GPS Satellite Orbits

|  |
| --- |
|  |
| Orbits of the GPS-Satellites (distances are to scale)  |

The satellites orbit the earth with a speed of 3.9 km per second and have a circulation time of 12 h sidereal time, corresponding to 11 h 58 min earth time. This means that the same satellite reaches a certain position about 4 minutes earlier each day. The mean distance from the middle of the earth is 26560 km. With a mean earth radius of 6360 km, the height of the orbits is then about 20200 km. Orbits in this height are referred to as MEO – medium earth orbit. In comparison, geostationary satellites like ASTRA or Meteosat – satellites orbit the earth at 42300 km, which is about twice the distance of GPS satellites.

The satellites are arranged on 6 planes, each of them containing at least 4 slots where satellites can be arranged equidistantly. Today, typically more than 24 satellites orbit the earth, improving the availability of the system. The inclination angle of the planes towards the equator is 55°, the planes are rotated in the equatorial plane by 60° against each other. This means that the orbits range from 55° north to 55° degrees south. Block I satellites had an inclination of 63° against the equator.

|  |
| --- |
|  |
| Inclination of the orbital planes  |

By this arrangement of the orbits it is avoided that too many satellites are to often over the north and south pole (like it was the case in the TRANSIT system, where the satellites ran on polar orbits).

However the orbits run far enough to the north and south to guarantee GPS availability in polar regions. Furthermore this arrangement leads to a rather stable constellation, as orbit disturbing factors like solar winds and gravitation fields have about the same influence on all of the satellites.
The number and constellation of satellites guarantees that the signals of at least four satellites can be received at any time all over the world.

The closer you get to the poles, the lower over the horizon the satellites are located. They can still be received very well, but in no case they are directly above. This may lead to a - typically insignificant - loss of the precision of the position determination. This effect, caused by the geometry of the satellite arrangement, happens from time to time on any spot of the earth surface and can be forecasted.

|  |
| --- |
|  |
| Ground-Track (sub satellite path) of the Satellite GPS BIIR-07 (PRN 18) of 18.10.2001 00:00 o'clock to 19.10.2001 00:00 o'clock.  |

The above picture shows the ground track of satellite BIIR-07 (PRN 18) from 2001-10-18, 00:00 ’o clock to 2001-10-19, 00:00 ’o clock. The yellow arrow marks the 00:00 ’o clock time point. It can be seen that the orbit time is slightly shifted (about 4 minutes) in 24 h.

The yellow dot marks the satellite position at 09:30 pm. The satellite is positioned over Ethiopia. The correlating “zone of sight”, within which the satellite signal can be received, is mark in light blue. The graph was compiled with the FreeWare software WinOrbit (and slightly modified).
For a deeper insight into orbits of GPS satellites and other satellites, the following link can be recommended: J-Track 3D. A Java applet on this page can illustrate orbits and information of more than 500 satellites.

History of NAVSTAR GPS

The GPS System was created and realized by the American Department of Defense (DOD) and was originally based on and run with 24 satellites (21 satellites being required and 3 satellites as replacement). Nowadays, about 30 active satellites orbit the earth in a distance of 20200 km. GPS satellites transmit signals which enable the exact location of a GPS receiver, if it is positioned on the surface of the earth, in the earth atmosphere or in a low orbit. GPS is being used in aviation, nautical navigation and for the orientation ashore. Further it is used in land surveying and other applications where the determination of the exact position is required. The GPS signal can be used without a fee by any person in possession of a GPS receiver. The only prerequisite is an unobstructed view of the satellites (or rather of the sky).
The correct name of the system is NAVSTAR (Navigation System for Timing and Ranging), but commonly it is referred to as GPS (Global Positioning System).

|  |  |
| --- | --- |
| 1973 | Decision to develop a satellite navigation system based on the systems TRANSIT, TIMATION und 621B of the U.S. Air Force and the U.S. Navy. |
| 1974 - 1979 | System tests |
| 1977  | First receiver tests are performed even before the first satellites are stationed in the orbit. Transmitters are installed on the earth’s surface called Pseudolites (Pseudo satellites) |
| 1978 - 1985 | A total of 11 Block I satellites are launched in this period. |
| 1979 | Decision to expand the GPS system. Thereupon the resources are considerably shortened and the program is restructured. At first only 18 satellites should be operated. 1988 the number of satellites is again raised to 24, as the functionality is not satisfying with only 18 satellites. |
| 1980 | Launching of the first Block I satellite carrying sensors to detect atomic explosions. This satellite is meant to control the abidance of the agreement of 1963 between the USA and the Soviet Union to refrain from any nuclear tests on the earth, submarine or in space. |
| 1980- 1982 | The financial situation of the project is critical, as the usefulness of the system is questioned again and again by the sponsors. |
| 1983 | When a civilian airplane of the Korean Airline (Flight 007) was shot down after it had gone lost over Soviet territory, it was decided to allow the civilian use of the GPS system. |
| 1986  | The accident of the space shuttle "Challenger" means a drawback for the GPS program, as the space shuttles were supposed to transport Block II GPS satellites to their orbit. Finally the operators of the program revert to the Delta rockets intended for the transportation in the first place.  |
| 1989 | The first Block II satellite was installed and activated. |
| 1990 - 1991 | Temporal deactivation of the selective availability (SA) during the Gulf war. In this period civil receivers should be used as not enough military receivers were available. On July 01, 1991 SA is activated again. |
| 08.12.1993 | The Initial Operational Capability (IOC) is announced. In the same year it is also definitely decided to authorize the world wide civilian use free of charge. |
| March 1994 | The last Block II satellite completes the satellite constellation. |
| 17.07.1995 | Full Operational Capability (FOC) is announced. |
| 01.05.2000 | Final deactivation of the selective availability and therefore improvement of the accuracy for civilian users from about 100 m to 20 m.  |
| 20.03.2004 | Launching of the 50st GPS satellite. |
| 25.09.2005 | Launch of the first IIR-M GPS-satellite. This new type supports the new military M-signal and the second civil signal L2C. |

GPS Satellites

The GPS system can be divided into three basic segments. The space segment will be discussed below, the control segment will be explained on a separate page:

* Space segment (satellites)
* Control segment (control stations)
* User segment (GPS receiver)

Space Segment

|  |
| --- |
|  |
| GPS-Block IIF Satellite(Credits: NASA)  |

The space segment consists of at least 24 satellites. The first of the satellites was brought to its orbit as early as 1978. During the years the satellites became more and more sophisticated and meanwhile five different types of these satellites exist (Block I, Block II, Block IIA, Block IIR und Block IIF).

Block I Satellites

|  |
| --- |
|  |
| GPS-Block I Satellite(Credits: NASA)  |

From 1978 to 1985 11 Block I satellites were launched from California, each having a weight of 845 kg. None of those still operates today. Their lifespan was supposed to be 4.5 years and all of them exceeded this lifespan about another 5 years. The oldest of the satellites, in the beginning designed as prototype for the testing of the system, has been operating for 13 years.

All signals of the Block I satellites were accessible for civil users. Solar panels served as power supply with a power of 400 W. During the satellite’s way through the earth shadow, nickel-cadmium batteries served as reserve. Steering thrusters are operated with Hydrazine.
Further information about the Block I satellites can be obtained here, however with the satellites being out of operation the most recent available information is from 1996.

Block II Satellites

|  |
| --- |
|  |
| GPS-Block IIA Satellite(Credits: NASA)  |

Block II satellites weigh more than 1500 kg, which is about twice the weight of Block I satellites. The first of these satellites was launched in 1998 from Cape Canaveral. They have a wingspan of approximately 5.1 m and are constructed for a service life of 7.5 years. A total of 9 Block II and 18 Block IIA satellites were launched till September 1996. Although the satellites are still in six different orbits, each with the same angle to the equator, the newer Block IIA satellites have a slightly different constellation in space. In 1990 the first Block IIA satellite (A for "advanced") was launched. Further information about the Block II satellites can be obtained here. The status of the total system can be found here.

