Navigating the GNSS Landscape for More Precise, Power-Efficient Receivers

By Andrew Brierley-Green, Senior MTS, Applications, Core Products Group, Maxim Integrated

Introduction

Now that we can easily access location and directions from our smartphones, it's even more amazing to consider the measures that ancient explorers had to take to find their way around, especially at sea when out of sight of land. Early navigation relied on astrolabes and then sextants to measure latitude. By the late 1800s, chronometers were widely used with astronomical observation to determine longitude for marine navigation. Satellites came into the navigation picture in 1964, with the Transit system (also called NAVSAT or NNSS). Today, a network of 31 satellites orbiting the Earth at an altitude of 20,000km serves as the U.S. Global Positioning System (GPS), providing directional guidance to anyone in the world with a GPS receiver.

The basic principle of the GPS system is that radio signals are transmitted from satellites to receivers on or close to the Earth. The positions of the satellites at any given instant are known, and the distance or "range" from the receiver to each satellite can be calculated from the propagation delay of the radio signal from that satellite. Knowing the distance to each of several reference points (i.e., satellites) allows calculation of the receiver's spatial coordinates.

Figure 1. Like chronometers used in the late 1800s for marine navigation, today's satellite navigation systems help us find our way.

The GPS system consists of three segments: space (satellites), control (ground stations), and user (receivers). The space segment is the constellation of satellites. The GPS core constellation features a baseline of 24 satellites, each orbiting the Earth in about 12 hours. In this system are six orbital planes with four satellites per plane, as well as extra satellites in orbit for redundancy. In the control segment, ground stations monitor the satellites, gathering information such as position in orbit, errors in satellite clocks, and signal delay due to atmosphere. The collected information is uploaded to the satellites and then retransmitted back to the receivers on Earth in the form of a navigation message (NAV).

Satellites provide the references for the position location, so a precise knowledge of any satellite's position in space as it orbits the Earth is required. The elliptical orbits around Earth that satellites follow are mostly determined by our planet's gravity. However, there are perturbations due to factors such as gravitational effects of the moon and sun, solar radiation pressure, and the Earth's non-uniform density. The orbital perturbations need to be accurately measured in order to yield sufficient accuracy in the estimate of the satellite's position.

GNSS receivers determine three spatial coordinates and the receiver clock offset. A technique called trilateration, which uses the geometry of circles and spheres, determines absolute or relative locations of points, based on measuring the distances to each satellite orbiting the Earth that is visible (above the horizon). For example, if the distance to a given satellite is calculated as 20,450km, then it is known that the receiver must lie somewhere on the surface of a sphere with a radius of 20,450km centered on this satellite. If the distance to a second satellite is calculated as 19,760km, then it is known the receiver must also lie somewhere on the surface of a sphere with a radius of 19,760km centered on the second satellite. Furthermore, since both conditions are simultaneously true, the receiver must be located on the intersection of these two spheres. Geometrically, the intersection of two spheres is a circle, so this constrains the receiver's position to somewhere on a circle. Extending this example to a third satellite, the position can be constrained to the intersection of two circles, or two points.

If one of these points is nonsensical, such as 400km inside the Earth, then by elimination, the receiver must be located at the other point.

A satellite orbit can be specified by a set of parameters called orbital elements. One of these elements is the eccentricity of the ellipse, for example. If not for the random perturbations, the position of a satellite in orbit at any time could be calculated precisely from the orbital elements. The general term for orbital data is ephemeris. The NAV contains the ephemeris, as well as the time (epoch) when the ephemeris data was obtained. Ground monitoring stations periodically measure the ephemeris of each satellite, uploading this data to each satellite. Every satellite broadcasts its ephemeris data and epoch in the NAV. Because of perturbations, orbital elements will have some inaccuracies. So, the NAV includes correction factors for these orbital perturbations.

