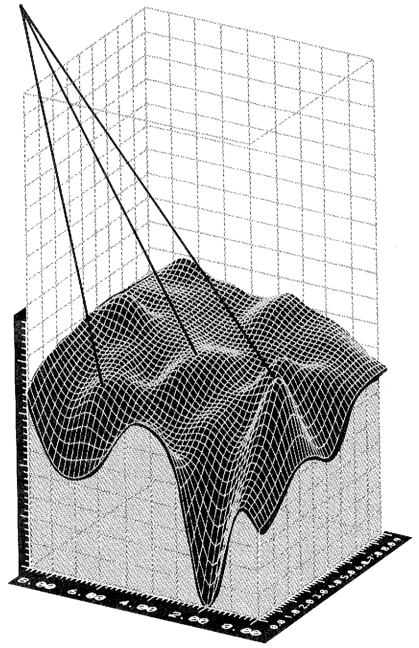
The GPS Receiver: An Introduction

Richard B. Langley
University of New Brunswick



In this issue, we resume our tutorial on different aspects of global positioning system technology with a look inside the GPS receiver. As with any piece of equipment, the more we know about how a GPS receiver works, the better prepared we are to judge the performance of a particular unit, assess its strengths and weaknesses, and compare the capabilities of the receivers in the ever-increasing product lines of the various manufacturers.

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 of topics for future columns.

In 1980 only one GPS receiver was available on the commercial market. Ten years later, considerably more than 100 different makes and models have become available. There are receivers for the military; receivers for civilians; receivers for waypoint navigation, for geodetic surveying, and for time transfer; receivers that use the C/A-code and those that additionally use the P-code; single- and dual-frequency receivers; and hand-held receivers and others more substantial in size. Although these receivers have differences in their design, construction, and capabilities, they share a number of basic principles in their operation. This column looks at the makeup of a generic GPS receiver and points out some of the features that are available in commercial units.

A GPS receiver consists of a number of basic building blocks (see Figure 1): an antenna and associated preamplifier, a radio frequency (RF) front-end section, a signal tracker block, a command entry and display unit, and a power supply. The overall operation of the receiver is controlled by a microprocessor that computes the receiver's coordinates as well. Some receivers also include a data storage device, an output to interface the receiver to a computer, or both. We'll examine each of these components in turn, starting with the antenna.

THE ANTENNA

The job of the antenna is to convert the energy in the electromagnetic waves arriving from the satellites into an electric current that can be handled by the electronics in the receiver. The size and shape of the antenna are very important, as these characteristics govern in part the ability of the antenna to pick up the very weak GPS signals. The antenna may need to operate at just the L1 frequency or at both the L1 and L2 frequencies. Also, because GPS signals are circularly polarized, all GPS antennas must be circularly polarized as well. Despite these restrictions, several different types of antennas presently are available for GPS receivers. These include monopole or dipole configurations, quadrifilar helices (also known as volutes), spiral helices, and microstrips.

The microstrip, because of its ruggedness and relative ease of construction, is perhaps the most common antenna. It may be circular or rectangular in shape and is similar in appearance to a small piece of copper-clad printed circuit board. Made up of one or more patches of metal, microstrips are often referred to as *patch* antennas. They may have either single- or dual-frequency capability, and their exceptionally low profile makes them ideal for airborne and some hand-held applications.

Other important characteristics of a GPS antenna are its *gain pattern,* which describes its sensitivity over some range of elevation and azimuth angles, and its ability to discriminate against multipath signals, that is, signals arriving at the antenna after being reflected off nearby objects. A particularly significant factor for antennas used in very precise positioning applications is the stability of an antenna's phase center, the electrical center of the antenna to which the position given by a GPS receiver actually refers.

Some antennas, such as the microstrip, require a ground plane for proper operation. The ground plane is usually a flat or shaped piece of metal on which the actual microstrip element sits. In high-precision surveying, the ground plane of the antenna is often extended with a metal plate or plates to enhance its performance in the presence of multipath signals.

GPS signals are very weak; they have roughly the same strength as those from geostationary TV satellites. The reason a GPS receiver does not need an antenna the size of those in some people's backyards has to do with the structure of the GPS signal and the ability of the GPS receiver to despread it (see "Why Is the GPS Signal So Complex?" in the May/June 1990 issue of *GPS World*). The power to extract a GPS signal out of the general background noise of the ether is concentrated in the receiver rather than in the antenna. Nevertheless, a GPS antenna must generally be combined with a low-noise preamplifier that boosts the level of the signal before it is fed to the receiver itself. In systems in which the antenna is a separate unit, the preamplifier is housed in the base of the antenna and receives power from the same coaxial cable along which the signal travels to the receiver.