In September 2005, the first satellite of a new generation (IIR-M, replacement, modernized) was successfully launched. The satellites of this type has the capacity to implement a second civil signal (L2C) and a new military signal with a new code (M-code on L1M and L2M). The satellite weighs 2 tons and costs $ 75 million.

|  |
| --- |
|  |
| Small atomic clock(Photographed at the Verkehrshaus Luzern)  |

Block II and Block IIA satellites are equipped with two rubidium and two cesium atomic clocks with a clock stability of at least 10 - 13s. From the base frequency of the atomic clocks (10.23 MHz) all other frequencies that are required for the GPS-satellite are derived. The newer satellites of Block IIR and IIR-M are equipped with three rubidium atomic clocks. Their extreme precision of ± 1 second in 1 million years is absolutely necessary for the functioning of the system. For an explanation please refer to the chapter “determination of a position”.

|  |
| --- |
|  |
| GPS-Block IIR and IIR-M Satellite(Credits: NASA)  |

Starting with the new Block IIR satellites only the so-called C/A-signal (Coarse/Acquisition) is accessible for civil use. The power supply and the propulsion system are the same as for the Block I satellites. However the solar panels have a higher capacity of 750 W. Initially, the satellites of the Block IIR generation should be brought to their orbit in groups of three by space shuttles. But after the challenger catastrophe in 1986 it was decided to take the satellites to the orbit in pairs with Delta rockets.

|  |
| --- |
|  |
| Start of a Delta rocket(Cedits: NASA)  |

Paradoxically, the first two satellites launched with a Delta rocket were lost when the rocket had to be destroyed due to a malfunction shortly after the lift off. This was the first malfunction ever of a rocket of this type.

Block II satellites have a couple of further features which are not related to the GPS system. For example they are equipped with sensors capable of detecting atomic explosions. The last of the Block IIa satellite launched so far was brought to space on January 30, 2001 from Cape Canaveral. The launch of one Block II satellite costs about $ 50 Mio. which points up the high investment costs the system implicates. The enormous budget is only granted by the US congress because the system can be used by the military as well as for civil tasks. Actual information about the status of Block II satellites can be obtained here or here.

The next generation (Block IIF) is planned to provide a third frequency for civil use (L5), allowing position determinations with even higher precision. This Block IIF satellites may be equipped with hydrogen maser clocks instead of atomic clocks due to their even higher precision.

The radiated signal power of the satellites is only about 50 W. For comparison: television satellites like ASTRA satellites irradiate with a power of 100 W, but focused on Europe, and still a dish antenna with a diameter of at least 50 cm is required for a good reception, whereas GPS antenna typically are only an inch across. However television satellites have to manage a considerably higher data transfer rate than GPS satellites.

Due to their high frequency, GPS signals cannot permeate stone or water. Even a thick foliation in forests may attenuate the signal to an extent that some (mainly older) GPS receivers have difficulties in receiving the signal. However, GPS works in any weather including a thick cloud cover. Problems may only arise in very heavy snowfall.

Control Segment (Monitor Stations)

The GPS-System is controlled by the US Army. The “master control station” (Schriever AFB) and four additional monitoring stations (on Hawaii, Ascension Islands, Diego Garcia and Kwajalein) were set up for monitoring the satellites.
During August and September 2005, six more monitor stations of the NGA (National Geospatial-Intelligence Agency) were added to the grid. Now, every satellite can be seen from at least two monitor stations. This allows to calculate more precise orbits and ephemeris data. For the end user, a better position precision can be expected from this. In the near future, five more NGA stations will be added so that every satellite can be seen by at least three monitor stations. This improves integrity monitoring of the satellites and thus the whole system.

|  |
| --- |
|  |
| Position of the monitor stations and the master control station(Earthmap:NASA; http://visibleearth.nasa.gov/)  |
|  |
| Satellite-tracking-station on Hawaii(Source: Schriever Air Force Base Satellite Flyer Vol. 6; No.12)  |

The passive monitor stations are GPS receivers which track all satellites in their range and collect data of the satellite signals. The raw data are then sent to the mater control station where the data are processed. The stations on Ascension Islands, Diego Garcia and Kwajalein are also transmitting stations for correction data. The "master control station" is located on the Schriever Air Force Base (formerly Falcon AFB), about 20 km south of Colorado Springs.

|  |
| --- |
|  |
| left:  Schriever AFB, Colorado; right: 50th Space Wing’s 2nd Space Operations Squadron(Credits: AFSPC Image Gallery)  |

The "50th Space Wing’s 2nd Space Operations Squadron" is responsible for the operation of the GPS system. Here the data from the monitor stations are processed 24 h a day in real time. As results, information about orbits and clocks of the satellites are obtained. Doing this, possible malfunctions can quickly be detected.
Additionally, from the raw data new ephemeris’s data are calculated. Once to twice a day, theses data and other commands are sent back to the satellites via the transmitting antennae on Ascension Islands, Diego Garcia or Kwajalein by means of a S-band signal (S-band: 2000 - 4000 MHz).
Block IIR satellites are capable of exchanging data with other satellites and can correct their orbit data on their own. In theory they only need a contact to a ground station every 180 days.

Latest information about the status of the GPS system can be obtained here and here.

User Segment (GPS-Receiver)

This will be a short chapter only giving rudimentary information since the receiver market is highly dynamic market with various models changing every few months.

|  |
| --- |
|  |
| Older GPS-Instrument; Magellan GPS 2000(Credits: Magellan)  |

Modern GPS satellite receivers can be built in such a compact way, that they even can be integrated in wrist watches. Most of the commercially available instruments today have about the size of a cell phone. All receivers today have at least 12 channels, meaning that they can receive and process the signals of at least 12 satellites in parallel. Earlier instruments had to process data in series, making them considerably slower and less accurate and more sensitive against disturbances. Today’s receivers can follow up to 66 GPS signals (more than are currently available) and deal with multipath and reflected signals as well as are sufficiently sensitive to work indoors. Instruments for the professional use (military and land survey) typically are bigger and considerably more precise.

Position Determination with GPS

In a considerably simplified approach, each satellite is sending out signals with the following content: I am satellite X, my position is Y and this information was sent at time Z. In addition to its own position, each satellite sends data about the position of other satellites. These orbit data (ephemeris und almanac data) are stored by the GPS receiver for later calculations.

For the determination of its position on earth, the GPS receiver compares the time when the signal was sent by the satellite with the time the signal was received. From this time difference the distance between receiver and satellite can be calculated.

If data from other satellites are taken into account, the present position can be calculated by trilateration (meaning the determination of a distance from three points). This means that at least three satellites are required to determine the position of the GPS receiver on the earth surface. The calculation of a position from 3 satellite signals is called 2D-position fix (two-dimensional position determination). It is only two dimensional because the receiver has to assume that it is located on the earth surface (on a plane two-dimensional surface). By means of four or more satellites, an absolute position in a three dimensional space can be determined. A 3D-position fix also gives the height above the earth surface as a result.

Simplified, the position determination by means of a GPS works on the sample principle as the distance of thunderstorms can be judged: the time is measured between lightning and the following thunder. The speed of light is so high that the delay between the time where the flash hits the ground and the time the observer sees the flash can be neglected. The speed of sound in the earth’s atmosphere is approximately 340 m/s. This means that for example a difference of 3 seconds between lightning and thunder corresponds to approximately 1 km distance to the thunderstorm.

However, this procedure is not yet a determination of a position, but only a determination of a distance. If different people on fixed positions would determine the time span between lightning and thunder, this would allow the determination of the position where the flash hit the ground!

In the following an explanation is given, how the position determination by GPS works. For simplification, in the first step we assume that the earth is a two-dimensional disk. This allows us to do some understandable sketches for illustration. The principle can then be transferred to the model of a three-dimensional globe.

|  |
| --- |
|  |
| Position determination with two satellites (in a 2-dimensional world)  |

In the example on the left, the time needed by a signal to travel from the first of two satellites to the receiver was determined to be 4 s. (In reality this value is far too high. As the signals travel with the speed of light (299 792 458,0 m/s), the actual time span for signals from the satellite to the receiver lies in the range of 0.07 s.) Based on this information, we can at state that the receiver is positioned somewhere on a circle with a radius of 4 s around the first satellite (left circle).

