The Global Positioning System

Abstract

GPS is a fantastic worldwide location system within the reach of anybody. It allows knowing almost instantly with a standard precision of only tens of meters, the coordinates and elevation where a proper receiver is located. Its basic operation is simple as reading a compass, and its whole theoretical fundamentals are just high-school level.

However, despite the inexpensiveness and apparent simplicity of any GPS receiver, a complex high technological system is hidden beyond its naïve box. This is the price paid to fulfil the very hard demands that real world compelled for the materialization of the easy concepts involved. Certainly, there is no other system for massive layman use that requires such indispensable sophisticated resources as GPS does.

This project report basically describes how GPS works, and why it operates the way it does in order to avoid, or alternatively to face and solve, its major practical problems. A detailed analysis of all intricate topics affecting GPS accuracy is far beyond the intention of this work, but special attention has been paid to relativistic corrections. Relativistic phenomena greatly affect GPS operation, becoming an unquestionable proof about the full pertinence of Einstein's legacy.

Figure 1
The complete constellation of the current 24 active GPS satellites

1 - Introduction

Precise location positioning, anywhere but especially in open waters, has always been an issue of uncertainty and concern. As a matter of fact, if just a single event were to be selected as the cornerstone in navigation's history, it surely has to be the widespread use of the compass occurred in XI century.

The remarkable general technological improvement blossomed along last century making life much easier, also came to collaborate in this old problem by realizing that radio signals could be of great helpfulness. Radio signals have been used for aid navigation as early as the 1920s.

A major step forward was the introducing by the middle of last century of LORAN (Long Range Aid to Navigation), a position-finding system that measured the time difference of arriving radio signals. Although achieving a fairly good accuracy (250-300 meters), LORAN coverage was just limited to where the system was established (about only 5 % of Earth's surface), and it only provided latitude and longitude information.

The very first idea about using satellites for location positioning came immediately after the Sputnik (1957) proved to properly use a radio transmitter to broadcast telemetry information. The revolutionary new idea was that any particular exact position on Earth could be accurately located by measuring the satellite's radio signals, on condition that the satellite's orbital position were precisely known.

By the middle of the 1960s -although rudimentary- several satellite-based US military positioning systems existed and were continually upgraded. Finally, in 1973, the Air-Force was selected to consolidate all military efforts into a single satellite navigation system, thus fructifying the NAVSTAR (Navigation Satellite Timing and Ranging) Global Positioning System. This is the official name for what is nowadays universally called just as the "GPS" (Global Positioning System).

The first NAVSTAR test-satellite was launched in 1974, and twenty years later the system was entirely developed and in full worldwide service, totalizing a constellation of 24 operational satellites. The GPS program has been completely successful and is still funded through the U.S. Department of Defence [***1***].

2 - Theoretical concepts of location systems based on radio signals

The basic idea buttressing any location positioning system is to properly know the distance to some reference points, that is, points whose exact location is previously known. This process is generally known as "triangulation". Geometrically, triangulation requires to know the distance to just one more reference point than the considered spatial dimension (i.e. the distance to two different reference points on a one-dimension space, three for 2D, or four for 3D).

For example, the easy case of a two-dimension flat space is described in the following figure.

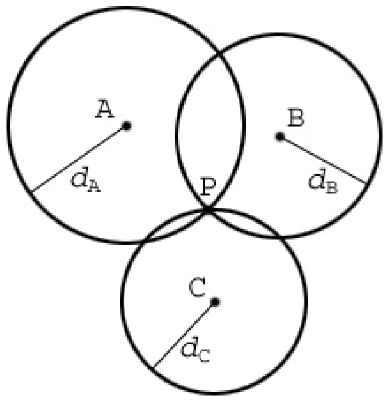


Figure 2
Triangulation for a flat two dimensions space
The unknown position of point P converges at the precise intersection of the three circumferences (radii d_A, d_B, d_C) around each reference point (respectively A, B, C)

The exact location of the point *P* is found out from the proper knowledge of the three respective distances (d_A, d_B, d_C) to each one of the three reference points (*A*, *B*, *C*). It is interesting to underline that just from the knowledge of only two distances (for example, d_A and d_B) the uncertainty in the exact location of P becomes reduced to only two possibilities -the two intersections of circles of radii d_A and d_B- , and the third data (d_C) is only necessary to resolve this "minor" ambiguity. In this theoretical context, any measurement to other reference points will supply no further information but to confirm the two coordinates already found.

Distances are measured using radio signals. Each reference point has a proper transmitter, and a corresponding receiver is placed at the unknown *P* point. The determination of the unknown distance (d_i) to each reference point (x_i ,y_i) is made by means of measuring the time (Δt_i) it took for the emitted radio signal to travel up to the receiver, and multiplying it by the signal's speed (s), becoming

$$d_i = s * \Delta t_i \qquad [1]$$

The determination of the unknown *P* position (x_0 ,y_0) in the flat plane results from the resolution of the following system of second grade equations

$$(x_0 - x_i)^2 + (y_0 - y_i)^2 = (d_i)^2 \qquad 1 \le i \le 3 \qquad [2]$$

where, as said, three equations are formally needed for the strict resolution of (x_0 ,y_0).

Referring to our 3D space, triangulation works exactly the same, but now there will be necessary to know the distances to four reference points. Therefore, the *P* position -of rectangular coordinates (x_0 ,y_0, z_0) centred in Earth's centre- will correspond to the unique intersection of four resulting spheres. This implies to resolve the system

$$(x_0 - x_i)^2 + (y_0 - y_i)^2 + (z_0 - z_i)^2 = (d_i)^2 \qquad 1 \le i \le 4 \qquad [3]$$

where four equations are needed in order to find out the (x_0 ,y_0, z_0) final solution, which in turn could be easily converted into corresponding longitude, latitude, and altitude.

In theory, adding new distance measurements to the previous perfect system will be redundant. In practice, measurements will have some discrepancies (error) in regard to exact distance values, thus translating into a much greater uncertainty in the final requested solution. Therefore, the final accuracy of the location (x_0 ,y_0, z_0) can always be improved far beyond individual distance measurement errors, at the simple cost of adding more distance measurements to the four required ones.