THE RF SECTION

The RF section of a GPS receiver translates the frequency of signals arriving at the antenna to a lower one, called an intermediate frequency (IF), that is more easily managed by the rest of the receiver. This is done by combining the incoming signal with a pure sinusoidal signal generated by a component in the receiver known as a local oscillator. Most GPS receivers use precision quartz crystal oscillators, enhanced versions of the regulators commonly found in wristwatches. The IF signal contains all of the modulation that is present in the transmitted signal; only the carrier has been shifted in frequency. The frequency of the shifted carrier is simply the difference between the original received carrier frequency and that of the local oscillator. It is often called a *beat frequency* in analogy to the beat note that is heard when two musical tones very close together are played simultaneously. Most receivers employ multiple IF stages, reducing the carrier frequency in steps. The final IF signal passes to the workhorse of the receiver, the signal trackers.

THE SIGNAL TRACKERS

The omnidirectional antenna of a GPS receiver simultaneously intercepts signals from all satellites above the antenna's horizon. The receiver must be able to isolate the signals from each particular satellite in order to measure the code pseudorange and the phase of the carrier. This isolation is achieved through the use of a number of signal channels in the receiver. The signals from different satellites may be easily discriminated by the unique C/A-code or portion of the P-code they transmit and are assigned to a particular channel.

The channels in a GPS receiver may be implemented in one of two basic ways. A receiver may have *dedicated channels*, each continuously tracking a particular satellite. A minimum of four such channels tracking the L1 signals of four satellites would be required to determine three coordinates of position and the receiver clock offset. Additional channels permit tracking of more satellites or the L2 signals for ionospheric delay correction, or both.

The other channelization concept uses one or more *sequencing channels*. A sequencing channel "listens" to a particular satellite for a period of time, makes measurements on that satellite's signal, and then switches to another satellite. A single-channel receiver must sequence through four satellites to obtain a three-dimensional position "fix." Before a first fix can be obtained, however, the receiver has to dwell on each satellite's signal for at least 30 seconds to acquire sufficient data from the satellite's broadcast message. The time to first fix and the time between position updates can be reduced by having a pair of sequencing channels.

A variation of the sequencing channel is the *multiplexing channel*. With this design, a receiver sequences through the satellites at a fast rate, essentially acquiring all the broadcast messages from the individual satellites at the same time. With a multiplexing receiver, the time to first fix is 30 seconds or less, the same for a receiver with dedicated multiple channels.

Receivers with single channels are less expensive but, because of their slowness, are restricted to low-speed applications. Receivers with dedicated channels have greater sensitivity because they can make measurements on the signals more often, but they have inter-channel biases that must be carefully calibrated. This calibration is usually done by the receiver's microprocessor.

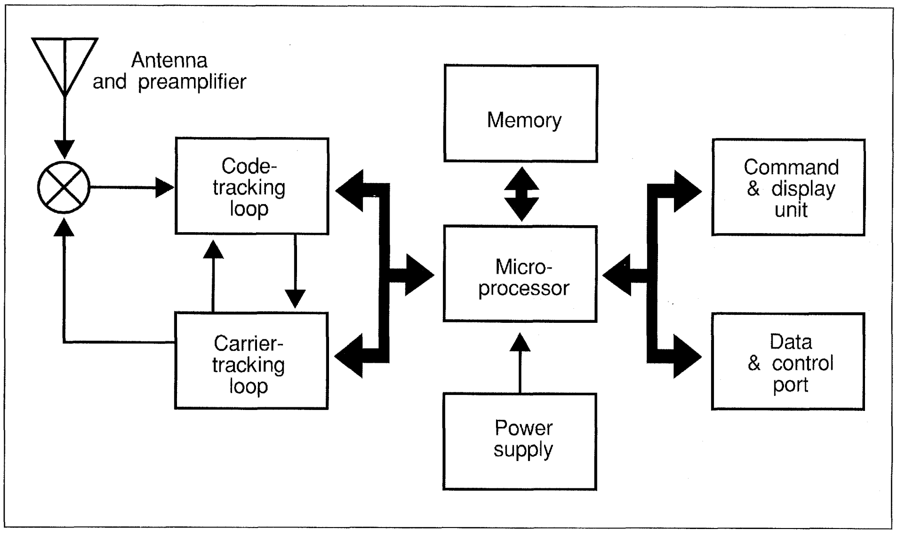


Figure 1. The major components of a generic one-channel GPS receiver.

