1993 IEEE INTERNATIONAL FREQUENCY CONTROL SYMPOSIUM

RELATIVITY IN THE FUTURE OF ENGINEERING

Neil Ashby
Department of Physics, Campus Box 390
University of Colorado
Boulder, CO 80309-0390

Abstract

Improvements in clock technology make it possible to develop extremely accurate timing, ranging, navigation, and communications systems. Three relativistic effects–time dilation, the Sagnac effect, and gravitational frequency shifts–must be accounted for in order for modern systems to work properly. These effects will be related in a non-mathematical way to fundamental relativity principles: constancy of the speed of light, and the principle of equivalence. Examples of current and future engineering applications will be discussed, such as in the Global Positioning System, in time synchronization systems, geodesy, and communications.

Atomic Clocks

Relativistic effects become important in applications requiring very accurate timing, time transfer, or synchronization. Many engineering systems are beginning to rely on modern atomic clocks which have fractional frequency stabilities of the order of 10^{-12} or 10^{-13}. An excellent example is the Global Positioning System (GPS), in which about a dozen relativistic effects must be accounted for in order for the system to work properly. Atomic clock technology not only provides the basis for the definition of the second as the unit of time, this technology is expected to improve rapidly in the future. Vessot *et at.*[1] have summarized potential future performance improvements in several promising devices including cryogenic H-masers, Cs fountains, and trapped Hg ions; these predictions are summarized in Fig. 1. These analyses show there is some hope that fractional frequency stabilities in the range 10^{-16} to 10^{-17} can be achieved. For this paper I shall however adopt a conservative fractional frequency stability figure of 10^{-15} as a guideline for determining what relativistic effects might be important in the future.

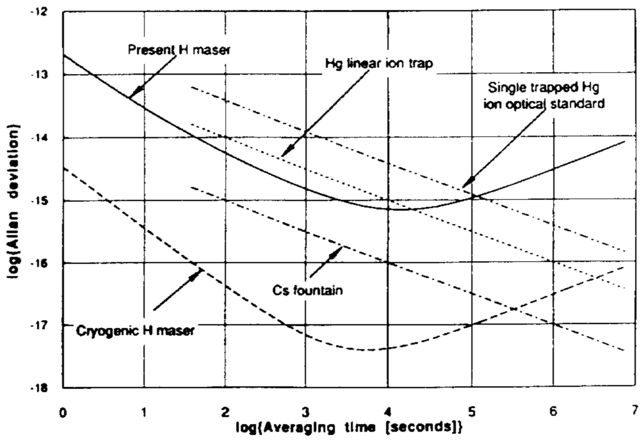


Fig. 1. Predicted Allan variance for future frequency standards. (This assumes no systematic effects in Cs and Hg devices.)

Constancy of the Speed of Light

Relativity enters metrology in a most fundamental way through the so-called 'Second Postulate' of the special theory of relativity, the principle of the constancy of the speed of light, c. This now widely accepted principle states that the speed of light in free space has the same value in all inertial systems, independent of the motion of the source. The speed of light is also independent of the motion of the observer. The numerical value of c has been defined by convention:

$$c = 299\ 792\ 498 \text{ meters/second.} \qquad (1)$$

In conjunction with the adopted unit of time, this value for c defines the SI unit of length, the meter. In thinking about the speed of light, a convenient alternative rule of thumb is that c is approximately equal to one foot per nanosecond (1 nanosecond = 1 ns = 10^{-9} second).

In an inertial frame of reference, the principle of the constancy of c provides a means of synchronizing remotely placed clocks. Consider two standard clocks, A and B, placed at rest a distance L (meters) apart. (The distance L could be found by measuring the time on clock A required for a light signal to propagate from A to B and back, and multiplying by $c/2$. This would not depend on the presence of a clock at B.) Now suppose a signal originates at clock A at time t_A. The time required for the signal to propagate in one direction from A to B is L/c. The clock at B will then be synchronized with that at A if the signal arrives at the time t_B given by

$$t_B = t_A + L/c. \tag{2}$$

This procedure is called the 'Einstein Synchronization Procedure' and clocks distributed at rest in any inertial frame will be presumed to be synchronized by this or an equivalent procedure.

Clearly in discussing electromagnetic signals as I have done above, I am ignoring quite a few practical difficulties. Signals must have sufficient spectral bandwidth that it is possible to reconstruct well-defined pulses in time. Noise in real clocks, frequency drifts due to environmental factors, etc., are not a concern here. Also I'm usually going to ignore effects on propagation speed which might arise because the signals propagate through a medium rather than through a vacuum.

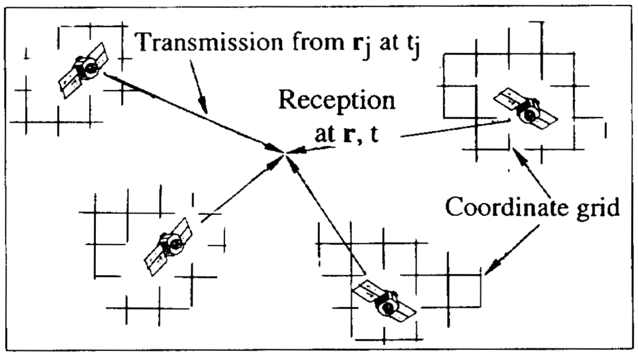


Fig. 2. Idealized conception of a navigation and time transfer system.

Navigation

Keeping these caveats in mind, the constancy of c leads to the following idealized conception of a navigational system. Referring to Fig. 2, suppose four transmitters, each with its own standard clock, are placed at known locations $\mathbf{r}_j$. Assume the clocks are synchronized by the Einstein procedure. There is a receiver at unknown position $\mathbf{r}$ carrying a standard clock which has not been synchronized. Let these transmitters rapidly transmit synchronized pulses which are tagged with the transmitter's position and time, so that a receiver can determine the time t_j and the location $\mathbf{r}_j$ of the pulse from transmitter j. The receiver's position $\mathbf{r}$ and clock time t can then be determined by solving four simultaneous propagation delay equations:

$$(\mathbf{r} - \mathbf{r}_j)^2 = c^2(t - t_j)^2; \qquad j = 1, 2, 3, 4, \tag{3}$$

for the unknowns $\mathbf{r}$ and t. These equations just express the principle of the constancy of the speed of light in an inertial frame. Clearly a timing error of one nanosecond would lead to a an error of about a foot in position determination.

Event Detection

There is a kind of reciprocity in this situation which can be used for event detection: suppose that instead of transmitters at the locations $\mathbf{r}_j$ there are receivers, tied to synchronized standard clocks. Suppose that an event occurs at the position $\mathbf{r}$ at time t causing a signal to be transmitted, which is received at the four receivers at the respective known positions $\mathbf{r}_j$ at the measured times t_j. Then by solving four propagation equations of the form of Eqs. (3), the position of the event and the time at which it occurs can be determined. If some information about the position of the event is available, it may be possible to locate the event by solving fewer than four propagation delay equations.