The concept works like follows. Errors that degrade distance measurements can be regarded as affecting all the measurements by the same unknown factor (ε), *at least in a first approximation*. Therefore, each measurement formally becomes [***2***]

$$d_i\,(\Delta t_i\,,\,\varepsilon) = s * \Delta t_i\,(\varepsilon) \qquad [4]$$

where a new unknown variable (ε) is incorporated to equation system [3]. Therefore, at least one more equation is required for its formal resolution, thus becoming

$$(x_0 - x_i)^2 + (y_0 - y_i)^2 + (z_0 - z_i)^2 = [d_i(\Delta t_i, \varepsilon)]^2 \qquad 1 \le i \le 5 \qquad [5]$$

Summarizing, the theoretical structure of a 3D location positioning system based on radio signals is very simple, and lies on the following fundamentals:

- At least four radio transmitters located at exact reference points (x_i, y_i, z_i) have to be accessed from the receiver, placed at the unknown point (x_0, y_0, z_0).
- The accuracy in the final (x_0, y_0, z_0) location depends on the respective distance measurement errors to each reference point.
- Introducing a fifth distance measurement to another independent reference point greatly improves the final accuracy of the required (x_0, y_0, z_0) location, as errors similarly affecting distance measurements can be properly eliminated.

3 - Atomic clocks

The very first practical difficulty in the GPS world is to accurately measure travel times. Considering that radio signals will be travelling at about $3x10^8$ m/s, from equation [1] it results that any minute timing error of only 1 millisecond will be "immediately" translated into 300 meters of distance inaccuracy. This poses a very high demand on the standard quality of the real clocks to be used, as all transmitter and receiver clocks should run at exactly the same rate.

Accurate clocks measure time by counting some kind of resonation. Precise resonation repeatability directly accounts for the quality of the measurement. Even precise quartz crystal resonators (with usual accuracies of 1 part in 10^5) are absolutely short for the extreme accuracy of GPS demands. A much more consistent resonator is required, and there is not such solution other than down in the atomic world.

Any atom or molecule that undergoes quantum transitions (energy changes) accordingly emits or absorbs electromagnetic radiation always at some exact frequency, characteristic of the considered atom or molecule. This is the reason that explains why the SI (International System) unit of time -the second- is currently defined as *"the interval of time taken to complete 9,192,631,770 oscillations of the cesium 133 atom exposed to a suitable excitation"* [1] [***3***]. Therefore, by definition, cesium 133 exactly resonates at 9,192,631,770 cycles per second.

Nowadays, the achievable accuracy of atomic clocks is better than one second per one million years (1 part in 32,000,000,000,000). This stability and precision makes it possible that atomic clocks can keep time better that any other artificial or natural clock (even better than Earth's rotation).

Atomic clocks become an essential gear of the GPS program. However, the construction of an atomic clock implies high-standard technology (basically, an ultra stable crystal oscillator locked to the atomic resonation) which in turns means a heavy and expensive instrument (about U$S 70,000 to 100,000).

[1] Definition established in 1967 by the 13th General Conference on Weights and Measures (SI).

4 - Development of a simple satellite-based location positioning system

So far, for a suitable location system it has been proved that atomic clocks are required at both each transmitting (reference) points, and the receiving (unknown) point. Let us analyse the operation of a simple GPS by assuming that each reference point is fixed in the sky and then appraise its whole suitability.

Current satellite technology permits to place and sustain satellites in precise locations on geostationary orbit -that is, permanently placed on the same terrestrial coordinates, somewhere above Earth's equator. The altitude of geostationary orbit can be derived from Newton's form of Kepler's third law:

$$p^2 = \left[\frac{4\pi^2}{G(m_1 + m_2)} \right] a^3 \qquad [6]$$

where p is de sidereal period of orbit, in seconds; a is the semimajor axis of orbit, in meters; m_1 and m_2 are the mass of the central and orbiting objects, in kilograms; and G is the universal constant of gravitation. Considering that the mass of the satellite is minute compare to the mass of Earth ($m_1 \rangle\rangle m_2$), it becomes

$$p^2 = \left[\frac{4\pi^2}{6.67x10^{-11} x5.97x10^{24}} \right] a^3 = 9.91\, x10^{-14}\ a^3 \qquad [7]$$

Imposing that the orbit must be geostationary, the sidereal period has to be exactly equal to one day, that is p = 86,000 s. Substituting

$$a^3 = \frac{(8.64x10^4)^2}{9.91x10^{-14}} \Rightarrow a = 4.22\, x10^7\,\text{m} \qquad [8]$$

This result implies that the satellites should be placed about 35,860 km above the Earth's equator (42,200 km minus the terrestrial equatorial radius, which is 6,378 km). Assuming that for a given point on the equator there should be at least five geostationary satellites located in an arc spanning 120 degrees centred in the vertical (see Appendix 1), this implies that 14 satellites would be needed for the whole system.

This satellite arrangement would properly work for low terrestrial latitudes. But as soon as latitude grows, two problems begin to arise.

First of all, geostationary satellites will appear in the sky always "concentrated" from the same side (in the northern sky for negative latitudes, or in the southern sky for positive ones). This geometrical condition of having all reference points concentrated in just a common "region" of the sky instead of having them evenly scattered ahead, intrinsically derives in a more inaccurate final result.

Figure 3 visualizes the problem just for 2D. For a given error in each distance measurement, if two considered reference points lay themselves well separated with respect to the receiver, as presented on the left side, the resulting uncertainty becomes much lesser than if the reference sources were relatively closer from each other, as showed at the right side of the same figure.

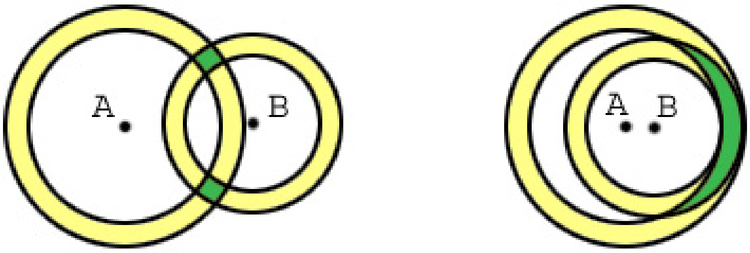


Figure 3
Geometry errors for location positioning in 2D
At left, where reference points A and B are greatly separated, the green area of their shared distance errors is notoriously smaller than for the case shown at right, where points A and B are much closer

The second problem, much more relevant, is that the local altitude of each satellite in the sky would go down and down as the terrestrial latitude enlarges (Appendix 2 shows the computation). For instance, at 60° of latitude the highest altitude that a geostationary satellite would appear in the local sky is slightly over 20°. Considering that this is just for the best case -a satellite located on the local meridian- the remaining requested satellites would reach altitudes of no practical use, thus making the geostationary option completely unsuitable for global use.