If we perform the same procedure with a second satellite (right circle), we get two points of intersection. On one of the two points the receiver must be situated. Now we have used two satellites. But the process is called trilateration, not dilateration so don't we need a third satellite? We may use a third satellite but we could also assume that the receiver is located somewhere close to the earth's surface and not deep in space, so we can neglect point B and know that the receiver must be found on point A. The area in the picture above which shaded grey is the region in which GPS signals are supposed to be “realistic”. Positions outside this area are discarded, so is point B. This assumption replaces the third satellite which would in theory be required for the process of trilateration. In this example an unequivocal position is obtained from only two satellites. So we just need a third satellite for a third dimension and that's it? Well, in principle yes. But…

The problem lies in the determination of the exact runtime of signals. As explained above, satellites impose a sort of time stamp on each transmitted data package. We know that all clocks of satellites are absolutely precise (they are atomic clocks after all) but the problem is the clock in our GPS receiver. Atomic clocks being too expensive, our GPS receivers are based on conventional quartz clocks which are comparatively inaccurate. What does this mean in practice?

|  |
| --- |
|  |
| 2D position determination with 2 satellites and clock error  |

Let's stick to our example and suppose the clock in our receiver is 0.5 seconds early compared to the clock in the satellite. The runtime of the signal seems to be 0.5 s longer than it actually is. This leads to the assumption that we are on point B instead of point A. The circles that intersect in point B are called pseudo-ranges. They are called “pseudo” as long as no correction of the synchronization errors (bias) of the clocks has been performed.

Depending on the accuracy of the clock in the GPS receiver, the determined position will be more or less wrong. For the practice of GPS based navigation this would mean that no determined position can ever be of any use, as the runtimes of the signals are so short, that any clock error has an overwhelming influence on the result. A clock error of 1/100 second, which is difficult to imagine but quite common from car races or skiing races, would in GPS navigation lead to a mistake in the position of about 3000 km. To achieve an accuracy of 10 m of the position, the runtime of the signal must be precise to 0.00000003 seconds.

As atomic clocks are no option in GPS receivers, the problem is solved in another and quite elegant way:

|  |
| --- |
|  |
| 2D position determination with 3 satellites and corrected clock error  |

If a third satellite is taken into account for the calculation of the position, another intersection point is obtained: in case that all clocks are absolutely precise, point A would be obtained, corresponding to the actual position of the receiver.

In case of the receiver clock being 0.5 s early, the three intersection points B are obtained. In this case the clock error stands out immediately. If now the time of the receiver clock is shifted until the three intersection points B merge to A, the clock error is corrected and the receiver clock is synchronized with the atomic clocks in the satellites.

The GPS receiver can now be regarded as an atomic clock itself. The distances to the satellites, formerly regarded as pseudo-ranges, now correspond to the actual distances and the determined position is accurate.

In case of the example – a two dimensional disc world – we therefore need three satellites for an unequivocal determination of our position. In the real world which has one additional dimension, we would need a fourth satellite.

Well, then why is it always said that three satellites are enough?
In practice you get a two-dimensional position determination (2D-fix) with three satellites. The position is bound to be located on the earth's surface. The fourth satellite is the geo-center; the distance to the “fourth satellite” corresponds to 6360 km (the radius of the globe). Therewith the fourth satellite necessary for the calculation is given, but the calculation is restricted to locations on the earth surface. However the earth is not a perfect sphere. The surface of the earth in this case means the earth geoid, corresponding to sea level. If the receiver is located on a mountain, the determined position again is afflicted with an inaccuracy, as the runtime of the satellite signals is wrong.

By constantly recalculating its position, the GPS receiver can additionally determine the speed and direction of a movement (referred to as "ground speed" and "ground track").

Another possibility of determining the speed is by using the Doppler's effect which occurs due to the movement of the receiver while receiving the signals. The principle is the same as for a moving siren on a police car: the tune is higher when the car moves towards the listener and it is lower when the car moves away.

Transmitted GPS Signals

The principle of position determination by GPS and the accuracy of the positions strongly depend on the nature of the signals. A variety of criteria was considered in the development of a suitable signal structure. In consequence the GPS signal is quite complex and offers the possibility of determining the following parameters: one-way (passive) position determination, exact distance and direction determination (Doppler effect), transmission of navigation information, simultaneous receiving of several satellite signals, provision of corrections for ionospheric delay of signals and insusceptibility against interferences and multi path effects. In order to fulfil all these requirements, the signal structure described below was developed.

 Choice of the carrier frequency

To transport data signals, a suitable carrier frequency is required. The choice of the carrier frequency is submitted to the following requirements:

* Frequencies should be chosen below 2 GHz, as frequencies above 2 GHz would require beam antennae for the signal reception
* Ionospheric delays are enormous for frequency rages below 100 MHz and above 10 GHz
* The speed of propagation of electromagnetic waves in media like air deviates from the speed of light (in vacuum) the more, the lower the frequency is. For low frequencies the runtime is falsified.
* he PRN-codes (explained below) require a high bandwidth for the code modulation on the carrier frequency. Therefore a range of high frequencies with the possibility of a high bandwidth has to be chosen.
* The chosen frequency should be in a range where the signal propagation is not influenced by weather phenomena like, rain, snow or clouds.

Based on these considerations, the choice of two frequencies proved to be advantageous.
Each GPS satellite transmits two carrier signals in the microwave range, designated as L1 and L2 (frequencies located in the L-Band between 1000 and 2000 MHz).
Civil GPS receivers use the L1 frequency with 1575.42 MHz (wavelength 19.05 cm). The L1 frequency carries the navigation data as well as the SPS code (standard positioning code). The L2 frequency (1227.60 MHz, wavelength 24.45 cm) only carries the P code and is only used by receivers which are designed for PPS (precision positioning code). Mostly this can be found in military receivers.

 Modulation of the carrier signals

C/A and P-Code

The carrier phases are modulated by three different binary codes: first there is the C/A code (coarse acquisition). This code is a 1023 “chip” long code, being transmitted with a frequency of 1.023 MHz. A “chip” is the same as a “bit”, and is described by the numbers “one” or “zero”. The name “chip” is used instead of “bit” because no information is carried by the signal. By this code the carrier signals are modulated and the bandwidth of the man frequency band is spread from 2 MHz to 20 MHz (spread spectrum). Thus the interference liability is reduced.

The C/A code is a pseudo random code (PRN) which looks like a random code but is clearly defined for each satellite. It is repeated every 1023 bits or every millisecond. Therefore each second 1023000 chips are generated. Taking into account the speed of light the length of one chip can be calculated to be 300 m.

 Pseudo Random Numbers (PRNs)

The satellites are identified by the receiver by means of PRN-numbers. Real GPS satellites are numbered from 1 – 32. To WAAS/EGNOS satellites and other pseudolites higher numbers are assigned. These PRN-numbers of the satellites appear on the satellite view screens of many GPS receivers. For simplification of the satellite network 32 different PRN-numbers are available, although only 24 satellites were necessary and planned in the beginning. For a couple of years, now more than 24 satellites are active, which optimizes the availability, reliability and accuracy of the network.

The mentioned PRN-codes are only pseudo random. If the codes were actually random, 21023 possibilities would exist. Of these many codes only few are suitable for the auto correlation or cross correlation which is necessary for the measurement of the signal propagation time. The 37 suitable codes are referred to as GOLD-codes (names after a mathematician). For these GOLD-codes the correlation among each other is particularly weak, making an unequivocal identification possible.

