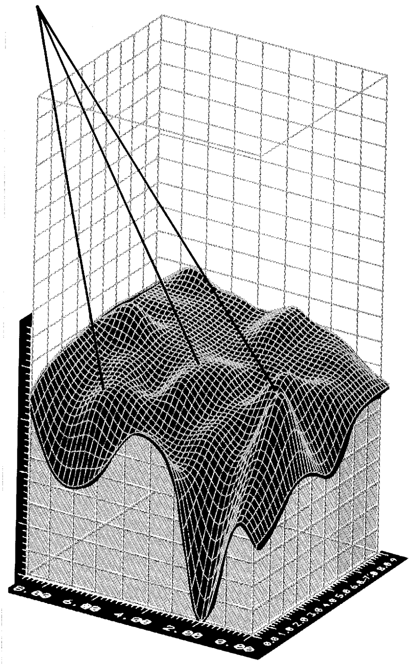
Why is the GPS Signal So Complex?

Richard B. Langley

Reprinted from GPS WORLD May/June 1990



Why is the GPS Signal So Complex?

Richard B. Langley

University of New Brunswick

In this issue we continue our introduction to GPS technology with a look at the signals transmitted by the satellites. A basic understanding of these signals is a prerequisite for getting the most out of a GPS receiver, whether it is used for navigating a fishing boat or for monitoring the deformations of the earth's crust. In answering the question posed in the title, we hope to uncover some of the mystery surrounding the GPS signal. This column is coordinated by Richard Langley and Alfred Kleusberg of the Department of Surveying Engineering at the University of New Brunswick. We welcome your comments and suggestions for future columns.

The radio signals intercepted by the antenna of a GPS receiver are amazingly complex. But this complexity gives GPS its versatility. Once fully operational, the NAVSTAR Global Positioning System will be available continuously, 24 hours a day, anywhere on or near the face of the earth, providing high accuracy positions and velocities in real time to both military and civilian users. To be able to provide such service, the GPS satellites must transmit signals that contain a number of different components. (The features that the GPS signal is called upon to provide are summarized in the sidebar accompanying this column.) Let's dissect the signal and see what it contains.

GPS is a one-way ranging system: signals are only transmitted by the satellites. This system characteristic is dictated primarily by the fact that GPS is, first and foremost, a military system. Military security requires that users not announce their position by sending signals to the satellites. However, the requirement that GPS be a passive system means that both the GPS receiver and the satellites must contain their own clocks that must be very closely synchronized. This synchronization is achieved using the signals themselves and is the reason why signals from four, rather than three, satellites are normally required to determine a three-dimensional position.

Another reason why GPS is a passive system is the need to serve a large number of users. The military requirement alone involves tens of thousands of users. Some projections place the eventual number of civilian users in the millions. For GPS satellites to process two-way communications from each user, the system would have to be much more complicated, which would likely reduce its versatility.

THE CARRIERS

Each GPS satellite transmits signals centered on two microwave radio frequencies: 1575.42 megahertz, referred to as Link 1, or simply L1; and 1227.60 megahertz, referred to as L2. These channels lie in a band of frequencies known as the L-band, which starts just above the frequencies used by cellular telephones. The International Telecommunications Union, the radio regulation arm of the United Nations, has set aside special subbands within the L-band for satellite-based positioning systems. The L1 and L2 frequencies lie within these bands.

Such high frequencies are used for several reasons. The signals, as we have said, consist of a number of components. A bandwidth of about 20 megahertz is required to transmit these components. This bandwidth is equal to the whole FM broadcast band! So a high, relatively uncluttered part of the radio spectrum is required for GPS-type signals.

GPS signals must provide a means for determining not only high-accuracy positions in real time, but also velocities. Velocities are determined by measuring the slight shift in the frequency of the received signals due to the Doppler effect – essentially the same phenomenon, albeit for sound waves, that gives rise to the change in pitch of a locomotive's whistle as a train passes in front of you at a level crossing. In order to measure velocities with centimeter-per-second accuracies, centimeter wavelength (microwave) signals are required.

Another reason for requiring such high frequencies is to reduce the effect of the ionosphere. As was mentioned in the last "Innovation" column, the ionosphere affects the speed of propagation of radio signals. The range between a satellite and a receiver derived by measuring travel times of the signal will therefore contain errors. The size of errors gets smaller as higher frequencies are used. But at the L1 frequency the error can still amount to 30 meters for a signal arriving from directly overhead.