Fault Location

As an example of event detection using only two synchronized clocks, consider the problem of determining the location and time of a fault that occurs in a power line stretching between two detectors a distance L apart. In Fig. 3, clocks at the ends of the line are synchronized from some independent primary reference clock. A fault occurring at distances L_1, L_2 from the respective detectors at the ends of the lines sends out a signal at time t which is received at times t_1, t_2 at the respective ends of the line. A previous survey would give

$$L = L_1 + L_2, \tag{4}$$

whereas from the constancy of c, the times t_1, t_2 are related to the time t by propagation delay equations:

$$t_1 = t + L_1/c, \quad t_2 = t + L_2/c. \tag{5}$$

Solution of only two propagation delay equations, in conjunction with Eq. (4), gives the time and position of the fault. To locate the fault to within one foot requires synchronization to better than a nanosecond.

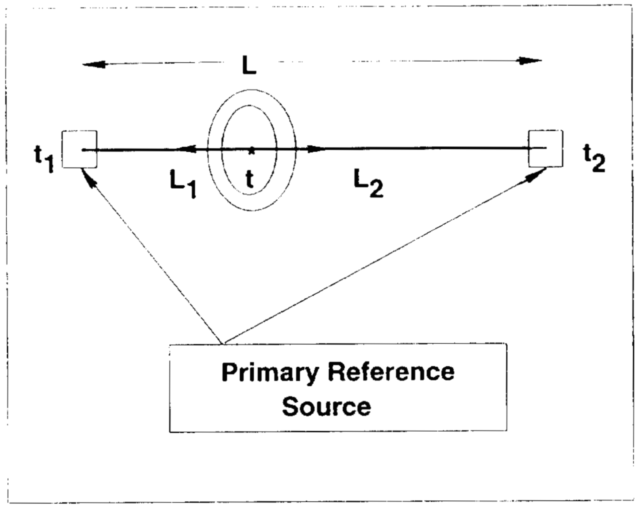


Fig. 3. Fault location using constancy of c.

Breakdown of Simultaneity

The discussion above assumes that the clocks are at rest in some inertial reference frame. Usually however clocks are in motion; for example in Fig. 2 the transmitters could be orbiting the earth. Relative motion introduces subtle new effects; perhaps the most profound of these is the breakdown of the concept of simultaneity. Events which appear to occur simultaneously in one inertial frame may not appear simultaneous to observers in some other inertial frame, which is moving with respect to the first. This is a direct consequence of the principle of the constancy of c.

In discussing measurements made by observers in two different, relatively moving inertial frames, one always imagines that each observer is equipped with his/her own measuring rods and standard clocks, that the clocks used by observers in one frame are at rest, and that they are synchronized by the Einstein procedure. In each of the inertial frames, any particular electromagnetic signal propagates with speed c.

Consider then as in Fig. 4. two events consisting of two lighting strokes which hit the two ends of a train of length $L = 2x$ simultaneously *as seen by observers on the ground*. The train is assumed to be moving to the right at speed v relative to the ground. For ease of discussion, I'll refer to the ground as the 'rest' frame, and the train as the 'moving' frame. Observers on the ground (in the rest frame) can determine the midpoint between the two lightning strokes, a distance x from either end of the initial position of the train. They will then find that light signals from the two events will propagate along the tracks and collide at the midpoint. This has nothing to do with the motion of the train.

Now look at the sequence of events involving a moving observer, sitting at the midpoint of the moving train. As the train moves forward, this observer approaches the oncoming light emitted from the event at the front of the train, and recedes from the light signal emitted from the event at the back of the train. Therefore the moving observer will encounter light from the front event first, and will have to conclude that the event at the front of the train occurred first. By the principle of the constancy of c, light must travel with speed c no matter what the value of the relative speed v is. So if light from event A arrives before that from event B, which is the same distance away, then event A must occur first.

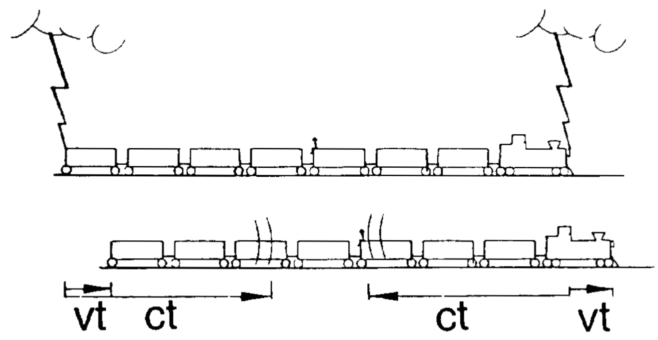


Fig. 4. Thought experiment illustrating relativity of simultaneity.

To analyze this approximately is not difficult. Suppose the zero of time for observers in both the rest and the moving frames is set to occur at the instant the midpoint of the train encounters the signal from the lightning stroke at the front of the train. I'll use primes to denote quantities measured by the moving observer. Then to the moving observer, the time t' of the stroke at the front of the train is

$$t' = -\frac{x}{c}. \tag{6}$$

To observers in the rest frame, however, the midpoint of the train is approaching the signal at the relative speed $c + v$, so to first order in v,

$$t = -\frac{x}{c+v} \approx -\frac{x}{c} + \frac{vx}{c^2}. \tag{6}$$

Therefore

$$t' = t - \frac{vx}{c^2}. \tag{8}$$

The term $-vx/c^2$ is a relativistic correction for breakdown of simultaneity. The effect is proportional to the relative velocity and proportional to the distance x.

Putting in some numbers, suppose $v = 1000$ km per hour (typical for a jet aircraft) and $x = 3500$ km. Then the correction is 108 ns. The negative sign in Eq.

(8) means that of two events simultaneous in the rest frame, to the moving observer the event farther out in front, at the more positive x, occurs earlier.

Sagnac Effect

The above discussion of the breakdown of simultaneity can be used to understand the peculiar physics on the edge of a slowly rotating disc. The prime engineering application is to time transfer and synchronization on the surface of the rotating earth. For purposes of illustration, therefore, I'll use the angular velocity of rotation of the earth, $\omega = 7.29 \times 10^{-5}$ rad/sec, and for the equatorial radius of the earth, $R = 6.378 \times 10^{6}$ meters.

In this case the rest frame is a local non-rotating frame, with axes pointing toward the fixed stars, but with origin at the center of the earth. The moving frame is a reference frame extending over a small portion of the rotating earth's surface, having velocity $v = \omega r$ relative to the rest frame, where r is the distance of the clocks from the rotation axis.

Now imagine two clocks fixed a small east-west distance x apart on the equator of the earth. Viewed from the nonrotating frame they will be moving with approximately equal speeds $v = \omega r$. If a clock synchronization process involving electromagnetic signals were carried out by two earth-fixed observers using Einstein synchronization in the moving frame, then the two clocks would not be synchronous when viewed from the nonrotating frame. The magnitude of the discrepancy is $vx/c^2 = \omega r x/c^2 = (2\omega/c^2)(rx/2)$. If this synchronization process is performed successively all the way around the circle, then effectively the distance x is $x = 2\pi r$, and the time discrepancy is thus

$$\Delta t = 2\omega/c^2 \times \pi r^2, \tag{9}$$

where πr^2 is the area enclosed by the path followed during the synchronization process. For example, synchronization around the earth's equator involves a discrepancy

$$\Delta t = \frac{2\omega}{c^2}\pi R^2 \approx 208 \text{ ns.} \qquad (10$$

upon arriving back at the starting point.