The C/A code is the base for all civil GPS receivers. The P code (p = precise) modulates the L1 as well as the L2 carrier frequency and is a very long 10.23 MHz pseudo random code. The code would be 266 days long, but only 7 days are used.
For protection against interfering signals transmitted by an possible enemy, the P-code can be transmitted encrypted. During this anti-spoofing (AS) mode the P-code is encrypted in a Y-code. The encrypted code needs a special AS-module for each receiving channel and is only accessible for authorized personnel in possession of a special key.
The P- and Y-code are the base for the precise (military) position determination. Since January 31, 1994 the AS-system is operating continuously and the P-code is only transmitted as Y-code.

 Transmission of data

In the GPS system data are modulated onto the carrier signal by means of phase modulations. Phase modulation is a rarely used technique, compared to amplitude modulation (AM) or frequency modulation. In the following, these three modulation techniques shall be explained shortly.

Amplitude modulation

|  |
| --- |
|  |
| Amplitude modulation of a data signal onto a carrier signal  |

For the amplitude modulation the amplitude, which corresponds to the strength of the signal, is changed in accordance to the data signal that shall be transported. If this principle would be applied to sound waves, the sound level would change in order to transport a signal. With increasing attenuation it becomes more and more difficult to filter the data from the signal. This kind of modulation is known from AM radio (that's what AM stands for: amplitude modulation).

Frequency modulation

|  |
| --- |
|  |
| Frequency modulation of a data signal onto a carrier  |

For the frequency modulation, the carrier frequency itself is changed by modulating the data signal onto it. If we stay with the example of the sound waves, the pitch of the tones would be changed while the volume would be kept constant. Frequency modulated signals are less susceptible for disturbances and provides a higher bandwidth than AM modulation. This kind of modulation is used for FM radio.

Phase modulation

|  |
| --- |
|  |
| Phase modulation of a data signal onto a carrier  |

When a data signal shall be modulated onto a carrier signal by phase modulation, the sine oscillation of the carrier signal is interrupted and restarted with a phase shift of e.g. 180°. This phase shift can be recognized by a suitable receiver and the data can be restored. Phase modulation leads to an extension of the frequency range of the carrier signal (leading to a spread spectrum) depending on how often the phase is shifted. When the phase changes, wave peaks are followed by wave minimums in a shorter distance than were in the original carrier signal (as can be seen in the graph). This kind of modulation can only be used for the transmission of digital data.

The following graph shows the composition of signals which are transmitted by GPS-satellites. The setup of the NAV/System data is explained in the chapter "data signal composition".

|  |
| --- |
|  |
| Composition of the signals from GPS satellites(according to Peter H. Dana; used with friendly permission)  |

Remark: Modulo 2 Sum means that sums are formed according to arithmetic rules. If the result is larger than 2, only the rest is kept which can not be divided by 2 (0+0=0; 0+1=1; 1+0=1; 1+1=0).

Composition of the Data Signal

In addition to the C/A code navigational information is modulated into the L1 signal. The information consists of a 50 Hz signal and contains data like satellite orbits, clock corrections and other system parameters (information about the status of the satellites). These data are constantly transmitted by each satellite. From these data receiver gets it's date, the approximate time and the position of the satellites.

The complete data signal consists of 37500 bit and at a transmission rate of 50 bit/s a total of 12.5 minutes is necessary to receive the complete signal. This time is required by a GPS receiver until the first determination of a position is possible, if no information about the satellites is stored or the information is outdated. The data signal is divided into 25 frames, each having a length of 1500 bit (meaning an interval of 30 seconds for transmission).

|  |
| --- |
|  |
| Structure of the GPS data of one "frame"  |

The 25 frames are divided into subframes (300 bit, 6 sec.), which are again divided into 10 words each (30 bit, 0.6 sec). The first word of each subframe is the TLM (telemetry word). It contains information about the age of the ephemeris data. The next word is the HOW (hand over word), which contains the number of counted z-epochs. These data contain the time since last “restart” of the GPS time on the previous Sunday 0:00 o’clock. As the P-code is 7 days long, the HOW is used by military receivers to locate their access to the P-code.

The rest of the first subframe contains data about status and accuracy of the transmitting satellite as well as clock correction data. The second and third subframes contain ephemeris parameters. Subframes 4 and 5 contain the so-called almanac data which include information about orbit parameters of all satellites, their technical status and actual configuration, identification number and so on. Subframe 4 contains data for the satellites number 25 – 32, ionospheric correction data, special information and UTC time information; subframe 5 contains almanac data for the satellites 1 – 24 as well as time and the number of the GPS week.

The first three subframes are identical for all 25 frames. Every 30 seconds the most important data for the position determination are transmitted with these three subframes. From the almanac data the GPS receiver identifies the satellites that are likely to be received from the actual position. The receiver limits its search to these previously defined satellites and hence this accelerates the position determination.
As mentioned earlier, the data signal contains correction parameter for the satellite clocks. Why is this necessary, if the atomic clocks are absolutely precise?

Each satellite carries several atomic clocks and has a very accurate time. However the atomic clocks of the individual satellites are not synchronized to the GPS reference time, but run on their own. Therefore correction data for the clocks of each satellite are required. Furthermore, the GPS reference time is different from UTC time (world time) which is synchronized with the rotation of the earth by means of leap seconds.
If a satellite does not transmit its data correctly or its orbit is unstable, it can be marked as in healthy by the control station. This information is transmitted by the satellite in its signal. Receivers then do not take the data from this satellite into account for the position determination. At least if their firmware is properly programmed.

A typical reason why satellites are marked as defective is the necessity of an orbit correction. In this case the thrusters of the satellite are ignited and the defective marking is removed as soon as the satellite has stabilized in its new orbit.

When ephemeris and almanac data are stored in the GPS receiver, it depends on their actuality how long the GPS needs for the first position determination. If the receiver has not had any contact to the satellites for long time, the first position determination will take longer. If the contact has only been interrupted for a short time (e.g. when driving through a tunnel), the position determination is restarted instantly and we speak of reacquisition.

If position and time are known and the almanac and ephemeris data are up-to-date, we speak of a hot start. This is the case when the receiver is turned on at approximately the same position within 2 – 6 hours after the last position determination. In this case a position fix can be obtained within approximately 15 seconds.

If the almanac data are available and the time of the receiver is correct but the ephemeris data are outdated, this is called a warm start. In this case it takes about 45 seconds to actualize the ephemeris data and obtain a position fix. Ephemeris data are outdated when more than 2 – 6 hours have elapsed since the last data reception from the satellites in view. The more new satellites have come into view since the last position determination, the longer the warm start takes.

If neither ephemeris nor almanac data and the last position are known, we talk of a cold start. Then in the first step all almanac data have to be collected from the satellites, this procedure takes up to 12.5 minutes. This happens when the receiver was switched off for several weeks, was stored without batteries or has travelled approximately 300 km or more since the last position fix.

In the last case no almanac data have to be collected, but as the “wrong” satellites are in view, the receiver has to screen all satellites till it finds the ones in view. For a lot of receivers the duration of a cold start can be shortened when the date and approximate position are entered manually.

If you want even more detailed information, please have a look here.

Runtime Measurement of the Signals

As explained earlier, each satellite transmits a pseudo random code (PRN) which is known to the receiver. The receiver can compare the PRN in its memory with the PRN it just received.

The following graph shows two identical codes. The colored rectangles symbolize binary 1, white gaps symbolize 0. The violet rectangles are the signal from the satellite; the orange rectangles are the signal from the receiver. Now it is determined how “far” the signals have to be shifted until they are aligned. The distance corresponds to a time – the runtime of the signal from the satellite to the receiver. By means of this runtime the distance between the satellite and receiver can be calculated.

|  |
| --- |
|  |
| Comparison of two signals. Top: shifted; Bottom: aligned  |

But how is the shifting done in practice? The received signals are very weak, all satellites transmit on one frequency and whatever reaches the receiver is a great mess of data. The solution to this problem is provided by a tricky algorithm named cross correlation which is rather insensitive against disturbances.

In the following the procedure shall be explained for a simple and well-defined signal.