For some GPS applications, an error of this size is tolerable. However, other applications require much higher accuracies. This is why GPS satellites transmit on two frequencies. If measurements made simultaneously on two well-spaced frequencies are combined, almost all of the ionosphere's effect can be removed.

Although high frequencies are desirable for the reasons just given, they should not be too high. For a given transmitter power, a received satellite signal becomes weaker as the frequency used becomes higher. The L-band frequencies used by GPS are therefore a good compromise between this so-called *space loss* and the perturbing effect of the ionosphere.

GPS signals, like most radio signals, start out in the satellites as pure sinusoidal waves or *carriers*. But pure sinusoids cannot be readily used to determine positions in real time. Although the phase of a particular cycle of a carrier wave can be measured very accurately, each cycle in the wave looks like the next one, so it is difficult to know exactly how many cycles lie between the satellite and the receiver. This ambiguity can be resolved using the differential technique pioneered by surveyors (see "GPS: A Multipurpose System" in the January/February issue of *GPS World*), but that can be time-consuming.

THE CODES

For a user to obtain positions independently in real time, the signals must be modulated; that is, the pure sinusoid must be altered in such a fashion that time-delay measurements can be made. This is achieved by modulating the carriers with *pseudorandom noise* (PRN) codes.

These PRN codes consist of sequences of binary values (zeros and ones) that, at first sight, appear to have been randomly chosen. But a truly random sequence can only arise from unpredictable causes over which, of course, we would have no control, and which we could not duplicate. However, using a mathematical algorithm or special hardware devices called *tapped feedback registers,* we can generate sequences that do not repeat until after some chosen interval of time. Such sequences are termed *pseudo*random. The apparent randomness of these sequences makes them indistinguishable from certain kinds of noise such as the hiss heard when a radio is tuned between stations or the "snow" seen on a television screen. Although noise in a communications device is generally unwanted, in this case the noise is very beneficial.

Exactly the same code sequences are independently replicated in a GPS receiver. By aligning the replicated sequence with the received one and knowing the instant of time the signal was transmitted by the satellite, the travel time, and hence the range, can be computed. Each satellite generates its own unique codes, so a GPS receiver can easily identify which signal is coming from which satellite, even when signals from several satellites arrive at its antenna simultaneously.

Two different PRN codes are transmitted by each satellite: the C/A, or coarse/acquisition, code, and the P, or precision, code. The C/A-code is a sequence of 1,023 binary digits, or *chips,* which is repeated every millisecond. This means that the chips are generated at a rate of 1,023 million per second and that one chip has a duration of about one microsecond. Each chip, riding on the carrier wave, travels through space at the speed of light. We can therefore obtain a unit of distance by multiplying the time interval by this speed. One microsecond translates to approximately 300 meters. This is the *wavelength* of the C/A-code. Because the C/A-code is repeated every millisecond, a GPS receiver can quickly lock onto the signal and begin matching the received code with the one generated by the receiver.

The precision of a range measurement is determined in part by the wavelength of the chips in the PRN code. Higher precisions can be obtained with shorter wavelengths. To get higher precisions than those afforded by the C/A-code, GPS satellites also transmit the P-code. The wavelength of the P-code chips is only 30 meters, one-tenth the wavelength of the C/A-code chips; the rate at which the chips are generated is correspondingly 10 times as fast: 10.23 million per second.

REQUIRED SIGNAL PROPERTIES

- Work with one-way measurements (receive only)
- Serve tens of thousands to millions of both military and civilian users
- Provide accurate, unambiguous, real-time range measurements
- Provide accurate Doppler-shift measurements
- Provide accurate carrier-phase measurements
- Provide a broadcast message
- Provide ionospheric delay correction
- Allow simultaneous measurements from many satellites
- Have interference protection
- Have multipath tolerance

The P-code is an extremely long sequence. The pattern of chips does not repeat until after 266 days, or about 2.35×10^{14} chips! Each satellite is assigned a unique one-week segment of this code, which is initialized at Saturday/Sunday midnight each week.