Having discarded the "easy" theoretical geostationary alternative, the next and final solution is to place satellites on convenient orbits for proper global coverage. This implies that (a) the new condition of variable locations for each reference point must be taken into account, and (b) a greater number of satellites must be used.

To know the correct location of a non-stationary satellite at a given time is quite simple, as orbital predictions can be perfectly done with an accuracy of only a few meters of error [***4***]. Therefore, each receiver should include a trustful satellite data base that allows calculating the corresponding orbital locations for any requested time.

The needful number of satellites will result from the determination of the optimal orbital distribution that assures global coverage with minimum satellites -which is not an easy problem. If 14 satellites on geostationary orbit were required for just a partial terrestrial coverage, at first sight it seems that conveniently selecting four or five different orbital planes, with about 6 to 8 satellites per orbit, could be enough. As previously stated, the greater the number of satellites, the better the final precision of the system.

5 - Practical improvements for a simple location positioning system

Now that a real global location system is being considered, one major issue to take into account is that computational work, no matter its demand on complexity, can be both easily and relative inexpensively performed with nowadays standard equipment. Therefore, replacing required inputs by adding extra computational work seems quite logical on condition that the same final accuracy will be achieved, especially if it results in the exemption of complex or expensive devices.

The first input saving that can be immediately applied to a real global location system is to eliminate one distance measurement and “replace” it just by software computation. This is possible because in real cases, the point $P\ (x_0, y_0, z_0)$ will necessarily be located somewhere “near” the Earth’s surface, so one of the two final alternatives will be discarded just by software (the one laying far away in “space”).

This software “trick” reduces in one unity the number of required equations in order to solve the location of the point $P\ (x_0, y_0, z_0)$ or which is equivalent, just only three accurate distance measurements are needed for the final solution.

The second input saving is to eliminate the expensive necessity of having an atomic clock at the receiver location by just incorporating a new distance measurement. This great idea is based on the concept that as long as all distance measurements to the reference points *are being performed simultaneously at the receiver location*, the receiver clock error -although unknown- is the same for all of them.

Therefore, the receiver clock error can be regarded as a new unknown to be resolved, which translates in the necessity of one more equation for solving the system. In practice, the earliest request of an expensive atomic clock at the receiver (point *P*) becomes exchanged by just adding of one more “imperfect” distance measurement.

In consequence, the real global location system will demand at least four simultaneous distance measurements to reference points to find out the unknown location of the point $P\ (x_0, y_0, z_0)$ with a final accuracy that will depend essentially on the distance measurement errors.

In order to achieve a distance error of, say 1,000 m, the margin error just from only considering timing inaccuracies should be less than 3.33 milliseconds (the time it takes radio signals to travel 1,000 m). With atomic clocks this becomes no big deal. The real problem lays in all remaining conditions to be satisfied: perfect synchronism, accurate satellite location, and exact knowledge of the speed of signals.

Concluding, a simple location positioning system will imply the use of about 30 satellites on convenient non-stationary orbits, onboard carrying synchronised atomic clocks, and continuously transmitting radio messages to be decipher by any receiver on Earth. With just four of such inputs, any receiver will be able to perform the required computations and finally obtain its current location with a reasonable expected error (about several hundred meters).

6 -Relativistic corrections

As strange as it sounds, truth is that GPS would not achieve an acceptable degree of accuracy without corrections provided by relativity. Although very subtle, relativistic effects make that real differences appear in the running times of clocks on non-stationary satellite orbits compared to clocks laying on Earth's surface. And, as already seen, the basic resource in the whole GPS affair is time's measurement exactitude.

Regarding the GPS world, relativity phenomena are involved due to two main facts:

(1) clocks on non-stationary satellite orbits and clocks on Earth's surface are moving with respect to each other, and according to Special Relativity, relative movement implies differences in running times; and

(2) clocks on satellite orbits and clocks on Earth's surface are lying at different distances from Earth's centre, and according to General Relativity, different gravity values also imply differences in running times.

In order to mathematically deal with spacetime above real Earth's surface -that is, to resolve the corresponding *Einstein's field equations*- it is necessary to firstly define a suitable simple theoretical model which properly describes it.

Although the *Schwarzschild metric* was actually developed for spacetime around a non-spinning spherically symmetric body, regarding GPS purposes it becomes a very good approximation, as Earth rotates slowly[2] and its overall shape is quite spherical [*5*].

Therefore, the equation that computes the timelike spacetime interval (τ), -that is, the proper time- above Earth's surface between two close enough events occurring on a spatial plane that passes through the centre of gravitational attraction is given by[3]

$$d\tau^2 = \left(1 - \frac{2M}{r}\right)dt^2 - \frac{1}{1 - \frac{2M}{r}}dr^2 - r^2 d\phi^2 \qquad [9]$$

where t is the time called "far-away time" (measured far away from the centre of attraction), r and ϕ are the spatial coordinates on the considered plane (respectively, the distance from Earth's centre, and the azimuthal angle with respect to some arbitrary initial direction), and M is Earth's gravitational mass, being all coefficients of the equation expressed "in units of length". Hence

[2] Effectively, the distortion of Earth's spacetime exclusively due to the rotational effect (the so called frame-dragging effect) in practice results orders of magnitude lesser compare to the gravitational effect.