The GPS receiver uses its tracking channels to make pseudorange measurements and to extract the broadcast message. This is done through the use of *tracking loops*. A tracking loop is a mechanism that enables a receiver to "tune into," or track, a signal that is changing either in frequency or in time. It is a feedback device that basically compares an incoming (external) signal against a locally produced (internal) signal, generates an error signal that is the difference between the two, and uses this signal to adjust the internal signal to match the external one in such a way that the error is reduced to zero or minimized. A GPS receiver contains two kinds 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 align a pseudorandom noise (PRN) code sequence (from either the C/A- or P-code) that is present in the signal coming from a satellite with an identical PRN code sequence generated within the receiver using the same algorithm that is employed in the satellite. Alignment is achieved by appropriately shifting the receiver-generated code chips in time so that a particular chip in the sequence is generated at the same instant its twin arrives from the satellite.

A correlation comparator in the delay-lock loop continuously cross-correlates the two code streams. This device essentially performs a multiply-and-add process that produces a relatively large output only when the code streams are aligned. If the output is low, an error signal is generated and the code generator is adjusted. In this way, the replicated code sequence is locked to the sequence in the incoming signal. The signals from other GPS satellites will have essentially no effect on the tracking process because the PRN codes of all the satellites were chosen to be orthogonal to each other. This orthogonality property means that a very low output is always produced by the correlator whenever the code sequences used by two different satellites are compared.

Because the P-code sequence is so long, a P-code tracking loop needs some help in setting its code generator close to the right spot for obtaining lock with the satellite signal. It gets this help from information included in the broadcast message available to the receiver by first tracking the C/A-code.

The time shift required to align the code sequences is, in principle, the time required for a signal to propagate from the satellite to the receiver. Multiplying this time interval by the speed of light gives us the distance or range to the satellite. But because the clocks in a receiver and in a satellite are, in general, not synchronized and run at slightly different rates, the range measurements are biased. These biased ranges are called *pseudoranges*. Because the chips in the satellite code sequences are generated at precisely known instants of time, the alignment of the receiver and satellite code sequences also gives us a reading of the satellite clock at the time of signal generation.

Once the code-tracking loop is locked, the PRN code can be removed from the satellite signal by mixing it with the locally generated one and filtering the resultant signal. This procedure despreads the signal, shrinking its bandwidth down to about 100Hz. It is through this process that the GPS receiver achieves the necessary signal-to-noise ratio to offset the gain limitation of a physically small antenna.

The despread IF signal then passes to the phase-lock loop, which demodulates, or extracts, the satellite message by aligning the phase of the receiver's local oscillator signal with the phase of the IF or beat frequency signal. If the phase of the oscillator signal is not correct, the demodulator in the phase-lock loop detects it and applies a correction signal to the oscillator. Once the oscillator is locked to the satellite signal, it will continue to follow the variations in the phase of the carrier as the range to the satellite changes. Most implementations of carrier tracking use the Costas loop, a variation of the phase-lock loop designed for binary biphase modulated signals such as those transmitted by GPS satellites.

Most GPS receivers use precision quartz crystal oscillators, enhanced versions of the regulators commonly found in wristwatches.

The carrier beat phase observable is obtained in principle simply by counting the elapsed cycles and by measuring the fractional phase of the locked local oscillator signal. The phase measurement, when converted to units of distance, is an ambiguous measurement of the range to the satellite. It is ambiguous because a GPS receiver cannot distinguish one particular cycle of the carrier from another and hence assumes an arbitrary number of full cycles of initial phase when it first locks onto a signal.

This initial ambiguity must be solved for mathematically along with the coordinates of the receiver if phase observations are used for positioning. Because this ambiguity is constant as long as the receiver maintains lock on the received signal, the time rate of change of the carrier phase is freed from this ambiguity. This quantity is related to the Doppler shift of the satellite signal and is used, for example, to determine the velocity of a moving GPS receiver such as that in an aircraft. After the carrier-tracking loop locks onto a satellite signal, the bits in the broadcast message are subsequently decoded using standard techniques of bit synchronization and a data detection filter.

The carrier beat phase can be measured by another method besides the code-tracking/Costas loop combination, but it comes with a penalty. This method is the so-called signal-squaring technique. The GPS signal is simply a constant carrier whose phase is shifted by exactly 180°, more than a million times each second, as a result of modulation by the PRN codes and the broadcast message. These phase reversals can be considered as a change in the amplitude of the signal from $+1$ to -1 or from -1 to $+1$, and the instantaneous amplitude is therefore either plus or minus one. Electronically squaring the signal results in a signal with a constant amplitude of unity, although with a frequency equal to twice that of the original. However, the phase of this signal is easily related to the phase of the original carrier. Of course, in the squaring process both the codes and the broadcast message are lost, so code-derived pseudorange measurements are not possible and the information in the message that describes the orbits, health, and other details about the satellites must come from another source. Some inherent signal-to-noise loss also occurs in the squaring process, which may result in phase measurements that are slightly noisier than those obtained with code tracking.