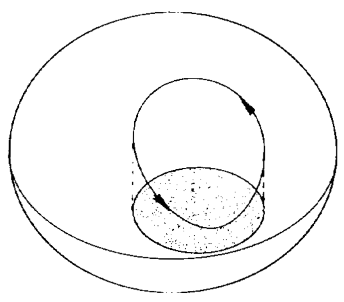
This effect is known as the *Sagnac effect*. If the synchronization path were westward around the earth rather than eastward, then the discrepancy would be of opposite sign. This means that Einstein synchronization in a rotating reference frame is not self-consistent: If A is synchronized with B and B is synchronized with C, then A is not necessary synchronized with C. In order to avoid difficulties with such non-transitivity it is best to adopt time in the *non-rotating* frame as the measure of time in the rotating frame. Thus one discards Einstein synchronization in the rotating frame.

To put it another way, if Einstein synchronization is used in the earth-fixed rotating frame, then it is necessary to apply a 'Sagnac correction' to the readings of clocks on the rotating earth, in order the obtain an internally consistent 'coordinate time' on earth's surface.

This is illustrated in Fig. 5, where there is a sketch of a flattened rotating earth. For a sequence of synchronization processes forming a closed circuit on the rotating earth, upon projecting the path onto the equatorial plane of the earth one can determine the projected area A_E. The Consultative Committee for the Definition of the Second and the International Radio Consultative Committee have agreed that, in order to obtain consistently synchronized clocks on the earth's surface at the subnanosecond level, the correction term to be applied is of the form

$$\Delta t = 2\omega/c^2 \times A_E, \tag{11}$$

where A_E is the projected area on the earth's equatorial plane swept out by the vector whose tail is at the center of the earth and whose head is at the position of the electromagnetic signal pulse. The area A_E is taken as positive if the head of the vector moves in the eastward direction. If two clocks located on the earth's surface are compared by using electromagnetic signals in the rotating frame of the earth, then Δt must be subtracted from the measured time difference (east clock minus west clock) in order to synchronize the clocks so they will measure coordinate time on the rotating earth. They will effectively measure time in the local non-rotating frame attached to the earth's center.



Sagnac Correction: $\Delta t = \dfrac{2\omega A_E}{c^2}$

Fig. 5. Projected area for a sequence of Einstein synchronization processed forming a closed circuit on the rotating earth's surface.

Lack of transitivity in synchronization has implications for devices which rely on accurate synchronization. Suppose a communications network distributes synchronization through a series of nodes, along two different paths, to the ends of a communication link as in Fig. 6. If the area enclosed by the path, projected onto the earth's equatorial plane, is not zero, then problems with inconsistent synchronization can arise. For example, suppose one synchronization link goes from San Francisco directly to New York, while a second link goes from San Francisco to Miami and then to New York. The discrepancy in synchronization between these two paths due to the Sagnac effect is about 11 ns. While this is not significant if the signal is 60 Hz as in a power grid, in an optical communications network operating at 10^{15} Hz the discrepancy amounts to 10^7 cycles of oscillation. Depending on the design of the system this may become significant in the future.

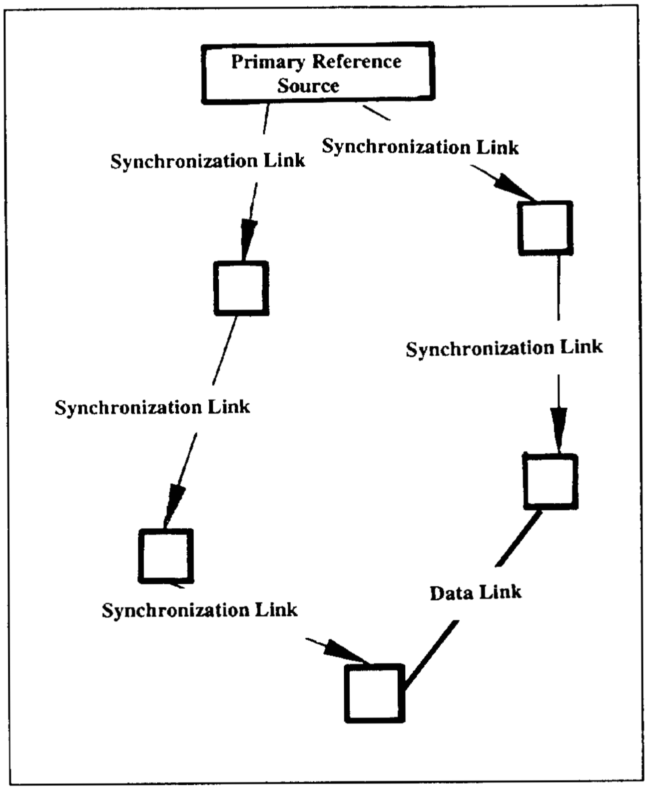


Fig. 6. Distribution of synchronization for a communications network.

Furthermore, if the trouble is taken to incorporate hardware delays to compensate for the Sagnac effect while sending in one direction, then when the communications are sent back the other way over the same link the effect will become twice as big. The effect is asymmetric. The same effect will occur in optical fiber communications networks where the speed of signal propagation may be significantly less than c. In the rotating reference frame the Sagnac effect is a property of space and time, not dependent on signal propagation speed.

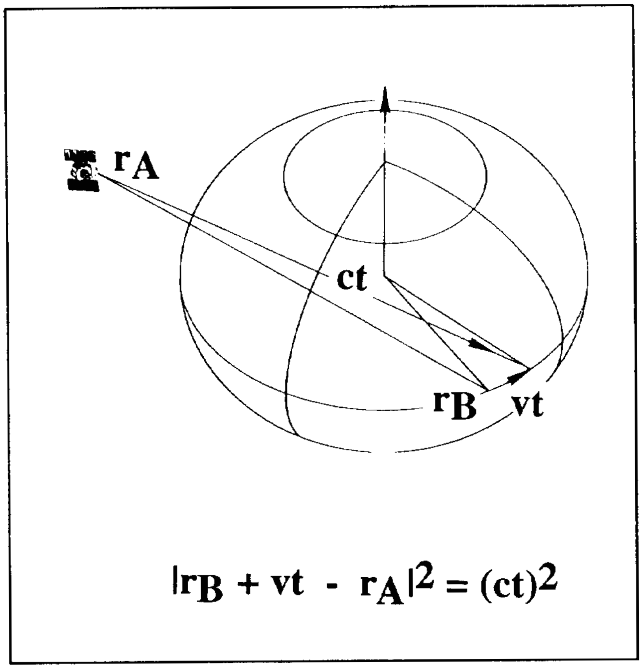


Fig. 7. The Sagnac effect will be automatically included if receiver motion due to earth rotation during signal propagation is accounted for.