Each satellite transmits a unique pseudo random noise (PRN) code, a repeating sequence of bits with a long enough repetition period that it has similar statistical properties to a truly random sequence. Using the PRN code, receivers can correlate the signal to the satellite. Assuming, for simplicity, that a receiver knows the satellite from which it is receiving the PRN, the receiver computes the autocorrelation of a locally generated replica of the PRN against the received PRN. By repeatedly shifting its local code replica and calculating the autocorrelation, eventually, the receiver finds a maximum correlation of 1.0. The receiver must synchronize its local code replica to the received code in order to despread the received data from the satellite. Once the receiver knows how much it must shift its local ranging code replica to obtain maximum correlation with received code, it then knows the propagation delay of the signal modulo 1ms. There is a "1% resolution" rule of thumb which states that the time delay can be measured to a resolution of 1%. For the GPS L1 C/A code, which is the most commonly used civilian GPS signal that all civilian receivers will decode, the rate of the PRN code is 1.023 million chips per second. The correlation process allows measurement of delay to 1% of one-chip accuracy corresponding to 0.01/1.023 × 106 = 9.8ns. At the speed of light, this corresponds to a range accuracy of 2.9m.

The other purpose for the PRN code is spread-spectrum modulation. Since all satellites are transmitting simultaneously at the same frequency, there must exist a means to prevent these signals from interfering with each other. The PRN codes are chosen such that the cross-correlation between any two codes is low. Multiplying two different PRN codes and integrating the result will result in a low value for any value of lag between them.

Types of Errors to Address for More Accurate Results

The distance to each satellite is calculated by simply multiplying the propagation delay of the radio wave by the speed of light in free space. However, rather than providing the true range, this will give an estimated range called the "pseudorange." For example, while the signal is travelling through free space for most of its journey from the satellite to the receiver, it also passes through the Earth's atmosphere. To the extent that the radio signal is slowed down by the atmosphere relative to its free-space velocity, this will result in an error in the calculated range. These error contributions must be compensated to produce a sufficiently accurate range estimate.

Each GPS satellite has a very accurate onboard atomic clock. The satellites are all synchronized to a common time-base called "GPS time." However, the atomic clocks have some drift and offset from GPS system time. Since range measurement is based on measurement of time, these errors in the clocks will cause error in the range measurement. To perform time error correction, ground stations monitor the satellite clock offset from GPS time, calculate correction terms, and upload them to the satellite. These are subsequently broadcast in the NAV. The receiver applies the correction terms received from the satellite, effectively compensating the offset of the satellite clock from GPS time.

Another large source of error is the local receiver clock offset. However, since this offset is common to all the signals received from the visible satellites, it can be calculated and compensated for. The three spatial coordinates and the receiver clock offset are the four variables that need to be solved by the receiver. For this reason, a minimum of four satellites is required for an accurate position fix.

Lonospheric error stems from the ionized plasma of free electrons and electrically charged atoms (ions) that extend from about 50 to 1000km in altitude and are due to the sun's ultraviolet light, which strips electrons from atoms. The ionosphere varies depending on season of year, time of day, and location on Earth. Density is typically higher around noon and also when the Earth is closest to the sun in its orbit. To further complicate matters, the ionosphere has multiple strata, each of which affect signals differently and each of which changes through the day. If a satellite is close to the horizon, the signal must pass through more ionosphere, further delaying it. The problem of accounting for this large, randomly varying delay source is solved by taking advantage of the fact that the delay of the radio waves depends on the carrier frequency: the higher the frequency, the lower the delay. Two carrier signals received from the same satellite will have travelled through the same part of the ionosphere and will be delayed by different amounts.

Measuring this delay difference allows calculation of the absolute delay. Hence, multicarrier reception allows accurate measurement of ionospheric delay, resulting in a reduction of the largest source of ranging error. The ground control network periodically measures the ionospheric delay at various points on Earth using the two-carrier technique, uploading the correction factors to the satellites for broadcast to receivers in the NAV message. If a single-carrier receiver is used, it has to rely on an ionospheric model for correction of ionospheric delay to frequency. However, this method is not very accurate as the model is often updated only once per day and the nearest ground station may be thousands of kilometers away.