[3] It is basically the metric for proper time in a flat spacetime $d\tau^2 = dt^2 - dr^2 - r^2 d\phi^2$ with the addition of the *curvature factor* (1-*2M/r*) affecting just the first two coordinates (*t* and *r*) but not the third one (ϕ), as the gravitational object (Earth) has been assumed to be spherically symmetric.

$$M = \frac{Gm}{c^2}$$

where G is the universal gravitation constant, m is Earth's gravitational mass expressed in kg, and c represents the speed of light, thus resulting

$$M = \frac{6.673x10^{-11}x5.974x10^{24}}{\left(2.998x10^{8}\right)^2} = 4.435x10^{-3}\,\text{m} \qquad [10]$$

Applying equation [9] to a clock on a satellite in circular orbit (distance $r_{satellite}$ always being constant, so $dr = 0$) and dividing by the square of the far-away time (dt^2), the corresponding differential proper time for the satellite clock ($d\tau = dt_{satellite}$) results

$$\left(\frac{dt_{satellite}}{dt}\right)^2 = \left(1 - \frac{2M}{r_{satellite}}\right) - r_{satellite}^2\left(\frac{d\phi}{dt}\right)^2 = \left(1 - \frac{2M}{r_{satellite}}\right) - v_{satellite}^2 \qquad [11]$$

where the tangential velocity along a circular path ($v = r\frac{d\phi}{dt}$) expressed in adimensional units (real speed divided by c) has been introduced.

Proceeding in the same way for a clock on Earth's equator[4], its corresponding differential proper time ($d\tau = dt_{Earth}$) becomes

$$\left(\frac{dt_{Earth}}{dt}\right)^2 = \left(1 - \frac{2M}{r_{Earth}}\right) - r_{Earth}^2\left(\frac{d\phi}{dt}\right)^2 = \left(1 - \frac{2M}{r_{Earth}}\right) - v_{Earth}^2 \qquad [12]$$

Diving equation [11] by [12], and taking square roots, it results that the relativistic difference in rates ($dt_{satellite}/dt_{Earth}$) between a clock placed on the orbiting satellites and a clock placed on Earth's equator becomes

$$\frac{dt_{satellite}}{dt_{Earth}} = \left(\frac{1 - \frac{2M}{r_{satellite}} - v_{satellite}^2}{1 - \frac{2M}{r_{Earth}} - v_{Earth}^2}\right)^{1/2} \qquad [13]$$

Equation [13] effectively accounts for the two expected relativistic effects: the incidence of the relative motion becomes explicitly represented by the presence of the two velocities $v_{satellite}$ and v_{Earth}, and the incidence of the different gravity values, represented by the presence of the different radii $r_{satellite}$ and r_{Earth} in their respective curvature factors.

[4] For which also $dr = 0$, as Earth has been assumed to be perfectly spherical.

In order to evaluate just the incidence of the effect of gravity in the different clock rates, equation [13] will be analyzed for supposedly "stationary" clocks (each $v = 0$) at radii $r_{satellite}$ and r_{Earth}, thus resulting

$$\left(\frac{dt_{satellite}}{dt_{Earth}}\right)_{(stationary)} = \frac{\left(1-\frac{2M}{r_{satellite}}\right)^{1/2}}{\left(1-\frac{2M}{r_{Earth}}\right)^{1/2}} \qquad [14]$$

From classical mechanics[5], the acceleration of a satellite moving in a circular orbit of radius r with constant speed V is a vector pointing towards the centre with a module equal to V^2/r. Therefore, that acceleration module times the satellite mass (m_{sat}) has to be equal to Newton's gravitational force exerted by Earth:

$$\frac{m_{sat}V_{sat}^2}{r_{satellite}} = \frac{Gm_{Earth}m_{sat}}{r_{satellite}^2} \Rightarrow V_{sat} = \sqrt{\frac{Gm_{Earth}}{r_{satellite}}} \qquad [15]$$

GPS satellites orbit Earth with a sidereal period ($T_{satellite}$) of 12 hours, thus their orbital speed V_{sat} can also be derived from the simple relation

$$V_{sat} = \frac{2\pi r_{satellite}}{T_{satellite}} \qquad [16]$$

Replacing [15] in [16], it results that

$$r_{satellite} = \left(\frac{T_{satellite}^2 Gm_{Earth}}{4\pi^2}\right)^{1/3} = \left[\frac{(12x3{,}600)^2 x6.673x10^{-11}x5.974x10^{24}}{4\pi^2}\right]^{1/3}$$

Therefore, $r_{satellite}$ results equal to $2.661x10^7$ m (that is, 20,230 km above Earth's surface, and about four times greater than r_{Earth}), making V_{sat} to be equal to 3,870 m/s.

As both terms $2M/r_{satellite}$ and $2M/r_{Earth}$ result much lees than unity (respectively $3.333x10^{-10}$ and $1.391x10^{-9}$), equation [14] can be approximated by applying the binomial theorem, thus remaining

[5] Finding out the orbital speed V by means of Euclidean geometry and Newtonian mechanics for later mixing in a relativistic computation a priori seems to be improper. However, it can be proved that the introduced difference with respect to the real speed, for the particular GPS conditions, is completely insignificant.

$$\left(\frac{dt_{satellite}}{dt_{Earth}}\right)_{(stationary)} \approx \left(1-\frac{M}{r_{satellite}}\right)\left(1+\frac{M}{r_{Earth}}\right) \approx 1-\frac{M}{r_{satellite}}+\frac{M}{r_{Earth}} \qquad [17]$$

$$\left(\frac{dt_{satellite}}{dt_{Earth}}\right)_{(stationary)} \approx 1-\frac{4.435x10^{-3}}{2.661x10^{7}}+\frac{4.435x10^{-3}}{6.378x10^{6}} = 1+5.287x10^{-10} \qquad [18]$$

Therefore, the fractional difference in rates between clocks placed at the position of GPS satellite orbits and at Earth's surface, that is, only due to their difference in altitude alone without considering the relative motion among them, becomes $5.287x10^{-10}$.

This important result shows that satellite clocks (where gravity is comparatively weaker) will run faster than Earth's surface clocks due to position effects alone by 45,680 nanoseconds per day ($5.287x10^{-10}$ times 86,400 s/day).

As "moving clocks run slower" -one of Special Relativity's mottos- the effect of the so far ignored real relative motion of satellite clocks with respect to Earth's clocks will certainly be opposite to the already computed consequences of different altitudes.

The speed of a clock fixed on Earth's equator ($V_{Equator}$) due to the once per day terrestrial rotation becomes 2π times r_{Earth} divided by 86,400 seconds, that is, 463.8 m/s.