An equivalent way of looking at this phenomenon is diagrammed in Fig. 7, which shows a signal transmitted from a satellite to a ground-based receiver. From the point of view of the nonrotating frame, the signal goes in a straight line with speed c, from the initial transmitter position $\mathbf{r}_A$ to the final receiver position. If in this frame one accounts for the motion of the receiver during the propagation of the signal, then the Sagnac effect will be automatically accounted for. This if the initial position of the the receiver is $\mathbf{r}_B$, the velocity of the receiver is $\mathbf{v}$, and the signal propagation time is t, constancy of c requires

$$|\mathbf{r}_B + \mathbf{v}t - \mathbf{r}_A| = (ct)^2. \qquad (12)$$

Iterative solution of Eq. (12) for t is equivalent to calculating the Sagnac correction.

Time Dilation

In the previous section I discussed two effects which are of first order in the velocity–the breakdown of simultaneity, and the Sagnac effect. In this section I shall discuss another famous effect–time dilation–which is of second order in the velocity. Imagine two inertial frames, a 'rest' frame or laboratory frame, and a moving frame. A clock in the moving frame beats more slowly than clocks in the rest frame with which it is successively compared. The following thought experiment will readily convince anyone that the principle of the constancy of the speed of light requires the 'moving' clock to beat more slowly. A prime denotes quantities measured in the moving frame.

Suppose that observers in the two inertial frames each possess a set of rectangular Cartesian coordinate axes which they orient so that the x, x' and y, y' axes are parallel. The direction of relative motion is parallel to the x, x' axes. The moving observer orients a rod of length L' along the y' axis, and sends a light signal up along this rod from one end to the other. The situation is diagrammed in Fig. 8. To simplify the discussion one assumes that the light starts out at the instant the origins of the two reference frames pass by each other.

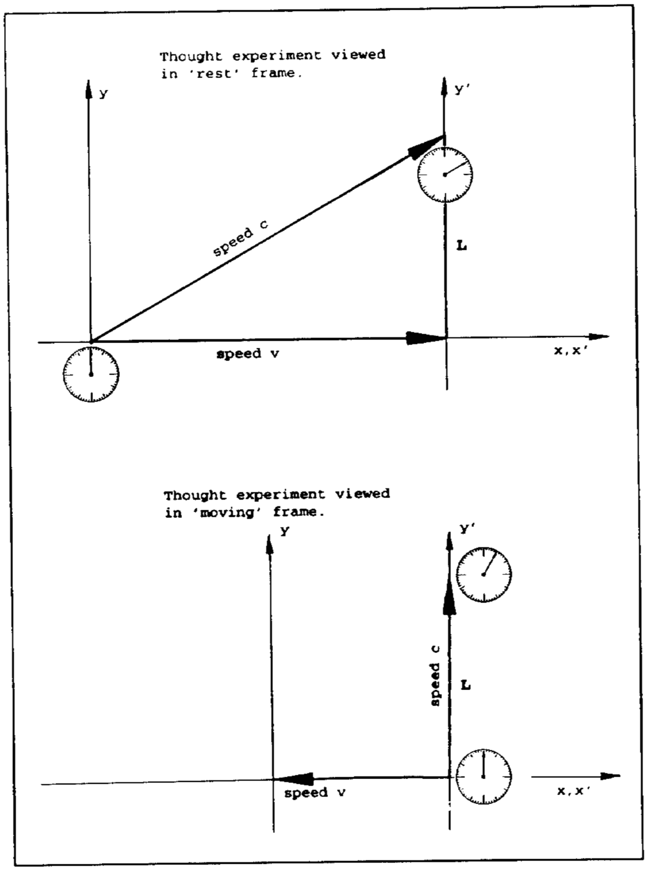


Fig. 8. Thought experiment showing that 'moving' clocks beat more slowly than clocks that remain 'at rest'.

The lower part of Fig. 8 shows the situation from the point of view of observers in the moving frame. The time t' required for light to travel along the rod is simply

$$t' = L/c. \tag{13}$$

The clock faces on the lower part of Fig. 8 indicate time at the beginning and end of the experiment.

The upper part of Fig. 8 shows the experiment from the point of view of observers in the rest frame. Breakdown of simultaneity would create difficulties for measurements of lengths oriented parallel to the relative velocity. But since this rod is oriented perpendicular to the relative velocity, by symmetry it is not possible for the rod to appear changed in length. So this rod has length $L = L'$ as it moves through the rest frame. The rod is moving to the right with speed v and the light travels along the rod, so there has to be a horizontal component of velocity of the light equal to v. The vertical component of the velocity of the light certainly has to be less than c; therefore the time required for the light to reach to upper end of the rod certainly has to be greater than L/c. This argument shows qualitatively that the clocks in the moving frame will beat more slowly than the sequence of clocks with which they are compared in the rest frame.

The top part of Fig. 8 actually gives the right answer, for by the principle of the constancy of the speed of light, the vertical component of the light velocity in the rest frame is just $\sqrt{c^2 - v^2}$. Thus for observers in the rest frame, the time t required for the light to reach the upper end of the rod is just

$$t = L/\sqrt{c^2 - v^2}, \tag{14}$$

so the relationship between t' and t obtained by eliminating L from Eqs. (13) and (14) and $L = L'$ is:

$$t' = \sqrt{1 - v^2/c^2}\,t. \tag{15}$$

Usually the ratio v/c is small, so the square root can be expanded, giving approximately

$$t' \approx \left(1 - \frac{1}{2}\frac{v^2}{c^2}\right) t. \tag{16}$$

The fractional slowing is given by the correction $v^2/2c^2$ in the above equation. This correction is also commonly called the second-order Doppler shift, or transverse Doppler shift.

Some examples of the size of this effect are as follows. For a clock at rest on the earth's equator, and viewed from the nonrotating frame,

$$-\frac{1}{2}\frac{v^2}{c^2} \approx -1.2 \times 10^{-12}; \tag{17}$$

this would accumulate to about 104 ns in one day. For a clock in a satellite orbiting the earth at 100 km altitude,

$$-\frac{1}{2}\frac{v^2}{c^2} \approx -3.4 \times 10^{-10}. \tag{18}$$

For a clock in a GPS satellite,

$$-\frac{1}{2}\frac{v^2}{c^2} \approx -8.34 \times 10^{-11}. \tag{19}$$

Keeping in mind that in the future the fractional frequency stability of orbiting clocks may approach a part in 10^{15}, these are very large effects. Even for clocks of frequency stability 1×10^{-13}, as in the GPS Block II satellites, the second-order Doppler shift for an earth-fixed clock is significant.

Gravitational Frequency Shifts

The Sagnac effect and the second-order Doppler shift are effects which can be understood on the basis of the Special Theory of Relativity. A third effect–the gravitational frequency shift–occurs when signals are sent from one location to another having a different gravitational potential. The effect can be understood in an elementary way using the fundamental assumption of the General Theory of Relativity–Einstein's Principle of Equivalence.[2]

The Principle of Equivalence

Einstein's Equivalence Principle states that over a small region of space and time, a fictitious gravitational field induced by acceleration cannot be distinguished from a gravitational field produced by mass. Thus the fictitious centrifugal force one feels in turning a corner in a vehicle has the same physical effects as a real gravitational field. An immediate consequence of the Equivalence Principle is that gravitational fields can be reduced to zero by transforming to a freely falling reference frame. The fictitious gravitational field due to the acceleration then exactly cancels the real gravitational field.