Multipath error is another error type to address. Ideally, the radio signal travels directly along the line-of-sight path from satellite to receiver. But, particularly for satellites close to the horizon, there's a good chance that the signal will reflect off various obstacles, such as buildings and trees. Multipath error can be mitigated via techniques such as antenna design and mask angle used. Antennas use ground planes to prevent signals reflecting off the ground from being received. Choke-ring antennas can attenuate surface waves that occur when a signal that bounces off the ground hits the edge of the ground plane. A mask angle ignores satellites with an elevation angle lower than some angle (typically 15°) above the horizon, since such satellites are more likely to suffer from multipath error.

GNSS satellites are subject to relativistic effects due to their high speed and altitude[1]. Left uncorrected, these effects would render the system unusable. Until the use of GPS receivers and smartphones became widespread, relativistic effects were not something that normally affected our daily life. Let's consider Einstein's Theory of Special Relativity, which predicts that time passes more slowly as a moving reference frame approaches the speed of light. Applying this theory to satellites, the atomic clock on a satellite will "tick" more slowly than a stationary clock on the Earth's surface, given that the satellite is traveling at a high velocity relative to the Earth-bound clock. The Lorentz transformation can be used to calculate time dilation:

$$\frac{1}{y} = \sqrt{1 - \frac{v^2}{c^2}}$$

Where:

v = velocity of satellite = 4km/s

c = speed of light = 2.998×10^8 m/s

1/y = relative time dilation = 10^{-10}

So, satellite clocks run slower than clocks on Earth by a factor of about 1 in 1010. Over one day, this difference accumulates to about 7µs. Now, let's consider Einstein's Theory of General Relativity, which predicts that time is dilated by gravity and clocks will tick more slowly in higher gravity. Applying this theory to satellites, we can surmise that because the satellites are at high altitude, they are subject to less gravity than clocks on the Earth's surface, so they will run faster. The gravitational time dilation equation for calculating this effect is as follows:

$$\frac{1}{y} = \sqrt{1 - \frac{2GM}{rc^2}}$$

Where:

G = universal gravitational constant = 6.674×10^{-11} Nm^2/kg^2

M = mass of Earth = 5.974×10^{24}kg

c = speed of light = 2.998×10^8m/s

r = distance from center of mass of Earth

1/y = relative time dilation

If we make an approximation to simplify the equation and calculate the difference in 1/y factors between a clock on Earth's surface and one in a GPS orbit, we will get the following equation:

$$\Delta = \frac{GM}{c^2}\left(\frac{1}{R_{Earth}} - \frac{1}{R_{gps}}\right)$$

Where:

R_{Earth} = radius of Earth = 6,357,000m

R_{gps} = orbital radius of GPS satellite = $20{,}184 \times 10^3 + R_{Earth}$ = 26,541,000m

Δ = difference in time dilation for satellite and Earth-bound observer = 5.3×10^{-10}

This accumulates to 45.85µs per day. The satellite's speed causes the satellite clock to lose 7µs per day, but gravity causes the clock to gain 46µs per day, so the net result is a gain of 38.6µs. The GPS satellites all use a fundamental frequency of 10.23MHz from which other clocks are derived. For example, the GPS L1 C/A code rate is one-tenth of this frequency of 1.023MHz. To compensate for the time dilation, the fundamental frequency used on the satellites is tuned to 10.229999995453MHz instead of 10.23MHz. The satellite clock will thus tick at 10.23MHz, on the nose, from the perspective of an Earth-based observer.

There used to exist a source of error in the pseudorange that was intentional. Originally, GPS satellites would intentionally add time-varying error to the transmitted coarse acquisition (C/A) signals in order to prevent enemies from using civilian GPS receivers for precision weapons guidance. This intentional degradation was called Selective Availability (SA). SA errors were typically 50m horizontally and 100m vertically. However, since the SA error affects all receivers in a region equally, if we use a GPS receiver with a known position, it can estimate the SA error and transmit it to other receivers. This technique is called differential GPS, and its use has made SA ineffective, so SA was turned off in 2000.