Therefore, the respective adimensional velocities for equation [13] result

$$v_{satellite} = \frac{V_{sat}}{c} = \frac{3{,}870}{2.998x10^{8}} = 1.291x10^{-5}$$

$$v_{Earth} = \frac{V_{Equator}}{c} = \frac{463.8}{2.998x10^{8}} = 1.547x10^{-6}$$

being $v_{satellite}$ about eight times greater than v_{Earth}.

Considering that both speed terms are much lesser than unity, the same valid approximation as before can be made to equation [13], thus resulting

$$\frac{dt_{satellite}}{dt_{Earth}} \approx 1-\frac{M}{r_{satellite}}-\frac{v_{satellite}^{2}}{2}+\frac{M}{r_{Earth}}+\frac{v_{Earth}^{2}}{2} \qquad [19]$$

$$\frac{dt_{satellite}}{dt_{Earth}} \approx 1-\frac{4.435x10^{-3}}{2.661x10^{7}}-\frac{\left(1.291x10^{-5}\right)^{2}}{2}+\frac{4.435x10^{-3}}{6.378x10^{6}}+\frac{\left(1.547x10^{-6}\right)^{2}}{2}$$

$$\frac{dt_{satellite}}{dt_{Earth}} \approx 1-4.44x10^{-10} \qquad [20]$$

This total discrepancy, accounted for a time period of 24 hours, implies that satellites clocks appear to run faster than terrestrial clocks by 38,370 nanoseconds per day ($4.44x10^{-10}$ times 86,400 s/day).

This overall difference in clock rates is really important for the operation of GPS. Assuming that satellite clocks and terrestrial clocks could be somehow synchronized as frequent as, say 1hour, just that tiny difference would introduce a potential uncertainty in time measurements of about 1,598 ns ($4.44x10^{-10}$ x 1h), equivalent to an error of 480 m.

This relativistic analysis has been made considering that Earth clocks were placed on the equator, and consequently, the speed v_{Earth} was the largest possible one. As Figure 4 depicts, anywhere else on Earth's surface its value will be smaller due to the shorter distance to the axis of rotation, down to becoming just zero on the poles, where Earth clocks are stationary relative to the gravity centre.

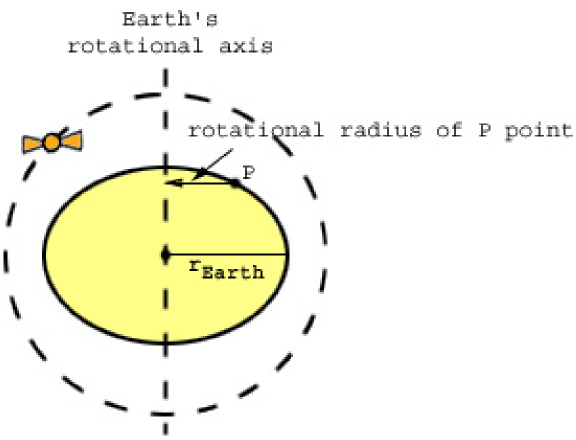


Figure 4
The rotational speed of any fixed clock on Earth surface depends on its distance to Earth's rotational axis, so that the real speed becomes variable with latitude. On the other hand, the real distance to Earth's centre is also latitude-dependent due to Earth's actual shape

However, the variation of the v_{Earth} value due to the actual latitude position has no practical incidence in equation [19] because it becomes compensated with the fact that Earth is not spherical, but bulging a bit at equator due to its rotation, as Figure 4 exaggeratedly shows. Therefore, the smaller distance to the gravitational centre at latitudes nearing the poles actually increases the M/r_{Earth} term by the same amount that the velocity term v_{Earth} decreases (due to the "rotational" distance dependence on latitude). This makes that the relativistic correction found out just for the equator really applies to all possible latitudes [***6***]. A compelling computation has been included in Appendix 3.

This analysis of GPS operation has properly identified and quantified the two main relativistic effects, although there are others of much less importance[6], *"too small to affect the system at current accuracy levels, but may become important as the system is improved"* [***7***].

Finally, having proved that time flowing on satellites runs at a different rate than time on Earth's surface, the found difference must be properly neutralized. The solution is to set atomic clocks on satellites to run at a rate $4.44x10^{-10}$ times slower than they would do on Earth's surface.

This is accomplished as Figure 5 proposes. Instead of considering that 1 second has passed after counting 9,192,631,770 cycles, atomic clocks on satellites are set so that 1 second is acquainted only after counting 9,192,632,774 cycles (9,192,631,770 divided by $1\text{-}4.44x10^{-10}$). With this correction, GPS real uncertainty can be reduced to only a few meters.

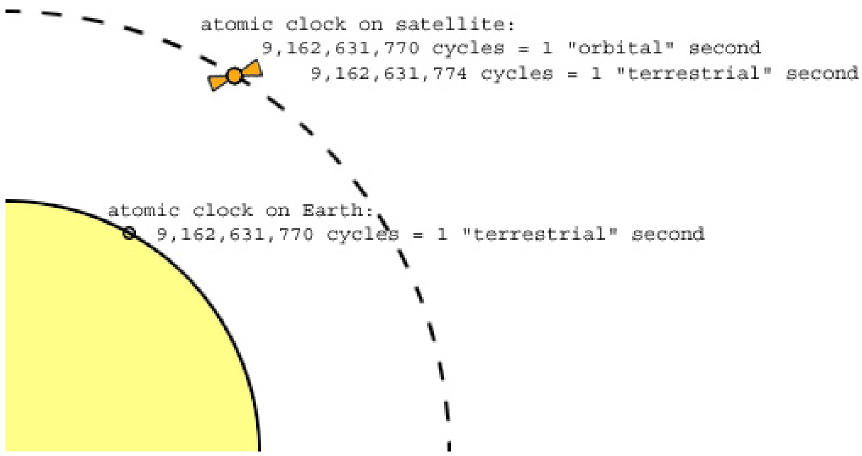


Figure 5
As time runs faster in satellite orbits compare to terrestrial time, the same 9,162,632,770 cesium atomic cycles that defines one second on Earth actually occur earlier onboard satellites, thus requiring an extra lapse of four more cycles in order to measure time like Earth's atomic clocks

Summarizing, for GPS operation (a) relativistic effects are certainly important and must be taken into account; (b) the incidence in clock rate discrepancies is opposite in sign but far greater from the fact that satellites are orbiting higher above the surface, than from the fact that satellites are moving with respect to ground; and, (c) the elegant solution to counteract both relativistic effects is to make atomic clocks onboard satellites to run at a slower rate than they would normally do around Earth's surface.