All experiments performed in a real gravitational field, such as in a laboratory on the surface of the earth where there is a gravitational field **g**, will have the same results at experiments performed in a laboratory in free space which is accelerated in the opposite direction with acceleration $\mathbf{a} = -\mathbf{g}$. In Fig. 9a are sketched some experiments performed in a laboratory fixed on the earth's surface. For example two objects of different compositions are observed to fall downward with equal accelerations g. (This is related to the deep experimental fact of the strict proportionality of inertial and gravitational mass, a subject we shall not go into here.[3]) In Fig. 9b, a similar experiment is performed in a laboratory in free space which is being pulled upward with acceleration g. A non-accelerated observer sees that the apple and the lead ball have no forces exerted on them so remain at rest with respect to each other and the laboratory is accelerated past the objects, whereas the observer in the accelerated frame sees the objects 'fall' downward with identical accelerations g.

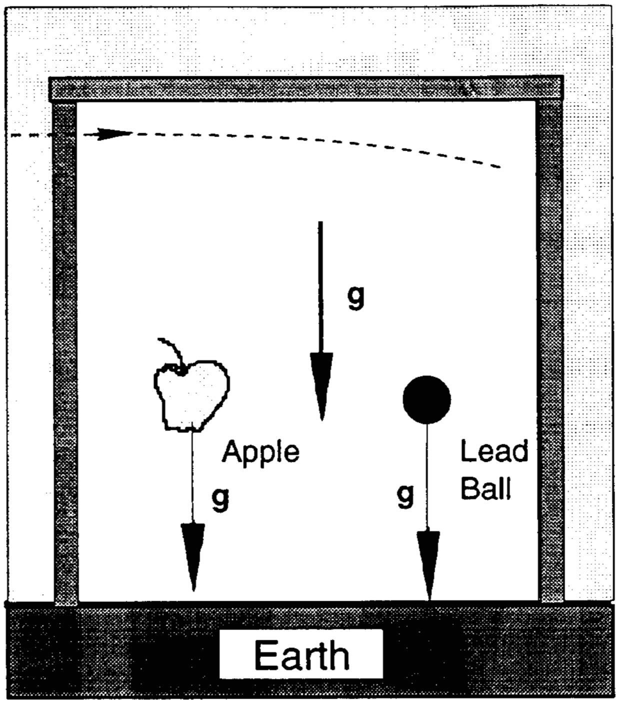


Fig. 9a. All objects fall with equal accelerations in a laboratory near the earth's surface.

The equivalence of the two laboratories implies that a beam of light is deflected toward the source of the gravitational field. Let a beam of light–which travels in a straight line in free space–enter the side of the accelerated laboratory (near the top, in Fig. 9b). The observer in this laboratory is accelerated past the light, so it must appear to fall down just as do the massive objects. The experiment must have the same outcome in the non-accelerated laboratory on earth, so to an observer in a real gravitational field light must fall down. A beam of light passing near any massive body will be deflected towards the body.

Time Delay

If one imagines the wavefronts in a beam of light as the beam is deflected toward the massive source of a gravitational field, then one can picture the portions of the wavefront nearest the mass being slowed down slightly with respect to the portions of the wavefront farther away from the source. The wavefront then tilts over and the beam is thereby deflected. This means that of two beams of light passing near a massive source, the one which passes closer will take longer to get by. Thus not only is light deflected, it is slowed down by a gravitational field.

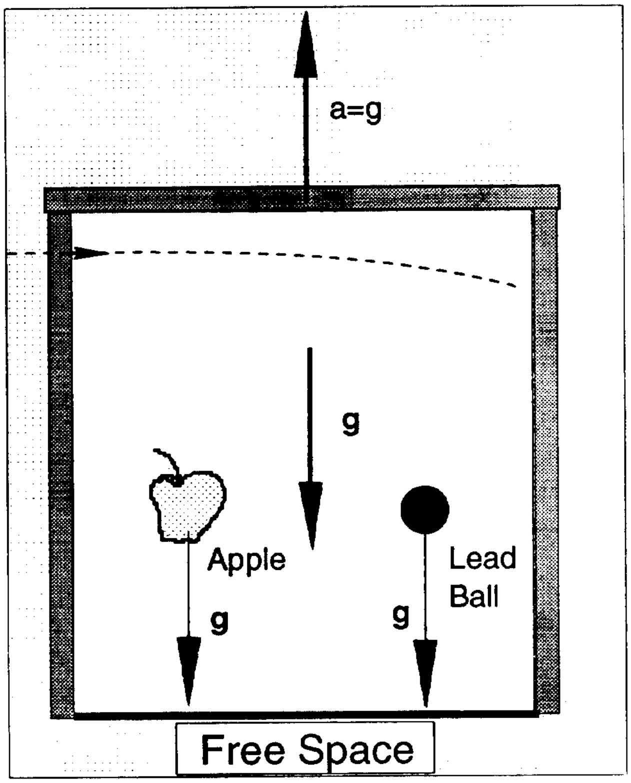


Fig. 9b. By the Equivalence Principle, experiments performed in an acclerated lab in free space have the same outcomes.

Time delays of signals in the neighborhood of the earth can be a few tenths of a nanosecond. Such time delays are determined by a complicated logarithmic function of signal path parameters, times the quantity $4GM_E/c^3$, where G is the Newtonian gravitational constant and M_E the earth's mass. For earth $GM_E/c^2 = 0.443$ cm, so the scale of such effects near earth is

$$\frac{4}{c}\frac{GM_E}{c^2} \approx \frac{1.77 \text{ cm}}{c} = .06 \text{ ns.} \qquad (20$$

This is not enough to worry about at the present time but could be significant in the future–a timing error of .1 ns in a navigational system would give rise to a 3 cm error in position.

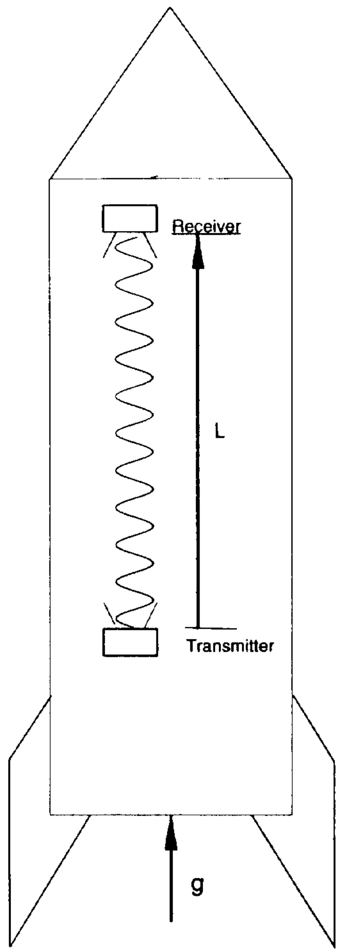


Fig. 10. A signal travelling upwards in a gravitational field is shifted towards lower frequencies.