Techniques for Higher Positioning Precision

Differential GPS (DGPS) provides one of several techniques for higher positioning precision, which is essential for applications such as surveying, precision agriculture, and self-driving vehicles. A typical civilian GPS receiver can achieve 2m to 5m position accuracy horizontally, but applications such as those just mentioned require sub-meter accuracy to be effective. With DGPS, a stationary reference receiver at a fixed known location calculates the pseudorange from each visible satellite and also calculates the timing error. Timing error correction information is transmitted over some channel to nearby "roving" receivers, which apply the correction terms to their pseudorange calculations. If the receivers are within a few hundred kilometers of the reference receiver, the signals received by both the reference and roving receivers will have travelled through the same atmosphere and, therefore, be subject to the same delays. Any error sources in common will be corrected (with the exception of multipath errors, since these are local to the receiver).

Carrier-based GPS is another path to higher precision. This method involves using the phase of the radio wave carriers, instead of the PRN codes, to estimate distance. Using the 1% rule of thumb for resolution, the carrier phase can be measured down to a resolution of 1%. For the GPS L1 C/A signal, the carrier frequency is 1575.42MHz, giving a wavelength of 19cm and a measurement resolution of 1.9mm. The accuracy of this method is so high that the continental drift of tectonic plates, of a few inches per year, must be accounted for in the position of the reference receiver.

Choosing the Optimal GNSS Receiver for Your Design

Processing signals broadcast by satellites, GNSS receivers determine user position, velocity, and precise time (PVT). A typical GNSS receiver, as shown in Figure 2, is composed of:

- An antenna
- An optional external low-noise amplifier (LNA) for low-noise amplification close to the antenna
- An optional SAW filter to reject jammers
- Temperature-controlled crystal oscillator (TCXO)
- RF front-end IC, which amplifies, down-converts, filters, and samples the GNSS signal
- Baseband digital signal processor (DSP), which is usually implemented in an FPGA for real-time receivers. The DSP outputs NAV bits and information such as carrier phase and code phase.
- Baseband processor subsystem, which performs all of the mathematical calculations to compute the navigation solution, interpret NAV messages, and apply corrections

What we've discussed so far in this article is the U.S. GNSS. Constellations developed by other countries and regions include GLONASS in Russia, Galileo in the European Union, BeiDou in China, and IRNSS in India. Each has its own signal structures and uses different bands, though some have overlap. Table 1 highlights various satellite navigation systems.

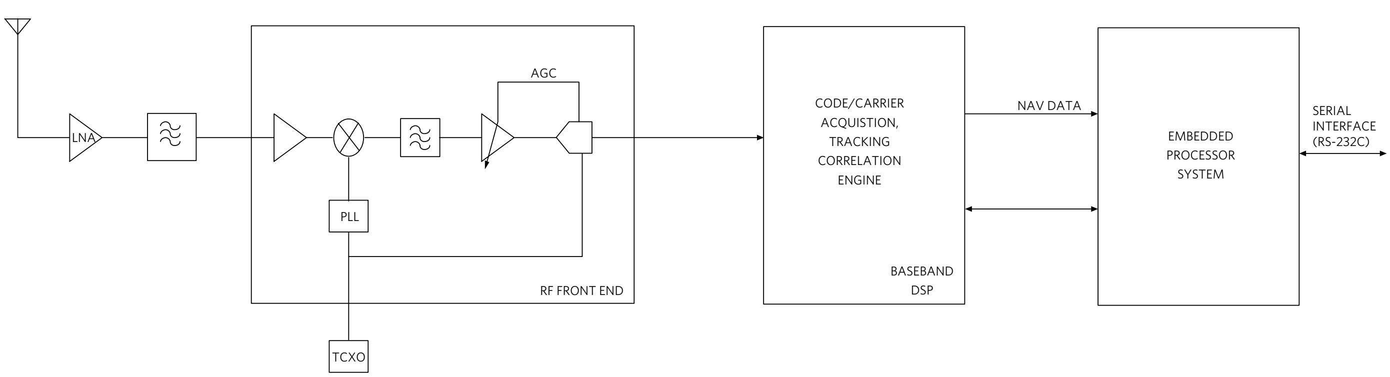

Figure 2. Typical GNSS receiver architecture.