The GPS PRN codes have additional useful properties. When a receiver is processing the signals from one satellite, it is important that the signals received simultaneously from other satellites do not interfere. The GPS PRN codes have been specially chosen to be resistant to such interference. Also, the use of PRN codes results in a signal that is essentially impervious to unintentional or deliberate jamming from other radio signals, a possibility that the U.S. Department of Defense, the owner of the system, has to worry about.

At the present time, the C/A-code is modulated onto the L1 carrier, whereas the P-code is transmitted on both L1 and L2. This means that only users with dual-frequency GPS receivers can correct the measured ranges for the effect of the ionosphere. Users of single-frequency receivers must resort to models of the ionosphere that account for only a portion of the effect. Access to the lower accuracy C/A-code is provided in the GPS *Standard Positioning Service (SPS),* the level of service authorized for civilian users. The *Precise Position Service (PPS)* provides access to both the C/A-code and the P-code and is designed primarily for military users. The SPS incorporates a further intentional degradation of accuracy, called *selective availability,* which will be discussed in a future "Innovation" column.

THE BROADCAST MESSAGE

To convert the measured ranges between the receiver and the satellites to a position, the receiver must know where the satellites are. To do this in real time requires that the satellites broadcast this information. Accordingly, there is a message superimposed on both the L1 and L2 carriers along with the PRN codes. Each satellite broadcasts its own message which consists of orbital information (the *ephemeris*) to be used in the position computation, the offset of its clock from GPS system time, and information on the health of the satellite and the expected accuracy of the range measurements.

The message also contains *almanac* data for the other satellites in the GPS constellation, as well as their health status and other information. The almanac data, a crude description of the satellite orbit, are used by the receiver to determine the location of each satellite. The receiver uses this information to quickly acquire the signals from satellites that are above the horizon but are not yet being tracked. So, once one satellite is tracked and its message decoded, acquisition of the signals from other satellites is quite rapid.

The broadcast message also contains another very important piece of information for receivers that track the P-code. As mentioned earlier, the P-code segment assigned to each satellite is seven days long. A GPS receiver with an initially unsynchronized clock has to search through its generated P-code sequence to try to match the incoming signal. It would take many hours to search through just one second of the code, so the receiver needs some help. It gets this help from a special word in the message called the *hand-over word (HOW),* which tells the receiver where in the P-code to start searching.

The GPS broadcast message is sent at a relatively slow rate of 50 bits per second, taking 12.5 minutes for all the information to be transmitted. To minimize the time it takes for a receiver to obtain an initial position, the ephemeris and satellite clock offset data are repeated every 30 seconds.

The C/A-code and P-code chip streams are separately combined with the message bits using *modulo 2 addition.* This is just the binary addition that computers and digital electronics do so well. If the code chip and the message bit have the same value (both 0 or both 1), the result is 0. If the chip and bit values are different, the result is 1. The carriers are then modulated by the code and message composite signal.

This process is readily accomplished for the L2 channel because it carries only the P-code. But the L1 channel has to carry both the P-code and the C/A-code. This result is achieved by a clever technique known as *phase quadrature*. The P-code signal is superimposed on the L1 carrier in the same way as on the L2 carrier. To get the C/A-code signal on the L1 carrier, the unmodulated carrier is tapped off and shifted in phase by 90 degrees. This quadrature carrier component is mixed with the C/A-code signal and then combined with the P-code modulated in-phase component before being transmitted by the spacecraft antenna.