Gravitational Frequency Shifts

It follows from the Equivalence Principle that an electromagnetic signal passing upwards in a gravitational field will be redshifted. In Fig. 10 is a sketch of an experiment performed in an equivalent laboratory, a rocket having acceleration g upwards in free space. Imagine the situation from the point of view of a non-accelerated frame. Suppose a signal leaves the accelerated transmitter at the initial instant, when the transmitter velocity is still zero. The signal upwards a distance L, and is received by the accelerated receiver. The time required for the signal to propagate from transmitter to receiver is:

$$t = L/c. \qquad (21)$$

During this time, the receiver has picked up a velocity

$$v = gt = gL/c. \qquad (22)$$

To the receiver, the signal appears to come from a receding source and is Doppler shifted. To a first approximation the fractional frequency shift is $\Delta f/f = -v/c$; therefore the fractional frequency shift in the 'effective' gravitational field g is

$$\frac{\Delta f}{f} = -\frac{v}{c} = -\frac{gL}{c^2}. \tag{23}$$

The quantity gL can be interpreted as the change in gravitational potential, $\Delta\phi$, of the signal.

At the surface of the earth,

$$g/c^2 = 1.09 \times 10^{-13} \text{ per km}, \tag{24}$$

which is very important for today's time standards. For example a signal of definite frequency originating at mean sea level would be redshifted by 1.79 parts in 10^{13} upon arriving at the altitude of the NIST frequency standards laboratory in Boulder, CO. Consequently the contribution of the NIST time standard to Universal Coordinated Time (UTC) requires that a paper correction of -15.5 ns/day be applied to the NIST clock before it can be compared to time standards at mean sea level.

For a clock in a satellite orbiting the earth at 100 km altitude compared to one on the geoid,

$$\frac{\Delta\phi}{c^2} = 1.08 \times 10^{-11}. \tag{25}$$

Not only will these effects be large in the future when clock stabilities approach a part in 10^{15} or better, it will be necessary to compute them quite accurately. This will mean, for example, that there will be a need for improved precision of the ephemerides of clock-carrying satellites.

The Global Positioning System

The best existing example of an engineering system in which relativity plays an essential role is the GPS. This consists of a constellation of perhaps 24 earth-orbiting satellites carrying atomic clocks which synchronously transmit navigation signals, much as described in the discussion of Fig. 2. The satellite orbits are at approximately 20,200 km altitude. Therefore clocks in the satellites will be significantly blueshifted in rate, compared to clocks on the ground. The second-order Doppler shift of such clocks was given in Eq. (20). Also, if the orbits are not perfectly circular (and they almost never are), the clocks' yo-yo motions towards and away from the earth will generate periodic additional gravitational frequency shifts, and second-order Doppler shifts. Further, observers on the ground who wish to make use of the navigational signals will experience the Sagnac effect due to earth's rotation.

A complete discussion of all the significant relativistic effects, with analytical expressions for the necessary corrections, can be found elsewhere.[4] Here I shall just indicate roughly the magnitudes of some of the corrections.

First, consider ground-based clocks in receivers which are at rest on the earth's surface. Standard clocks on the geoid are used to define the unit of time; however, from the point of view of a local, nonrotating frame, there is a frequency shift due to earth's mass; the fractional frequency shift is about -7×10^{-10}. The earth's oblateness is associated with a quadrupole contribution to the gravitational potential which cannot be neglected; the fractional frequency shift is about -4×10^{-13}. If earth-based clocks are not on the geoid they suffer a gravitational frequency shift (see Eqs. (23-24)). Finally there is a second-order Doppler shift due to the earth's rotation; the fractional frequency shift from this effect can be as large as -1.2×10^{-12} (see Eq. 17).

For GPS receivers in motion relative to the earth's surface, there is an additional second-order Doppler shift due to their speed with respect to the ground; this can be of the order of 10^{-12} depending on the ground speed. Also, the Sagnac effect–or motion of the receiver during propagation of the navigation signal may give rise to effects of several hundred nanoseconds magnitude.

The transmitters themselves suffer a frequency shift due to the earth's gravitational potential, and a second-order Doppler shift due to orbital motion; these effects are several parts in 10^{10}. The additional frequency shifts due to orbital eccentricities can be tens of nanoseconds; for a GPS satellite of eccentricity $e = .01$, the maximum size of the effect is about 23 ns.

Propagation of signals from transmitter to receivers are subject to the Sagnac effect, involving relativistic corrections of up to several hundred nanoseconds. Relativistic time delay of signals or relativistic deflection of signals is a few tenths of nanoseconds and is currently neglected in the GPS.

The Concept of Coordinate Time

With so many significant relativistic effects occurring on earth-fixed and earth-orbiting clocks, the problem of synchronization of the clocks becomes an acute one. Rates are affected by motional and gravitational effects; synchronization on the spinning earth is inconsistent if the Einstein procedure is used. How is it possible to synchronize a network of distributed, rapidly moving clocks so that a navigational system will work as conceived in Fig. 2? What has been found to work extremely well in the GPS is to use the time in a hypothetical underlying local inertial frame, with origin attached to the earth but not spinning, as the measure of time. This time is not time on any standard clock orbiting the earth; instead one makes use of general relativity to correct the readings of such clocks so they would agree with hypothetical clocks at rest in the local inertial frame. The time obtained by so correcting all the clocks in the system, is an example of *coordinate time*.

Thus, imagine an underlying nonrotating frame, or local inertial frame, unattached to the spinning earth, with with its origin at the center of the earth. This frame is sometimes called the "Earth-Centered Inertial" frame, or ECI frame. In this frame, introduce a fictitious set of standard clocks available anywhere, all synchronized via the Einstein procedure, and running at agreed upon rates such that synchronization is maintained. Gravitional effects are incorporated by choosing one clock as a Master Clock and requiring that all other clocks be syntonized to the Master clock by simple transmission of signals without any frequency shift corrections. The resulting time scale is called coordinate time.

Now introduce a set of standard clocks distributed around the surface of the rotating earth, or orbiting the earth. To each one of these standard clocks a set of systematic corrections is applied, so that at each instant the standard clock as corrected agrees with the time on a fictitious standard clock, at rest in the ECI frame, with which it instantaneously coincides. The set of corrected standard clocks will therefore be keeping coordinate time. In other words, coordinate time is equivalent to time measured by standard clocks in the ECI frame.

Time measured on coordinate clocks has two highly desirable properties. First, synchronization is reflexive: if A is synchronized with B, then B is synchronized with A. Second, synchronization is transitive: if A is synchronized with B, and B is synchronized with C, then A is synchronized with C. Internal inconsistencies are therely eliminated.

GPS time is an example of coordinate time. To an observer on the earth's geoid, a standard clock in a GPS satellite in a nominally circular orbit would appear to be blueshifted by .4465 parts per billion, or about 39,000 ns per day; this is a net effect of gravitational frequency shifts and motional Doppler shifts of satellite clocks relative to reference clocks fixed on the ground. To compensate for this, the 10.23 MHz reference frequency of satellite clocks is adjusted downward to 10.229 999 995 43 MHz. The adjustment is accomplished on the ground before the satellites are launched.