One of the first commercially available GPS receivers, the Macrometer, used the squaring technique, and a number of currently available dual-frequency receivers use this approach for measurements on the L2 frequency. A variation of this technique has been used in receivers that measure the *phase* of the code modulations without having to know the actual code sequences.

If antispoofing is turned on in the satellites, which results in an encryption of the P-code, then multiplying the signal by itself will be the only way to make measurements on the L2 frequency. Pseudorange measurements and the broadcast message would still be available from the code and carrier tracking on the L1 frequency.

THE MICROPROCESSOR

Although the bulk of a GPS receiver could be built using analog techniques, the trend in receiver development has been to make as much of the receiver as digital as possible, resulting in smaller, cheaper units. In fact, it is possible for the IF signal to be digitized and to perform the code and carrier tracking with software inside the microprocessor. Thus, in some respects, a GPS receiver may have more in common with a compact disc player than it does with an AM radio. Because a receiver has to perform many different functions, a GPS receiver's operation is controlled by a microprocessor. (Again, these functions include such things as acquiring satellite signals as quickly as possible once the receiver is turned on, tracking the codes and carriers of the signals, decoding the broadcast message, determining the user's coordinates, and keeping tabs on the other satellites in the constellation.) The microprocessor's software, which contains the instructions for running the receiver, is imbedded in memory chips within the receiver.

The microprocessor works with digital samples of pseudorange and carrier phase. These data samples are acquired as a result of analog-to-digital conversion at some point in the signal flow through the receiver. The receiver uses these samples to establish its positions and may record them for future processing. The microprocessor may run routines that do some filtering of these raw data to reduce the effect of noise or to get positions and velocities that are more reliable when the receiver is in motion.

The microprocessor may compute waypoint information or convert coordinates from the standard WGS 84 geodetic datum to a regional one. It also manages the input of commands from the user, the display of information, and the flow of data through its communication port (if the unit has one).

USER INTERFACE

The majority of self-contained GPS receivers have a keypad and a display of some sort to interface with the user. The keypad can be used to enter commands for selecting different options for acquiring data, for monitoring what the receiver is doing, or for displaying the computed coordinates, time, or other details. Users may also key in auxiliary information, such as that required for waypoint navigation or weather data and antenna height for geodetic surveying. Most receivers have well-integrated command and display capabilities with menus, prompting instructions, and even on-line help. Some receivers have a basic default mode of operation that requires no user input and can be activated simply by turning the receiver on.

Some GPS receivers are designed as sensors to be integrated into navigation systems and, therefore, don't have their own keypads and displays; input and output are accessed only via data ports.

DATA STORAGE AND OUTPUT

In addition to a visual display, many GPS receivers provide a means of saving the carrier-phase and pseudorange measurements and the broadcast messages for use in postprocessing. This feature is a necessity for receivers used for surveying and for differential navigation.

In surveying applications, the pseudorange and phase observations must be stored for combination with like observations from other simultaneously observing receivers and subsequent analysis. Usually the data are stored internally in the receiver using semiconductor memory. Some receivers store data on magnetic tape or directly on a floppy disk using an external microcomputer.

Some receivers, including those that store their data internally for subsequent analysis and those used for real-time differential positioning, have an RS-232 or some other kind of communications port for transferring data to and from a computer, modem, or data radio. Some receivers can be remotely controlled through this port.

THE POWER SUPPLY

Most GPS receivers have internal DC power supplies, usually in the form of rechargeable nickel-cadmium (NiCd) batteries. The newest receivers are designed to draw as little current as possible, extending the operating time between battery charges. Most receivers also make a provision for external power in the form of a battery pack or AC-to-DC converter.

CONCLUSION

The state-of-the-art GPS receiver is a remarkable achievement in electronic engineering. What took a rack of equipment costing a couple of hundred thousand dollars in 1980 can now be done with a hand-held unit costing less than $3,000. We will see further advances in GPS receiver technology in the next decade as even higher-speed, lower-power integrated circuits come out of the research labs. Particularly promising are the developments in fast microwave integrated circuits based on the compound gallium arsenide (GaAs) rather than traditional silicon. The first receivers using GaAs chips are, in fact, already on the market. We will have smaller, lighter, more rugged units with more options and better user interfaces that cost less money.

Where will the advances in GPS receiver design end up? Don't be surprised to see "Dick Tracy"–style wristwatch GPS receivers with integrated two-way voice communications at a cost of less than $500 before the 21st century is very old. What was the product of the fertile mind of a comic strip creator will soon become reality. ■