[6] Basically, the already mentioned frame dragged effect, the non-isotropy of speed light (the so called Sagnac effect), and frequency shifts of clocks in satellites due to earth's quadrupole potential.

7 - Description of the current GPS program

Current GPS program still runs under the authority of the U.S. Department of Defence, who is in full charge of running, monitoring and controlling its permanent operation.

GPS worldwide coverage is presently accomplished by a constellation of 24 operational satellites, and four additional spare ones. They are deployed in six evenly spared planes (four satellites per orbit) with an inclination of 55° with respect to the equator, flying in circular orbits at an altitude about 20,230 km above Earth with a sidereal period of 12 hours, as shown on page 1 (Figure1).

With such arrangement, it is assured the simultaneous presence of four to eight satellites above 15° elevation at any time, everywhere on Earth's surface. If the elevation mask is reduced to 10°, occasionally up to 10 satellites will be visible, while even two more could be seen at elevations between 10° and 5° [***8***].

Instead of carrying just one atomic clock, each GPS satellite actually makes use of four onboard time standards (two cesium atomic clocks, plus two rubidium atomic clocks), thus assuring a long term precision of about a few parts in 10^{-14} over one day. These highly accurate frequency standards produce the fundamental frequency (10.23 MHz) from which the two presently carriers (L1 = 1575.42 MHz and L2 = 1227.60 MHz) are derived [***9***].

Effectively, all GPS satellites are permanently transmitting by frequencies L1 and L2 their corresponding information (a unique pattern for timing purposes which also allows identifying them, plus some other orbital information), technically called their respective *pseudo random code*, or PRC.

PRC signals are very complex for two reasons. Firstly, its complexity avoids the chance of confusion of the proper identity of the emitting satellite; which in turn allows using the same frequencies for all satellites. Secondly, its particular layout permits to be efficiently detected at the receiver, thus avoiding any necessity of having impractical big reception dishes for GPS signals [***10***].

Each GPS receiver internally generates an identical PRC signal corresponding to each one of the satellites being received (at least four channels will be needed to perform the measurements simultaneously), so that by means of digital processing the travelling delay for each signal can be obtained after some minimum required time (basically in order to properly acknowledge PRC signals from each accessible satellite).

Current accuracy for GPS standard receivers (costing less than U$S 200) is about 15 meters in coordinates and 30 meters in elevation. Higher cost equipment (which makes use of especial sophisticated techniques) easily reaches accuracies better than 3-5 meters in coordinates and 8-10 meters in elevation.

8 - Analysis of error sources in real GPS

GPS error sources are derived due to inevitable real-world "imperfections". Thus, errors can arise, and they certainly do, from all possible stages of the system, becoming in: (1) space errors, (2) atmospheric delay errors, (3) reception errors, (4) geometry errors.

Space errors

They derive basically from inaccuracies in satellite clocks and in satellite positions. Satellite clocks are continuously monitored from GPS's operation authority, by comparison to master control clock systems in order to detect and calculate all satellite clock drifts. In the same way, satellite orbits are also permanently monitored and measured.

Messages with those found differences are then sent to the satellites, so that they could be included in the information that satellites are permanently transmitting. This is the key that later allows GPS receivers to take them into account, thus improving the exactness of the measurements. *"Any remaining satellite clock errors accumulate typically to about a few nanoseconds, which cause a distance error of about one meter"* [***11***].

Atmospheric delay errors

The medium that GPS radio signals have to travel is not uniform at all, as about the first 95% is space and the final 5% is the complete Earth's atmosphere. However, this tiny part is far from being homogenous, permanently varying local densities and compositions, which introduces unpredictable delays with respect to speed of light in vacuum. Hence, the actual speed of GPS signals is always uncertain.

For analysing the travelling speed of GPS signals two atmospheric layers must be specifically considered -the ionosphere and troposphere- with completely different characteristics. Figure 6 graphically shows those layers, obviously out of scale.

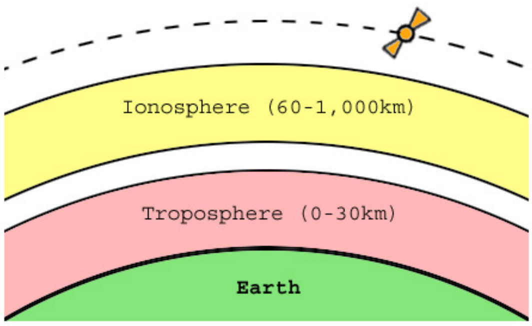


Figure 6
The two atmospheric variable layers that differently affect the speed of GPS radio signals

The upper layer of the atmosphere, called the ionosphere, contains charged particles whose variable density affects differently the code and carrier GPS signals (slowing down the first, while speeding up the later) [***11***]. Its overall effect also varies from day to night, and has a cyclical period of about 11 years, so it is greatly unpredictable. However, as the ionosphere effect depends on the transmitted frequency, its delay uncertainty can be completely eliminated by the simple solution of transmitting the same information by two different frequencies (the already mentioned L1 and L2). Once the receiver analyses the difference on their respective arrival times, the ionosphere delay is calculated and removed.

The lower layer of the atmosphere, called the troposphere, contains water vapors that slow down both code and carrier GPS signals. This effect can only be removed by proper computation once current water vapour content, temperature and pressure are known.

All space errors and atmospheric delay errors affect receivers placed not far away from each other in the same way and with identical magnitude. This fact gives the chance to compensate them, just after a relative closer receiver located on an exact fixed position would have measured them and automatically transmitted the corresponding corrections in almost real time. This technique is called "differential mode" or DGPS, and as long as the receiver is around 500 km the monitor station, *"accuracies at the 1-5 m level can be routinely achieved"* [***12***].