Also, if the orbit of the satellite clock is not perfectly circular, there will be additional gravitational and motional rate shifts which have to be accounted for. The additional correction required to achieve synchronization when the orbit eccentricity is not zero is given by the expression[4]

$$\Delta t = +4.428 \times 10^{-10} e\sqrt{a}\,\sin E \text{ sec}, \qquad (26)$$

where a is the semi-major axis in meters and E is the eccentric anomaly. Usually the software in the user's receiver makes this correction.

Application of Satellite Navigation in Geodesy

The motivation to obtain accurate measurements of movements of the earth's crustal plates is intense. Knowledge of these very slow motions is crucial to the development of improved earthquake prediction capability; the potential impact on construction codes, building restrictions, etc., is considerable. In recent years the GPS has been successfully used to measure very long baselines between fiducial points on different crustal plates by a method described as "carrier phase double difference." Two receivers are placed at the ends of a baseline of interest, and signals from two satellites are then "double differenced" in a manner to be described below. Differenceing removes the need for some systematic corrections but as will be seen, there are residual relativistic effects which must be accounted for.

Referring to Fig. 6 and the propagation time t given in Eq. (12), let the satellite position at the instant of transmission t_S be denoted by $\mathbf{r}_S$ and the receiver or observer position at the same instant be denoted by $\mathbf{r}_O$. Let the coordinate time of arrival of the signal at the observer be denoted by t_O. Then solving Eq. (12) for the propagation time gives

$$t_O = t_S + K + \frac{|\mathbf{r}_S - \mathbf{r}_O|}{c} + \frac{2\omega}{c^2} \cdot \left[\frac{1}{2}\mathbf{r}_S \times \mathbf{r}_O\right]. \qquad (27)$$

The last term is the Sagnac correction and K represents a possible time offset or error of the receiver's clock. The rate adjustment applied to satellite clocks means that the quantity t_S will have the correct scale when received on the geoid. There is a further correction, from the non-circular motion of the satellite, given by Eq. (26). Thus when all relativistic effects are incorporated,

$$
\begin{aligned}
t_O = t_S + K &+ \frac{|\mathbf{r}_S - \mathbf{r}_O|}{c} \\
&+ \frac{2\omega}{c^2} \cdot \left[\frac{1}{2}\mathbf{r}_S \times \mathbf{r}_O\right] \\
&+ 4.428 \times 10^{-10} e\sqrt{a}\,\sin E.
\end{aligned}
\tag{28}
$$

Let subscripts 1 and 2 denote the two different satellites and the two different observers. Suppose there are receivers at two different positions which receive a time signal originating from a single satellite. Upon taking the first differece of the arrival times, it is immediately seen that the eccentricity term cancels out, leaving the expression:

$$
\begin{aligned}
t_{O2} - t_{O1} = K_2 - K_1 &+ \frac{|\mathbf{r}_S - \mathbf{r}_{O2}|}{c} - \frac{|\mathbf{r}_S - \mathbf{r}_{O1}|}{c} \\
&+ \frac{2\omega}{c^2} \cdot \left[\frac{1}{2}\mathbf{r}_S \times (\mathbf{r}_{O2} - \mathbf{r}_{O1})\right].
\end{aligned}
\tag{29}
$$

The Sagnac correction is still needed. The time of transmission of the signal, t_S, cancels out which lessens the impact of selective availability.

Now the same set of measurements is taken, at essentially the same time, using a second satellite. Writing another equation similar to Eq. (29) for the second satellite and taking the difference, it can immediately be seen that even the clock offsets in the receivers cancel out, leaving only the usual propagation delay terms with relativistic corrections due to the rotation of the earth:

$$
\begin{aligned}
(t_{O2} - t_{O1})_{S2} &- (t_{O2} - t_{O1})_{S1} \\
&= \frac{|\mathbf{r}_{S2} - \mathbf{r}_{O2}|}{c} - \frac{|\mathbf{r}_{S2} - \mathbf{r}_{O1}|}{c} \\
&- \frac{|\mathbf{r}_{S1} - \mathbf{r}_{O2}|}{c} + \frac{|\mathbf{r}_{S1} - \mathbf{r}_{O1}|}{c} \\
&+ \frac{2\omega}{c^2} \cdot \left[\frac{1}{2}(\mathbf{r}_{S2} - \mathbf{r}_{S1}) \times (\mathbf{r}_{O2} - \mathbf{r}_{O1})\right].
\end{aligned}
\tag{30}
$$

The Sagnac correction is still necessary. In this application the correction is largest when the baseline is at right angles to the line between the satellites; it can be several hundred nanoseconds.

In Fig. 11 are plotted some baseline measurement data taken repeatedly on baselines in the Southwest Pacific, of lengths up to 2500 km.[5] Only the length of the baseline is shown here. The vertical scatter in the plotted points gives a measure of the errors involved. For the 2500 km baseline the spread is only a few cm.

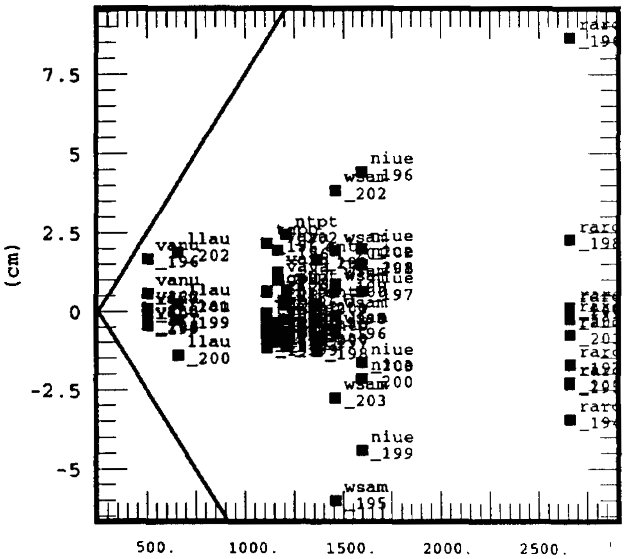


Fig. 11. Scatter in baseline lengths for several different baselines measured during the Southwest Pacific 1992 GPS Campaign. The data were provided by UNAVCO.

Impact on Fundamental Metrology

The previous sections have been devoted almost exclusively to the impact of relativity on the measurements of time, with distance derived by multiplying by c. At the level of a centimeter or less, there are additional effects on the measurement of position which arise because space in the neighborhood of a massive body is distorted. Consider as in Fig. 12 an attempt to establish a system of spatial coordinates in the neighborhood of earth, against which to measure the positions of the earth's crustal plates. Suppose that we wish to measure angles in the usual Euclidean way, so that a circle of coordinate radius r centered on the earth would have a circumference $2\pi r$, measured with standard rods or with the help of the constancy of c. Two such circles, of coordinate radii r_1 and r_2 are indicated in Fig. 12. The standard distance from the inner circle straight out along a radius to the outer circle is not $r_2 - r_1$; instead one finds the standard distance d is

$$
d = r_2 - r_2 - \frac{2GM_E}{c^2}\ln(r_2/r_1). \tag{31}
$$

The correction due to space curvature is of the order of 1 cm.