The upper row shows a segment of a PRN-code of a satellite, the middle row shows the same segment of the receiver. The green column symbolizes the start of the code. It can be seen that in the first example the code of the receiver is a little behind time. In the cross correlation the signals are multiplied with each other, resulting in the signal in the lowest row. The signals in the lowest row are summed up, giving a value of 9 (9 x 1 plus 39 x 0).

|  |
| --- |
|  |
| Top row: signal of the satelliteMiddle row: Signal of the receiver, delayed against the signal of the satellite.Bottom row: Both signals multiplied. If the multiplied signals of each position are summed up, the correlation value of 9 is received.  |

Now the signal of the receiver is shifted step by step and after each shift a cross correlation is done. This results in a figure for each shift. In the second example the signals are congruent. The sum at the end of the cross correlation is bigger than before.

|  |
| --- |
|  |
| Top row: signal of the satelliteMiddle row: Signal of the receiver, aligned with the signal of the satellite.Bottom row: Both signals multiplied. If the multiplied signals of each position are summed up, the correlation value of 25 is received.  |

If the signal is shifted further, the cross correlation again leads to smaller correlation values.

|  |
| --- |
|  |
| Top row: Signal of the satelliteMiddle row: Signal of the receiver, ahead of the signal of the satellite.Bottom row: Both signals multiplied. If the multiplied signals of each position are summed up this results in a correlation value of 9.  |
|  |
| Correlation of both signals for shifts from -7 to 13  |

The drawing on the left shows correlation values for shifts from -7 to 13. A clear maximum is obtained for a shift of 3. The function is normalized to 1 what means a 100 % correlation of the signals.

In reality, the process of cross correlation is a little more complex. For example, when moving the GPS receiver, the signal is compressed or stretched by the Doppler effect. On the one hand, this fact allows a determination of the speed, on the other hand this complicates correlation, as the signals do not only have to be shifted, but must also be stretched or compressed.

Now we want to use the principle explained above on the GPS signals. We already learned that the C/A code is composed of 1023 chips, being transmitted with 1.023 MHz and therefore being repeated every 1000 microseconds. At the speed of light, 1000 microseconds correspond to a distance of about 300 km. this means the signal is repeated every 300 km. Each column in the graph above corresponds to one chip of the GPS signal. The calculated shift of 3 therefore means 3 chips or a distance of 0.9 km.

Now what does this distance of 0.9 km mean? And: as the signal shift is only known within an accuracy of 1 microsecond, the distance is known with an accuracy of 300 m. How can GPS then be much more precise?

The answer to the second question first: modern GPS receiver are capable of calculating the signal shift as precise as 1 % of one chip. Therefore the distance to the satellite can be calculated with a precision of 3 m.

Concerning the first question:

All data packets of the C/A-Code are numbered (z-count). As soon as a packet is received and decoded, the receiver can calculate the time this packet was under way. From this, a pseudorange is calculated which means the distance to the satellite but containing the clock error. With at least four of the pseudo ranges it then can calculate the actual distance and thus the position. Regarding pseudo ranges, also see here.

Sources of Errors in GPS

Selective Availability

The most relevant factor for the inaccuracy of the GPS system is no longer an issue. On May 2, 2000 5:05 am (MEZ) the so-called selective availability (SA) was turned off. Selective availability is an artificial falsification of the time in the L1 signal transmitted by the satellite. For civil GPS receivers that leads to a less accurate position determination (fluctuation of about 50 m during a few minutes). Additionally the ephemeris data are transmitted with lower accuracy, meaning that the transmitted satellite positions do not comply with the actual positions. In this way an inaccuracy of the position of 50 – 150 m can be achieved for several hours. While in times of selective availability the position determination with civil receivers had an accuracy of approximately 10 m, nowadays 20 m or even less is usual. Especially the determination of heights has improved considerably from the deactivation of SA (having been more or less useless before).

The reasons for SA were safety concerns. For example terrorists should not be provided with the possibility of locating important buildings with homemade remote control weapons. Paradoxically, during the first gulf war in 1990, SA had to be deactivated partially, as not enough military receivers were available for the American troops. 10000 civil receivers were acquired (Magellan and Trimble instruments), making a very precise orientation possible in a desert with no landmarks.

Meanwhile SA is permanently deactivated due to the broad distribution and world wide use of the GPS system.

The following two graphs show the improvement of position determination after deactivation of SA. The edge length of the diagrams is 200 m, the data were collected on May 1, 2000 and May 3, 2000 over a period of 24 h each. While with SA 95 % of all points are located within a radius of 45 m, without SA 95 % of all points are within a radius of 6.3 m.

|  |
| --- |
|  |
| Plot of the position determination with and without SA (Diagram from http://www.igeb.gov/sa/diagram.shtml (page no longer available) With friendly permission of Dr. Milbert (NOAA))  |

 "Satellite geometry"

Another factor influencing the accuracy of the position determination is the "satellite geometry". Simplified, satellite geometry describes the position of the satellites to each other from the view of the receiver.

If a receiver sees 4 satellites and all are arranged for example in the north-west, this leads to a “bad” geometry. In the worst case, no position determination is possible at all, when all distance determinations point to the same direction. Even if a position is determined, the error of the positions may be up to 100 – 150 m. If, on the other hand, the 4 satellites are well distributed over the whole firmament the determined position will be much more accurate. Let’s assume the satellites are positioned in the north, east, south and west in 90° steps. Distances can then be measured in four different directions, reflecting a „good“ satellite geometry.

The following graph shows this for the two-dimensional case.

|  |
| --- |
|  |
| Good geometrical alignment of two satellites  |

If the two satellites are in an advantageous position, from the view of the receiver they can be seen in an angle of approximately 90° to each other. The signal runtime can not be determined absolutely precise as explained earlier. The possible positions are therefore marked by the grey circles. The point of intersection A of the two circles is a rather small, more or less quadratic field (blue), the determined position will be rather accurate.

|  |
| --- |
|  |
| Bad geometrical alignment of two satellites  |

If the satellites are more or less positioned in one line from the view of the receiver, the plane of intersection of possible positions is considerably larger and elongated- The determination of the position is less accurate.

The satellite geometry is also relevant when the receiver is used in vehicles or close to high buildings. If some of the signals are blocked off, the remaining satellites determine the quality of the position determination and if a position fix is possible at all. This can be observed in buildings close to the windows. If a position determination is possible, mostly it is not very accurate. The larger the obscured part of the sky, the more difficult the position determination gets.

Most GPS receivers do not only indicate the number of received satellites, but also their position on the firmament. This enables the user to judge, if a relevant satellite is obscured by an obstacle and if changing the position for a couple of meters might improve the accuracy. Many instruments provide a statement of the accuracy of the measured values, mostly based on a combination of different factors (which manufacturer do not willingly reveal).

To indicate the quality of the satellite geometry, the DOP values (dilution of precision) are commonly used. Based on which factors are used for the calculation of the DOP values, different variants are distinguished:

* GDOP (Geometric Dilution Of Precision); Overall-accuracy; 3D-coordinates and time
* PDOP (Positional Dilution Of Precision) ; Position accuracy; 3D-coordinates
* HDOP (Horizontal Dilution Of Precision); horizontal accuracy; 2D-coordinates
* VDOP (Vertical Dilution Of Precision); vertical accuracy; height
* TDOP (Time Dilution Of Precision); time accuracy; time

HDOP-values below 4 are good, above 8 bad. HDOP values become worse if the received satellites are high on the firmament. VDOP values on the other hand become worse the closer the satellites are to the horizon and PDOP values are best if one satellite is positions vertically above and three are evenly distributed close to the horizon. For an accurate position determination, the GDOP value should not be smaller than 5. The PDOP, HDOP and VDOP values are part of the NMEA data sentence $GPGSA.

The satellite geometry does not cause inaccuracies in the position determination that can be measured in meters. In fact the DOP values amplify other inaccuracies. High DOP values just amplify other errors more than low DOP values.