Reception errors

Reception errors are due to external or internal causes. The typical external cause is when the receiver receives the signal emitted by the satellite not only in the proper direct way, but also from indirect paths after being reflected in nearby surfaces. This case is exemplified in Figure 7, where besides the blue direct path, two other signal reflections arrive at the receiver interfering the original signal.

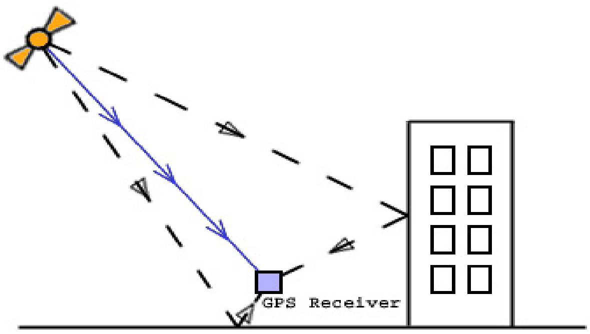


Figure 7
Multipath occurs when not only the direct GPS signal reaches the receiver, but also reflections

If not completely eliminated, multipath errors can be at least minimized by means of signal processing techniques (especially suitable for longer indirect paths) or multipath rejection devices (especially suitable for reflections below the GPS antenna).

Receiver internal errors are those introduced just by the receiver itself in the measuring or processing of the GPS signals, and could become minimized at the price of improving the processing quality of GPS receivers.

Geometric errors

Strictly speaking, the geometric effect is not an error by itself, but just the inevitable uncertainty in the final location that basically is derived from the particular geometrical arrangement of the satellites at the time of the measurements, but also from their respective altitudes in the sky.

This important fact can change a position error of just 1 meter due to all previous analysed errors (*range error*), into 1.5 meters for the best case (four satellites evenly scattered in the sky around the zenit), or up to tens or hundreds of meters for the worst case (four satellites clustered near each other and at relative low altitudes).

The geometric effect is technically called Geometric Dilution Of Precision (GDOP) and can be roughly interpreted as the ratio of the position error to the range error. The larger the angular area covered by the satellites, or the larger their number, the better (smaller) the GDOP number. A typical value for GDOP is 2, which means that the position final accuracy will be twice as much as the resulting error from all measurements [***13***].

9 - Conclusions

The key to GPS accuracy is the fact that all signal components are precisely controlled by atomic clocks. Those are placed onboard a constellation of non-geostationary satellites that assures proper worldwide coverage. Thus, the only restriction for the correct operation of any GPS receiver is to place it in an open area where at least four satellite signals can be received. In return, accurate coordinates and altitude will be supplied after only a couple of minutes at most.

In order to allow the remarkable precision that nowadays any standard GPS receiver achieves, not matter its location anywhere on Earth (the term "remoteness" no longer belongs to GPS dictionary), many hard technical problems had to be conveniently solved. Most interestingly, relativistic effects, which for common terrestrial issues have no practical incidence at all, had to be especially taken into account in order to cancel their manifest role in the GPS play.

GPS precision will continue to improve, at the price of launching more satellites, multiplying DGPS monitoring stations, and developing better algorithms for neutralizing all errors sources.

References

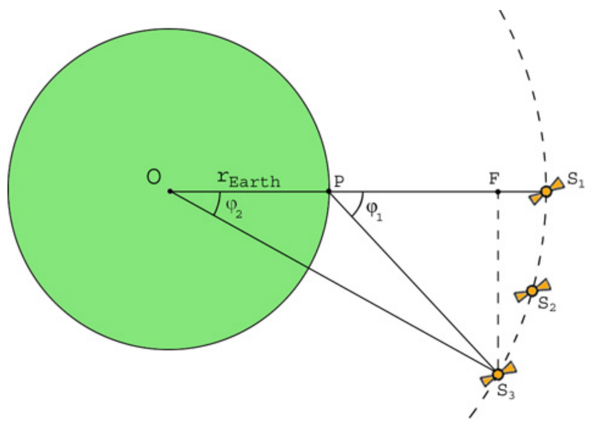
[**1**] *"GPS for Dummies"* by Joel McNamara, 2004, Wiley Publishing, Inc., page 50
[**2**] *"Global Positioning System: The Mathematics of GPS Receivers"* by Richard B. Thompson, Mathematics Magazine, Vol.71, N° 4, October 1998, pages 260-269
[**3**] *"How Atomic Clocks Work"* by Douglas Dwyer, at http://science.howstuffworks.com/atomic-clock.htm
[**4**] *"Field Methods in Ecology and Environmental Sciences"* by Charles Randolph Carlson, 2001, at www.vcrlter.virginia.edu/~crc7m/envsci796/ENVSCI796_1.doc
[**5**] *"Exploring Black Holes: Introduction to General Relativity"* by Edwin F. Taylor and John Archibald Wheeler, 2000, Addison Wesley Longman, Inc., pages A-2 and A-3
[**6**] *"Exploring Black Holes: Introduction to General Relativity"* by Edwin F. Taylor and John Archibald Wheeler, 2000, Addison Wesley Longman, Inc., page A-8
[**7**] *"General Relativity in the global positioning system"* by Neil Ashby at http://www.phys.lsu.edu/mog/mog9/node9.html
[**8**] *"Global Positioning System: Theory and Practice"* by Bernard Hofmann-Wellenhof, Herbert Lichtenegger, and James Collins, 5th ed, 2000, Springer, page 12
[**9**] *"Global Positioning System: Theory and Practice"* by Bernard Hofmann-Wellenhof, Herbert Lichtenegger, and James Collins, 5th ed, 2000, Springer, page 14
[**10**] *"GPS Technology, Random code"* at http://www.beaglesoft.com/gpstechnologhy.html
[**11**] *"Sources of Inaccuracy"* at http://www.topconps.com/gpstutorial.Chapter3.html
[**12**] *"Operation and Application of the Global Positioning System"* by Colleen H. Yinger at http://www.aero.org/publications/crosslink/summer2002/02.html
[**13**] *"Geometric Dilution Of precision"* at http://www.topconps.com/gpstutorial.Chapter3.html

Figure 1 obtained from www.vcrlter.virginia.edu/~crc7m/envsci796/ENVSCI796_1.doc

All other figures drawn by the author.

