nevertheless, speculation based on R&D activities being undertaken at present gives us some clues:

- The third generation of GPS receivers presently available already exhibit significant gains in miniaturisation, reduction in power consumption, and portability, over the earlier models. *We can expect this trend to continue.*

- The choice confronting GPS manufacturers is whether to maintain a broad-based development program, and market a product that is versatile enough to satisfy many applications (including the provision of interchangeable components such as antennas, RF units, memory, etc., to accomplish this), or to focus on specialist applications (military, navigation, differential navigation, surveying, kinematic, etc.). To date, most manufacturers are split between those with products for the high precision surveying market, and those focussing on the low cost, high volume navigation market. *Only a few address all markets.*

- There have been many predictions of low cost GPS receivers. However there is an enormous range of prices from the "top-of-the-line" surveying receivers to the "bare-bones" GPS "engines" implemented on one or more chips. GPS boardsets with basic navigation functionality are available at less than one hundred U.S. dollars each (when purchased in high volumes).

- There is a trend towards product differentiation, with many different configurations of tracking channels, data recording options, and, in particular, software options. Although some of these trends may be due to manufacturers wanting to give their product "an edge" in the marketplace, it is equally valid to suggest that this is in response to the different demands (some very specialist) of the market. *However, even in the case of complex ITS systems, it is possible to identify the "basic GPS component", and distinguish it from the "add-ons".*

- An exciting marriage is possible between satellite navigation and satellite communication, combining real-time positioning with instantaneous transmission of position. GPS receivers may be fitted to many different platforms (some unmanned, for example rail rolling stock and ship cargo containers), and their locations remotely monitored at a central site via a satcom link. In addition, for land navigation these could include electronic map displays to aid the driver. *What is clear is that for many applications the navigation data provided by a GPS receiver will*

merely be the first link in a complex information system.

- Market surveys suggest that the greatest growth is expected to be in the commercial and consumer markets. Consumer applications such as ITS, integration of GPS and cellular phones, and portable GPS for outdoor recreation and similar activities will account for more than 60% of the market by the year 2000..

- New tracking electronics, such as "narrow-correlator" technology, would improve pseudo-range measurement precisions, as would a combination of phase and pseudo-range measurement. Both are, however, still relatively expensive technologies and are not found on "bare-bones" GPS boardsets. Further electronic tracking refinements will lead to more multipath-resistant receivers.

References

[1] PARKINSON, B.W., 1994. GPS eyewitness: the early years. **GPS World, 5(9)**, 32-45.
[2] KAPLAN, E. (ed.), 1996. **Understanding GPS: Principles & Applications**. Artech House Publishers, Boston London, 554pp.
[3] SPILKER Jr., J.J. & PARKINSON, B.W. (eds.), 1995. **Global Positioning Systems: Theory & Applications**. American Institute of Aeronautics & Astronautics (AIAA), 1995, Vol.1(694pp), Vol.2(601pp).
[4] SEEBER, G., 1993. **Satellite Geodesy: Foundations, Methods & Applications**. Walter de Gruyter, Berlin New York, 531pp.
[5] SPILKER Jr., J.J., 1980. GPS signal structure and performance characteristics. In: Global Positioning System, papers published in **Navigation,** reprinted by the (U.S.) Inst. of Navigation, Vol.1, 29-54.
[6] WELLS, D.E., BECK, N., DELIKARAOGLOU, D., KLEUSBERG, A., KRAKIWSKY, E.J., LACHAPELKLE, G., LANGLEY, R.B., NAKIBOGLU, M., SCHWARZ, K.P., TRANQUILLA, J.M. & VANICEK, P., 1987. **Guide to GPS Positioning**. 2nd. ed. Canadian GPS Associates, Fredericton, New Brunswick, Canada, 600pp.
[7] GEORGIADOU, Y. & DOUCET, K.D., 1990. The issue of Selective Availability. **GPS World, 1(5)**, 53-56.
[8] N.R.C., 1995. The Global Positioning System: a shared national asset. Rept. by National Research Council, National Academy Press, 264pp.
[9] N.A.P.A, 1995. The Global Positioning System: charting the future. Rept. by National Academy of Public Administration & the National Research Council, National Academy Press, 332pp.
[10] GIBBONS, G., 1996. A national GPS policy. **GPS World, 7(5)**, 48-50.
[11] VAN DIERENDONCK, A.J., RUSSELL, S.S., KOPTIZKE, E.R. & BIRNBAUM, M., 1980. The GPS navigation message. In: Global Positioning System, papers published in **Navigation,** reprinted by the (U.S.) Inst. of Navigation, Vol.1, 55-73.
[12] LANGLEY, R.B., 1991b. The orbits of GPS satellites. **GPS World, 2(3)**, 50-53.
[13] LANGLEY, R.B., 1993. The GPS observables. **GPS World, 4(4)**, 52-59.
[14] LANGLEY, R.B., 1991d. Time, clocks, and GPS. **GPS World, 2(10)**, 38-42.
[15] LANGLEY, R.B., 1991c. The mathematics of GPS. **GPS World, 2(7)**, 45-50.
[16] LANGLEY, R.B., 1991a. The GPS receiver - An introduction. **GPS World, 2(1)**, 50-53.

Footnotes:

[1] Transit is an early satellite-based navigation system based on Doppler measurements (see [4]).