The error in the position determination caused by the satellite geometry also depends on the latitude of the receiver. This is shown below in the two diagrams. The diagram on the left side shows the inaccuracy of the height (at the beginning of the curve with SA), recorded in Wuhan (China). Wuhan is situated on 30.5° northern latitude were ideal satellite constellation can be found at all time. The graph on the right side shows the same interval recorded by the Casey-Station in the Antarctica (66.3° southern latitude). Due to the satellite constellation from time to time the error is much larger. Additionally the falsification by the atmospheric effect gets more significant the closer the position is to the poles (for an explanation see “atmospheric effects”).

|  |
| --- |
|  |
| Error in the height determination at different latitudes(Diagrams from http://www.ngs.noaa.gov/FGCS/info/sans\_SA/world/.With friendly permission of Dr. Milbert (NOAA))  |

Satellite Orbits

Although the satellites are positioned in very precise orbits, slight shifts of the orbits are possible due to gravitation forces. Sun and moon have a weak influence on the orbits. The orbit data are controlled and corrected regularly and are sent to the receivers in the package of ephemeris data. Therefore the influence on the correctness of the position determination is rather low, the resulting error being not more than 2 m.

Multipath effect

|  |
| --- |
|  |
| Interference caused by reflection of the signals  |

The multipath effect is caused by reflection of satellite signals (radio waves) on objects. It was the same effect that caused ghost images on television when antennae on the roof were still more common instead of today’s satellite dishes.

For GPS signals this effect mainly appears in the neighborhood of large buildings or other elevations. The reflected signal takes more time to reach the receiver than the direct signal. The resulting error typically lies in the range of a few meters.

Atmospheric effects

|  |
| --- |
|  |
| Influenced propagation of radio waves through the earth's atmosphere  |

Another source of inaccuracy is the reduced speed of propagation in the troposphere and ionosphere. While radio signals travel with the velocity of light in the outer space, their propagation in the ionosphere and troposphere is slower.

In the ionosphere in a height of 80 – 400 km a large number of electrons and positive charged ions are formed by the ionizing force of the sun. The electrons and ions are concentrated in four conductive layers in the ionosphere (D-, E-, F1-, and F2-layer). These layers refract the electromagnetic waves from the satellites, resulting in an elongated runtime of the signals.

These errors are mostly corrected by the receiver by calculations. The typical variations of the velocity while passing the ionosphere for low and high frequencies are well known for standard conditions. Theses variations are taken into account for all calculations of positions. However civil receivers are not capable of correcting unforeseen runtime changes, for example by strong solar winds.

It is known that electromagnetic waves are slowed down inversely proportional to the square of their frequency (1/f2) while passing the ionosphere. This means that electromagnetic waves with lower frequencies are slowed down more than electromagnetic waves with higher frequencies. If the signals of higher and lower frequencies which reach a receiver are analyzed with regard to their differing time of arrival, the ionospheric runtime elongation can be calculated. Military GPS receivers use the signals of both frequencies (L1 and L2) which are influenced in different ways by the ionosphere and are able to eliminate another inaccuracy by calculation.

The tropospheric effect is a further factor elongating the runtime of electromagnetic waves by refraction. The reasons for the refraction are different concentrations of water vapor in the troposphere, caused by different weather conditions. The error caused that way is smaller than the ionospheric error, but can not be eliminated by calculation. It can only be approximated by a general calculation model.

The following two graphs visualize the ionospheric error. The left data were collected with a one-frequency receiver without ionospheric correction, the right data were collected with a two-frequency receiver with ionospheric correction. Both diagrams have approximately the same scale (Left: latitude -15 m to +10 m, longitude -10 m to +20 m, Right: latitude -12 m to +8 m, longitude -10 m to +20 m). The right graph clearly shows less outliers, while the mean accuracy of the position for 95 % of the data is not considerably enhanced by the correction of the ionospheric error.

|  |
| --- |
|  |
| Position determination without and with atmospheric corrections by using the second frequency on a dual-frequency receiver(diagrams from http://www.ngs.noaa.gov/FGCS/info/sans\_SA/iono.With friendly permission of Dr. Milbert (NOAA))  |

With the implementation of WAAS and EGNOS it s possible to set up „maps“ of the atmospheric conditions over different regions. The correction data are sent to the receivers, enhancing the accuracy considerably.

 Clock inaccuracies and rounding errors

Despite the synchronization of the receiver clock with the satellite time during the position determination, the remaining inaccuracy of the time still leads to an error of about 2 m in the position determination. Rounding and calculation errors of the receiver sum up approximately to 1 m.

 Relativistic effects

The following section shall not provide a comprehensive explanation of the theory of relativity. In the normal life we are quite unaware of the omnipresence of the theory of relativity. However it has an influence on many processes, among them is the proper functioning of the GPS system. This influence will be explained shortly in the following.

As we already learned, the time is a relevant factor in GPS navigation and must be accurate to 20 - 30 nanoseconds to ensure the necessary accuracy. Therefore the fast movement of the satellites themselves (nearly 12000 km/h) must be considered.

Whoever already dealt with the theory of relativity knows that time runs slower during very fast movements. For satellites moving with a speed of 3874 m/s, clocks run slower when viewed from earth. This relativistic time dilation leads to an inaccuracy of time of approximately 7,2 microseconds per day (1 microsecond = 10-6 seconds).

The theory of relativity also says that time moves the slower the stronger the field of gravitation is. For an observer on the earth surface the clock on board of a satellite is running faster (as the satellite in 20000 km height is exposed to a much weaker field of gravitation than the observer). And this second effect is six times stronger than the time dilation explained above.

Altogether, the clocks of the satellites seem to run a little faster. The shift of time to the observer on earth would be about 38 milliseconds per day and would make up for an total error of approximately 10 km per day. In order that those error do not have to be corrected constantly, the clocks of the satellites were set to 10.229999995453 Mhz instead of 10.23 Mhz but they are operated as if they had 10.23 MHz. By this trick the relativistic effects are compensated once and for all.

There is another relativistic effect, which is not considered for normal position determinations by GPS. It is called Sagnac-Effect and is caused by the movement of the observer on the earth surface, who also moves with a velocity of up to 500 m/s (at the equator) due to the rotation of the globe. The influence of this effect is very small and complicate to calculate as it depends on the directions of the movement. Therefore it is only considered in special cases.

The errors of the GPS system are summarized in the following table. The individual values are no constant values, but are subject to variances. All numbers are approximative values.

|  |  |
| --- | --- |
| Ionospheric effects  | ± 5 meters |
| Shifts in the satellite orbits  | ± 2.5 meter |
| Clock errors of the satellites' clocks  | ± 2 meter |
| Multipath effect  | ± 1 meter |
| Tropospheric effects | ± 0.5 meter |
| Calculation- und rounding errors  | ± 1 meter |

Altogether this sums up to an error of ± 15 meters. With the SA still activated, the error was in the range of ± 100 Meter. Corrections by systems like WAAS and EGNOS, which mainly reduce ionospheric effects, but also improve orbits and clock errors, the overall error is reduced to approximately ± 3 - 5 meters.

Achievable Accuracy

A standard GPS receiver for civil use offers an accuracy down to a few meters. In praxis the number and geometry of the received satellites influences the accuracy considerably, and in daily use, accuracies of about 20 m can be expected. More sophisticated GPS receivers as they can be found for land survey cost several thousand dollars and achieve an accuracy of a few centimeters.

With selective availability activated, receivers achieve accuracies of approximately 100 m (these declarations are true for 95 % of all cases). After deactivation of the selective availability the accuracy rose to approximately 15 m, depending of the number and position of available satellites.

 See also here for up-to-date Information about local accuracy in southern Germany (page is in German but numbers are numbers).

 Differential GPS (DGPS)

The technique called differential GPS (DGPS) enables civil receivers to achieve accuracies of 5 m or less. Therefore a second stationary GPS receiver is applied for correcting the measurements of the first receiver. If the position of the stationary receiver is known very accurately, by means of a long wave transmitter (283.5 - 325.0 kHz) a correction signal can be sent which is received and analyzed by a receiver connected to the mobile GPS. The correction signal is like the GPS signals themselves free of charge, the only costs of the long wave receiver arise. This receiver is connected to the GPS and transfers the correction data in a serial data format (RTCM SC-104). The transmission partially is restricted to coastal areas, as it is often sustained by the coastguard of a country.