Name of System	Country	Medium Access	Accuracy (m)	Global/ Regional
GPS	US	CDMA	5	Global
GLONASS	Russia	FDMA	5 to 7	Global
BeiDou	China	CDMA	10	Regional
Galileo	EU	CDMA	1	Global
IRNSS	India	CDMA	10	Regional

Table 1. Satellite navigation systems around the world.

When evaluating GNSS receivers, be sure to first consider your target application. That will influence the types of features and levels of performance, accuracy, and power consumption you will need. If your application requires high-precision positioning, for example, then a receiver that supports multiple frequencies and multiple constellations is needed. Maxim Integrated has a history of delivering proven GNSS technologies used by some of the world's Tier 1 GNSS companies. Maxim's products include both GNSS RF front-end ICs and also GNSS LNAs. For example, Maxim offers its MAX2769 fully programmable universal GNSS receiver, which covers GPS, GLONASS, and Galileo systems on a single chip. A single-conversion, low-IF GNSS receiver, the MAX2769 eliminates the need for external IF filters via its on-chip monolithic filters and provides total cascaded noise figure as low as 1.4dB. An automotive-qualified version, the MAX2769B, is also available. The MAX2769C universal GNSS receiver covers the L1/E1, B1, and G1 bands for GPS, Galileo, BeiDou, and GLONASS satellite systems on single chip. Like the MAX2769, the MAX2769C (shown in Figure 3) also provides a total cascaded noise figure as low as 1.4dB and doesn't require external IF filters.

To implement a full solution, the MAX2769C can be connected to a microcontroller, such as the MAX32631, which is running GNSS baseband software in order to implement a software-based receiver. Alternatively, the ADC samples could be input to an FPGA and the lower-level baseband processing done within the FPGA for a higher performance, hardware-based solution.

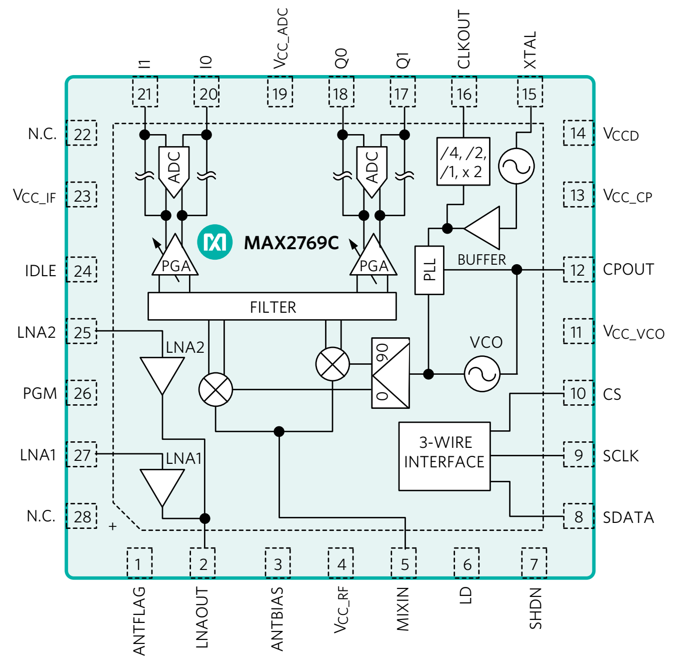

Figure 3. MAX2769C block diagram.

Summary

More electronic devices now offer location-based services, helping us navigate the world and also bringing us useful insights based on where we are. Inside these devices are sophisticated GNSS technologies such as receiver ICs that provide the pinpoint precision we need. This article provided some historical background on GNSS technology, and discussed errors to correct, techniques for more precise positioning, and receiver ICs that support multiple bands and satellite systems.

Sources

1. Error analysis for the Global Positioning System

Maxim Integrated
160 Rio Robles
San Jose, CA 95134 USA
408-601-1000
maximintegrated.com/techdocs