Appendix 1

In order to calculate how many geostationary satellites are needed to locate in orbit, it is necessary to firstly find out how the angle φ_1 (the angular separation from the zenith for the lowest satellite seen from an equatorial position P) relates to the angle φ_2 (the corresponding angle for the same satellite at the centre of Earth), as shown in the attached



figure, representing Earth seen from above (North Pole) and three geostationary satellites.

From the triangle *OFS₃*, it results

$$tg\varphi_2 = \frac{FS_3}{FO} = \frac{FS_3}{r_{satellite}\cos\varphi_2} \Rightarrow FS_3 = r_{satellite}\sin\varphi_2$$

From the triangle PFS_3, it results $tg\varphi_1 = \frac{FS_3}{FP} = \frac{FS_3}{h - FS_1}$

where $h = r_{satellite} - r_{Earth}$, and $FS_1 = (r_{satellite} - FO) = r_{satellite}(1 - \cos\varphi_2)$, so substituting

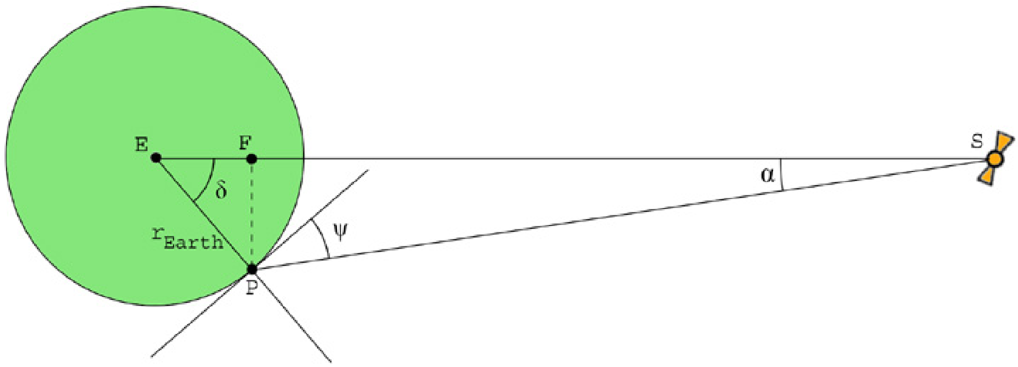
$$tg\varphi_1 = \frac{r_{satellite}\sin\varphi_2}{h - r_{satellite}(1 - \cos\varphi_2)}$$

Assuming that for a given point on the equator the lowest geostationary satellite should be located higher than an altitude of 30 degrees ($\varphi_1 = 60°$), the solution of the previous equation gives $\varphi_2 = 52°$. Therefore, if five satellites must be placed in just 104° of the full orbit, this means that (*360x4/104*) 14 satellites are requested.

The conservative election of 30° as the minimum altitude becomes justified by the fact that it was chosen for the best possible case, that is, for a point on the equator. As soon as latitude begins to enlarge, the observed altitude of geostationary satellites quickly drops, as shown in the next Appendix.

Appendix 2

The attached figure represents Earth as seen from the equatorial plane. The point *P* is located on the terrestrial surface at a latitude δ. From there, a geostationary satellite *S* appears to be observed at a local altitude ψ. From the triangle *EPS*, the unknown angle ψ becomes equal to (90° – δ – α).



From the triangle *EFP*, it results $FP = r_{Earth} \sin\delta$, and $EF = r_{Earth} \cos\delta$

From the triangle *SFP*, it results $tg\alpha = \dfrac{FP}{SF} = \dfrac{FP}{ES - EF}$, so substituting it becomes

$$tg\alpha = \frac{r_{Earth} \sin\delta}{r_{satellite} - r_{Earth} \cos\delta}$$

Replacing for r_{Earth} = 6,378 km, and $r_{satellite}$ = 42,233 km, it results that geostationary satellites are no longer visible from latitudes greater than δ = 81°, implying that this system can not theoretically cover areas 10° around the poles.

In practice, just for an intermediate latitude like δ = 60°, the angle α results 8.1°, making ψ = 21.9°. Being this low elevation the highest possible altitude than any geostationary satellite can reach in the sky at latitudes of 60° (meaning that the remaining required satellites will appear even lower), this shows that the geostationary solution would only properly work for low latitudes (up to about 50 – 60°).

Appendix 3

The fundamental equation found for the relativistic correction [19] is

$$\frac{dt_{satellite}}{dt_{Earth}} \approx 1 - \frac{M}{r_{satellite}} - \frac{v^2_{satellite}}{2} + \frac{M}{r_{Earth}} + \frac{v^2_{Earth}}{2}$$

As previously said in its numerical substitution, the two last terms were calculated just for the corresponding values on the equator.

Earth's real radius is not constant, varying from a maximum value at the equator (6,378.5 km) down to a minimum at the poles (6,357 km). An intermediate value is defined as the *mean radius* (R_{mean} = 6,371.3 km) resulting from the square root of Earth's surface divided by 4π, which corresponds to Earth's actual radius at about 35° of latitude.

As shown in Figure 5, the rotational speed varies according to the latitude (δ) of the considered point on the terrestrial surface, as the real distance to Earth's axis of rotation becomes the actual distance to the centre (r_{Earth}) times ($\cos\delta$).

The following table computes the numerical values for the two last terms of equation [19] corresponding to the three most particular latitudes: at the equator, at an intermediate one (35°), and at the poles. Earth's gravitational mass in units of length (M) was already found in equation [10], resulting $M = 4.435x10^{-3}$ m.

	r_{Earth} (m)	$V_{Earth} = \frac{2\pi r_{Earth}\cos\delta}{86,400}$	$v_{Earth} = V_{Earth}/c$	$M/r_{Earth} + v^2_{Earth}/2$
equator (δ = 0°)	*6,378,000*	*463.86* m/s	*1.547x10⁻⁶*	*6.97x10⁻¹⁰*
δ = 35°	*6,371,300*	*379.54* m/s	*1.266x10⁻⁶*	*6.97x10⁻¹⁰*
poles (δ = 90°)	*6,357,000*	*0* m/s	*0*	*6.97x10⁻¹⁰*

As previously stated, the fact that Earth is not spherical compensates the incidence of the term corresponding to terrestrial velocity, thus resulting that latitude has no practical effect on equation [19].