Wide Area Augmentation System (WAAS)

WAAS (Wide Area Augmentation System) has been operating since 1999 in the United States and is available for portable GPS receivers since 2001. WAAS consists of approximately 25 ground stations controlling the GPS signals and two reference stations that collect the data of the ground stations and calculate correction data. These data contain corrections for the satellite orbits, clock drift and signal delay of the satellites caused by the ionosphere and troposphere. The data are sent to the receivers via to geostationary satellites.

WAAS has been working since December 1999 with nearly no interruption. It was developed for the aeronautical authority FAA to enable precise instrument approaches for landing. The WAAS signal is accessible for civil use and offers a better coverage on land and on sea than the land-based DGPS systems. Unlike DGPS, the reception of WAAS requires no additional receivers. Only the software of the GPS receiver must support the reception of WAAS correction signals. However it is important that one of the geostationary satellites is in view of the receiver. This is more problematic, if the receiver is positioned further north, as the altitude of geostationary satellites above the horizon decreases. Therefore WAAS is especially useful for navigation in open land, aviation and navigation on sea.

In Europe a system corresponding to WAAS is operated, called EGNOS (Euro Geostationary Navigation Overlay Service). In Asia, a Japanese system called MSAS (Multi-Functional Satellite Augmentation System) is planned. As all those systems operate with the same principle, a GPS receiver supporting WAAS can also benefit from EGNOS and MSAS. More about the WAAS/EGNOS-System can be read here.

Overview of Typical Accuracies

|  |  |
| --- | --- |
| Accuracy of GPS system with SA activated | ± 100 Meter |
| Typical accuracy with SA deactivated | ± 15 Meter |
| Typical accuracy of differential GPS (DGPS) | ± 3 - 5 Meter |
| Typical accuracy with WAAS/EGNOS | ± 1 - 3 Meter |

Accuracy Values by Garmin Receivers

|  |
| --- |
|  |
| Accuracy of position determination  |

The declaration of the accuracy by Garmin GPS receivers often leads to confusion. What does it mean if the receiver states an accuracy of 4 m? This readout refers to the so-called 50 % CEP (Circular Error Probable). This means that 50 % of all measurements are within a radius of 4 m. On the other hand, 50 % of all measured positions are outside of this radius. Furthermore, 95 % of all measured positions are within a circle of twice this radius and 98.9 % of all positions are within a circle of 2.55 the radius. In the given example, nearly all positions are within circle with a radius of 10 m. The determined position is in the worst case accurate to 10 m.

WAAS and EGNOS

|  |
| --- |
|  |
|  The Logos of WAAS und EGNOS (From: FAA und ESA)  |

Very much simplified, WAAS (Wide Area Augmentation System) is a satellite based differential GPS system (DGPS). The difference is, that no additional long-wave receiver is necessary to receive the correction data and there is no need for an endless number of DGPS beacons that transmit these correction data.

 Differences between WAAS, EGNOS and MSAS

In principle, all three systems are the same and even more astonishing, the three systems are compatible to each other. This can be called astonishing since WAAS is maintained by north America, EGNOS (European Geostationary Navigation Overlay Service) is maintained by the European community and MSAS (Multi-Functional Satellite Augmentation System) is developed by Japan and other Asiatic countries.

While WAAS is operational (IOC = Initial Operational Capability) since January 2003 (although not yet approved by the FAA), the EGNOS system made some huge steps forward during 2002 but is still (May 2005) in test operation (ESTB = EGNOS Satellite Test Bed). The development of the MSAS system had a major drawback in 1999 after the first of two satellites planned for the system was lost during launch because of a malfunction of the rocket. The start for the replacement satellite was initially planned for August 2004 but was delayed until further notice to investigate the reasons for the malfunction of the rocket in 1999. Up-to-date information on the project are sparse (see here)

While alle systems can be called SBAS (Satellite Based Augmentation Systems), this name is seldomly used.

 How the SBAS work

Background

It is no surprise that WAAS, EGNOS and MSAS have not been developed to the increase the accuracy of GPS for hikers and geocachers. The main reason is to increase the safety for aviation. The GPS system is neither accurate nor reliable enough to be accepted as a sole means of navigation. One of the reasons is that there is no reliable and quick (within seconds) information to the user if problems with the system occur. As a consequence, for landing approaches, GPS can’t be used. Airplanes still have to use ILS-systems (Instrument Landing Systems) if visibility is poor. But the installation and maintenance of ILS-systems on every airport is expensive. With the SBAS systems, CAT I approaches (limited visibility) will be possible without additional ILS systems. For CAT III approaches (zero visibility) even the SBAS will not suffice and ILS are still required.

 Infrastructure and Principle of the System

The SBAS shall provide additional accuracy and reliability for the GPS system. To achieve this, a number of GPS receiving stations are necessary. In the US , 25 station are used, Europe uses 10 stations during the test operation and will have 34 when EGNOS is fully operational. The position of these RIMS (Ranging and Integrity Monitor Stations) must be known very precise. This means that the position of the receiving antenna needs to be know exactly to a few centimeters. The RIMS station receive the standard GPS signal (and also the signal from the Russian GLONASS system and the GALILEO system in future). That way it is possible to calculate the difference between the known position of the station and the position as calculated by the GPS receiver. And since the RIMS use receivers that use both GPS frequencies (L1 and L2), the signal delay through the ionosphere can be calculated for every single satellite.

Additionally, if the signals from more than four satellites are received, more information than needed for a position determination is available and this information may be used to check for possible problems with the satellites or deviations in their orbits or time.

The data from all RIMS are sent to a Central Processing Centre. For the EGNOS test bed (ESTB) this centre is in Toulouse ( France ) and a backup system is located in Hønefoss ( Norway ). Once EGNOS is fully operational there will be control centers, called MCC (Mission Control Centre) in Germany (Langen near Frankfurt ), Spain (Torrejon near Madrid ), Italy (Ciampino near Rome ) and Great Britain (Swanwick near London ). At these stations, the data will be collected and the following data will be calculated:

* Long term errors of the satellite orbits
* Short term and Long term errors of the satellite clocks
* IONO correction grids
* Integrity information

By use of the integrity information, it is possible to inform the users within 6 seconds on problems that occur with the GPS system.

|  |
| --- |
|  |
| Example of a TEC-Map of the Ionosphere over northern America(Source: JPL)  |

The most important feature of the SBAS for common GPS users is the IONO correction grid. Since SA (selective availability) is deactivated, the largest single source of error in GPS position determination is the signal delay in the ionosphere. Being able to correct these errors significantly increases the accuracy of every GPS receiver that is able to process WAAS/EGNOS data.

From the measured data of the RIMS, a ‘map’ of the Total Electron Content (TEC) in the ionosphere for the area covered by the RIMS station is calculated. With decreased accuracy the area where the TEC map is calculated can even be expanded further.

This TEC map is now transmitted to a geostationary satellite that itself acts like a GPS satellite, that means can be used for position determination but also provides the receiver with the information it needs for the correction of the ionospheric effects.

For the EGNOS test system (ESTB) from Aussaguel (near Toulouse in France ) data is transmitted to INMARSAT AOR-E and from Fucino ( Italy ) to INMARSAT IOR. Later when EGNOS is fully operational, data will be sent from Aussaguel and Goonhilly ( Great Britain ) to AOR-E and from Fucine and Goonhilly to IOR-F5. From the stations Torrejon ( Spain ) and Scanzano ( Italy ) data will be transmitted to the satellite Artemis.

The satellite Artemis had an interesting history when in January 2003 it finally arrived at its designated position. After problems with the fourth stage of the Ariane-5 rocket after the launch in July 2001, Artemis had almost to be abandoned when but the engineers managed to “walk” the satellite to its planned position by making extensive use of its newly developed ion propulsion system.