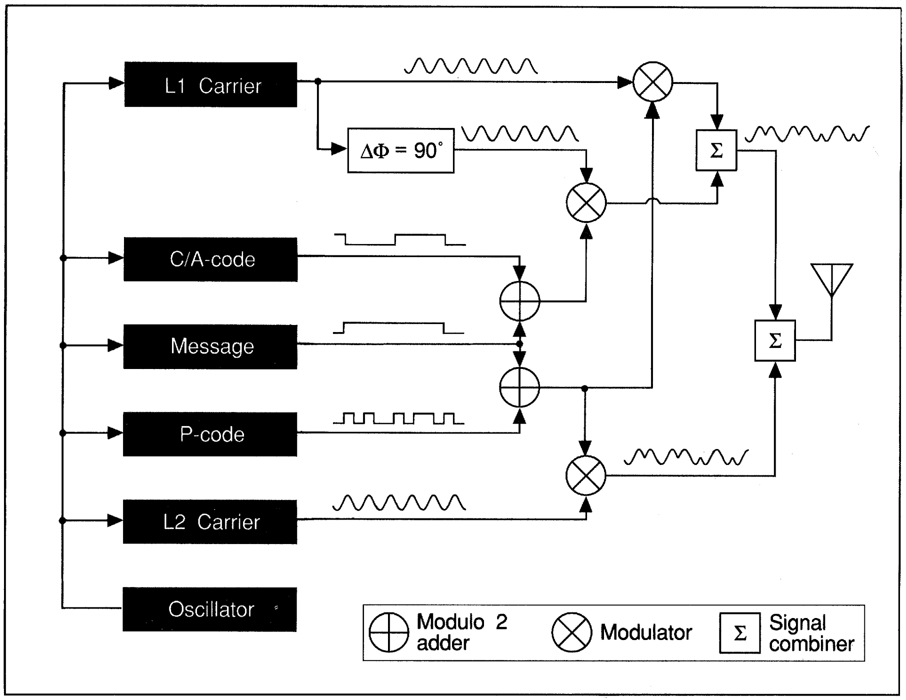


Figure 1: How the different components of the GPS signal are combined. Note that the various waveforms are not to scale.

BINARY BIPHASE MODULATION

Carrier waves can be modulated in a number of ways. The amplitude of the carrier can be varied, the frequency can be varied, or the phase can be varied. Phase modulation is the approach used for the GPS signals. Because the PRN codes and the message are binary streams, there must be two states of the phase modulation. These two states are the normal state, representing a binary 0, and the mirror-image state, representing a binary 1. The normal state leaves the carrier unchanged. The mirror-image state results in the unmodulated carrier being multiplied by -1. Therefore, a code transition either from 0 to 1 (normal to mirror image) or from 1 to 0 (mirror image to normal) involves a phase reversal or a phase shift of 180 degrees. This technique is known as *binary biphase modulation*.

An interesting property of binary biphase modulation was exploited by one of the first commercially available GPS receivers, the Macrometer. By electronically squaring the received signal, all of the modulation was removed, leaving a pure carrier. The phase of the carrier could then be measured to give the ambiguous range measurements used by surveyors. Of course, the broadcast message was lost in the process and, consequently, orbit data had to be obtained from an alternate source.

The timing and frequency control for the carriers, the PRN codes, and the message all come from an atomic oscillator on board the satellite. Each satellite carries four oscillators, one of which is selected by the GPS Master Control Station to control the signal.

The composite GPS signal then consists of carriers modulated by the PRN C/A- and P-codes and the broadcast message. The combining of these different components is illustrated in Figure 1.

Forgetting for a moment that GPS is a ranging system, we could consider the satellites to be simply broadcasting a message in an encoded form. The bits of the message have been camouflaged by the PRN code chips. The effect of this camouflaging is to increase the bandwidth of the signal. Instead of occupying only a fraction of one kilohertz, the signal has been spread out over 20 megahertz.

Inside a GPS receiver, the code-matching operation "de-spreads" the signal, allowing the message to be recovered. Clearly, this can only be done if the receiver knows the correct codes. The de-spreading operation conversely spreads out any interfering signal, considerably reducing its effect. Common in military circles for ensuring security and combating interference, this technique is known as *direct-sequence spread-spectrum communication*. Spread-spectrum signals have the additional property of limiting the interference from signals reflected off nearby objects — a phenomenon known as *multipath*.

It should be pointed out that the U.S. Department of Defense has reserved the right to encrypt the P-code — something it calls *anti-spoofing (AS)*. When AS is activated, the Y-code is transmitted instead of the P-code. The Y-code is created by combining a secret W-code with the normal P-code. Only authorized (i.e., military) PPS users will know the W-code. It is believed that AS will only be activated in times of national emergency and for brief test purposes.

CONCLUSION

We have examined the structure of the signal transmitted by a GPS satellite and have identified its components and their purpose. The NAVSTAR Global Positioning System is an extraordinarily versatile system that can serve the needs of navigators, surveyors, timekeepers, and others in both the civilian and military communities. The manifold nature of the GPS signal makes this all possible.

For further information on the structure of the GPS signals, readers might wish to consult the very detailed paper by J. J. Spilker in volume 1 of the (U.S.) Institute of Navigation series on GPS or the *Guide to GPS Positioning* published by Canadian GPS Associates. ■