The geostationary satellites do provide a signals very similar to that of the GPS-satellites and on the same frequency. Therefor these satellites may be used for position calculation and additionally, the correction data sent out can be used to improve accuracy for position calculation with all GPS satellites.

|  |
| --- |
|  |
| Two dimensional simplified diagram of a IONO correction grid  |

Using the TEC map transmitted by the geostationary satellites, the GPS receiver can now calculate the ‘pierce point’ and signal delay of the signal of each satellite used for position calculation and then correct the data for higher accuracy in position determination.

The ionosphere is not static but depends on the sun’s activity. For example it is known that single frequency receivers are more accurate shortly after midnight than they are during the day.

The other functions that the SBAS provide like integrity check of the GPS system and transmission of warnings in case of problems with the system will probably be never evaluated by standard handheld GPS receivers since the calculations are complex and the information is of not much interest to the common GPS user.

Differences to DGPS

For land-based people, the main difference between D GPS and the SBAS systems ist he calculation of the TEC map for ionospheric corrections. This brings some of the benefits of an expensive dual frequency receiver to a cheap single frequency receiver.

With D GPS , every single reference station compares its own precisely known position with the position calculated from the GPS signals. The station then transmits this information on a certain long wave band as correction data. A D GPS receiver receives the correction information and applies this correction to the signals received from the GPS satellites. With increasing distance of the receiver to the D GPS reference station, the atmospheric influences on the signals get more and more different and the correction get less and less accurate. If the distance between the reference station is large, the signals from the satellites travel through different parts of the atmosphere, being influenced in different ways. Even worse, due to the large distance the receiver may receive data from completely different satellite where no correctional information are provided in the correctional data. This effect where the reference station does not provide the right data for correction due to large distance to the receiver is called ‘spatial decorrelation’. Because of these phenomena, the typical range for DGPS stations is 70 – 200 km with good accuracy.

For the SBAS systems (WAAS, EGNOS, MSAS) this is different. Here, the monitor stations do not provide single isolated corrections but from all stations together a correction map is calculated for a wide area. Every single receiver then corrects its own position itself by use of this data. That way, the accuracy that can be achieved is even better than with D GPS .

If the receiver is outside of the area where valid TEC map data are provided, the receiver should use the build in standard ionospheric correction and then there should be no difference between WAAS/EGNOS switched on or off in the receiver setup menu in these cases. But since most manufacturers recommend switching WAAS/EGNOS off if not within the area where the signals are provided for, it may be assumed that the receivers do not make full use of all information provided and do not check whether the correction should be used or not. That way, the position may be even worse when WAAS/EGNOS is switched on when no proper correction signal is provided.

Coverage of the geostationary satellites

The area covered by WAAS, EGNOS and MSAS depends on where RIMs station are located and if signals from geostationary satellites are being received. For the transmission of the SBAS signals currently some INMARSAT satellites are used. These satellites are positioned in geostationary orbits (about 36000 km) and are primarily being used as relay satellites for telephone calls from and to ships. The following map shows the ‘footprint’ of these satellites, that means the area where their signals may be received. Before EGNOS is completely operational, there will be some changes, especially concerning the coverage over Europe.

|  |
| --- |
|  |
| INMARSAT-Satellites and their coverage  |

The following table lists the WAAS/EGNOS satellites and their identification numbers:

| **Satellite** | **Satellite location** | **GPS PRN No.** | **Garmin Sat ID** |
| --- | --- | --- | --- |
| INMARSAT 3 F2 (AOR-E)(Atlantic Ocean Region East) | Western Africa | 120 | 33 |
| INMARSAT 3 F4 (AOR-W)(Atlantic Ocean Region West) | East coast of Brazil | 122 | 35 |
| INMARSAT 3 F1 (IOR)(Indian Ocean Region) | Indian Ocean | 131 | 44 |
| INMARSAT 3 F3 (POR)(Pacific Ocean Region) | Pacific | 134 | 47 |
| INMARSAT IOR-W (III-F5)(Indian Ocean Region West) | Africa (Kongo) | 126 | 39 |
| Artemis | Africa (Kongo) | 124 | 37 |
| MTSAT-1R(Multifunction Transportation Satellite) | planned | 129 | 42 |
| MTSAT-2 | planned | 137 | 50 |

If you have the WAAS/EGNOS function activated on your GPS and you receive signals of other satellites than no. 33 or no. 44 while being in Europe you should be careful. Under certain circumstances, signal of satellites being designated for the America (especially no. 35) may be received in Europe . But these satellites only transmit correctional data for northern America so you do not have any advantages.

As already said, the satellite constellation over Europe will change before EGNOS is fully operational. It is planned that the ESA (European Space Agency) satellite ARTEMIS will be used for EGNOS while AOR-E won’t be used in future. The satellite IOR will be moved towards the pacific.

There is one major disadvantage of the correction systems that are based on geostationary satellites. For a GPS near ground and in central Europe or northern America , all geostationary satellites are located in the south and quite low over the horizon. For example, if you are located in Munich , AOR-E is about 35° above the horizon, IOR on at 16°. This easily leads to blocking of the signals by buildings or trees. In forested or hilly areas, WAAS or EGNOS probably will never work perfectly. This disadvantage results in the system being developed for aviation where it does not matter if the satellites are a little low over the horizon.

The INMARSAT satellite III-F5 that will also be used for EGNOS, will be located about 35° over the horizon viewed from Munich. How good this will be for signal reception will be seen in future.

WAAS, EGNOS and Garmin GPS

|  |
| --- |
|  |
| Sky view page of the Garmin etrex Vista with ESTB-Satellites  |

Beginning with April 1st, 2003 , the EGNOS or ESTB signal is transmitted in WAAS compatible data format (SBAS mode 0/2). Since that time, the receivers from Garmin are able to use the data. It should be repeated that Garmin receivers are only capable of doing WAAS/EGNOS corrections if they are set to normal mode. It is not enough to switch to WAAS-yes, you also have to got to normal mode. Being in energy-saving mode they do not evaluate the WAAS/EGNOS data. Unfortunately, normal mode does use significantly more power and thus greatly reduces the life time of your battery. The picture on the right shows the satellite screen of Garmins extrex vista with EGNOS correction applied. The letter “D” in the signal bar indicates that the data from this particulate satellite is being corrected by the EGNOS signal. With an accuracy of 2 meters (RMS) the positioning is quite good.

 Current Status of ESTB/EGNOS

as of June, 5th 2005:

ESTB tests on satellite IOR (PRN 131; ID 44) have been stopped on Mai, 27 th. The satellite will now be moved towards the pacific and there will be a transition from ESTB to EGNOS. At the moment ESTB signals are being transmitted through AOR-E (PRN 120; ID 33).

The final EGNOS constellation will consist of the satellites ARTEMIS (PRN 124; ID 37) , Inmarsat AOR-E (PRN 120; ID 33) and Inmarsat IOR-W (PRN 126; ID 39). Until the official start of EGNOS these satellites do send signals from time to time that will not be evaluated by end-user receivers, although they sometimes show up as grey bars in the sky view with a GPS signal.

The current transmission status of the ESTB-Satellites can he seen here.

The following diagram shows the planned activity for ESTB/EGNOS.

|  |
| --- |
|  |
| Planning for ESTB/EGNOS (Source: [esa](http://esamultimedia.esa.int/docs/egnos/estb/schedule.htm))  |

|  |
| --- |
|  |
| Position and coverage of the planned EGNOS-satellites  |

For more information click [here](http://esamultimedia.esa.int/docs/egnos/estb/schedule.htm).

Links:

[SBAS-page](http://www.faa.gov/asd/international/sbas.htm) of the FAA (Federal Aviation Administration)
[ESA satellite navigation](http://www.esa.int/export/esaSA/navigation.html)
[EGNOS-page of ESA](http://www.esa.int/export/esaEG/estb.html)
[ESA-list of GPS-Receivers capable to receive SBAS-signals (WAAS and EGNOS) as pdf-File](http://esamultimedia.esa.int/docs/egnos/estb/SBAS_receivers.pdf) [INMARSAT](http://www.inmarsat.org/)