More generally, the fact that c has a defined numerical value means that the physical unit of length depends on the clock used to define the unit of time. For example, in Barycentric Dynamical Time (TDB), the unit of time is the same as that of clocks on earth, in orbit around the sun, and the point of view taken is that of an observer in a reference frame at rest with respect to the solar system barycenter. The clocks on earth beat more slowly than clocks at rest at infinity in this system by the factor[6]

$$1 - L = 1 - 1.55 \times 10^{-8}. \qquad (32)$$

Therefore, the meter is physically longer, so the length of a physical object is numerically smaller by this factor. The mass of the earth can be used to construct a quantity having the physical dimensions of a length, namely GM_E/c^2. However c has a defined value; this means that in TDB coordinates, GM_E is numerically smaller than in SI units:

$$(GM_E)_{TDB} = (1 - L)(GM_E)_{SI}. \qquad (33)$$

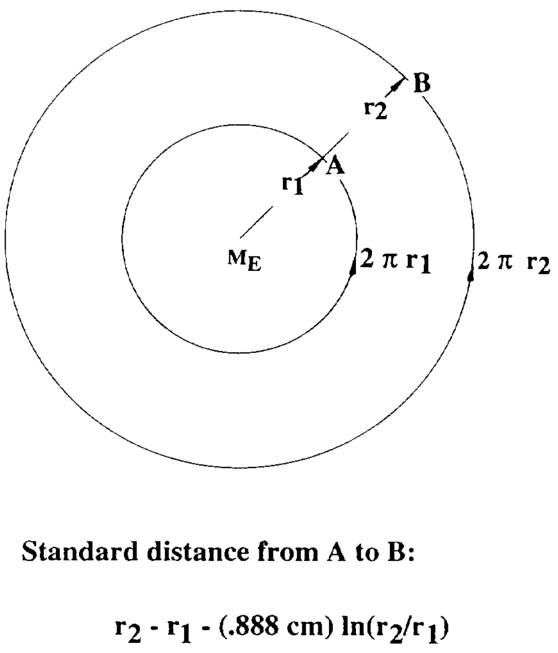


Fig. 12. Effect of spatial curvature on standard distance measurements.

Some Remarkable Cancellations

So far in this discussion I have ignored the fact that the earth is actually an oblate ellipsoid; clocks near one pole will be closer to the center of the earth and will therefore be subject to a gravitational redshift; on the other hand in the ECI frame such clocks are moving more slowly than clocks near the equator and are subject to less second-order Doppler shift. This is diagrammed in Fig. 13. Over the ages the earth's surface has assumed the approximate shape of a hydrostatic equipotential in the rotating frame–the average shape of the ocean's surface defines the geoid. it is a remarkable fact that on the geoid, there is a very precise cancellation of gravitational frequency shifts and motional Doppler shifts, so that all clocks at rest on the geoid beat at the same rate! Therefore it is possible to construct a network of standard clocks on the earth's geoid, all beating at the same rate. However, to synchronize these clocks consistently it is necessary to correct for the Sagnac effect, due to the earth's rotation.

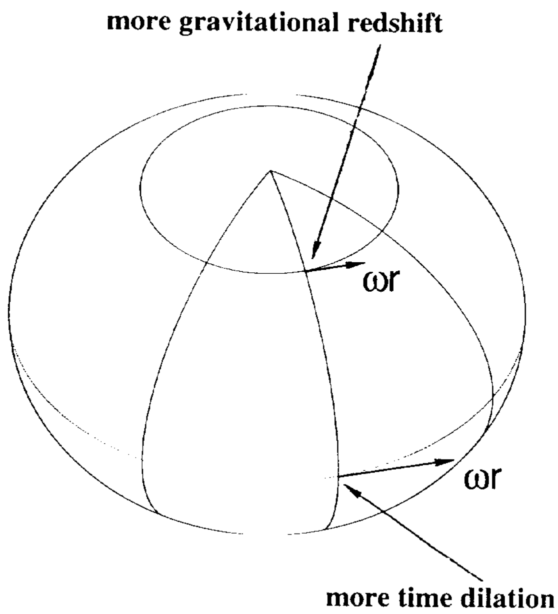


Fig. 13. On the oblate rotating earth's geoid, changes in gravitation frequency shift are precisely compensated by second-order Doppler shifts.

So far I have also ignored the possibility that the sun, moon, or other planets might contribute to gravitational frequency shifts. Also, the earth's orbit is not perfectly circular so one might expect a yo-yo effect on the rates of earth-orbiting clocks somewhat analagous to the correction given in Eq. (25) for GPS clocks. For example, when a satellite is in earth's shadow its clock should be gravitationally blueshifted as compared to a satellite-borne clock between the sun and earth. For such a configuration, the fractional frequency shift between clocks in the two satellites, due to the sun, is about three parts in a trillion, which in an hour would cause a 12 ns timing error to build up. Fortunately we do not have to worry about this! This effect is cancelled to high precision by other relativistic effects arising because the entire system of earth plus satellites is in free fall around the sun. By the principle of equivalence, we should not be surprised that for a system in free fall, the gravitational effects of the sun are transformed away. Detailed analysis of this situation is rather delicate; when comparing clocks in the ECI frame–which is falling around the sun–with clocks in the solar system center-of-mass frame, there is disagreement about the meaning of simultaneity in the two frames. Using coordinate time in the ECI frame, with clocks synchronized by the Einstein procedure (modified by gravitational effects), the gravitational effects due to other solar system bodies will cancel to high accuracy. The residual gravitational effects are due to tidal potentials only, and are less than a part in 10^{16}.

Conclusion

In this paper, numerous examples of relativistic effects which are important for current and future navigation, timing, and communications systems have been discussed. Relativistic effects are always systematic, but depend on knowledge of the positions and velocities of the various clocks in the given reference frame. These effects are not noise; they are well-understood, and can be corrected for to a high level of accuracy. As clock stability and accuracy continues to improve it will become increasingly important for system designers and practitioners to become familiar with these effects so they will be accounted for properly. I hope this paper helps in a small way to educate those for whom the mathematical apparatus of general relativity is excessively cumbersome.

References

[1] R. F. C. Vessot, E. M. Mattison, M. W. Levine, and R. L. Walsworth, "Status of local oscillators for operating ultra-high resolution frequency discriminators as frequency standards," 24^{th} Annual PTTI Applications and Planning Meeting, McLean, VA, December 1-3, 1992.

[2] A. Einstein, *The Meaning of Relativity*, 3^{rd} ed., Princeton Univ. Press, Princeton, NJ, (1950).

[3] C. W. Misner, K. S. Thorne, J. A. Wheeler, *Gravitation*, W. H. Freeman & Co., San Francisco, CA (1993), p. 13 ff.

[4] N. Ashby, "A Tutorial on Relativistic Effects in the Global Positioning System," NIST Contract No. 40 RANB9B8112, Final Report, Feb. 1990. Copies of this report are available from the author on request.

[5] John Braun, UNAVCO, private communication.

[6] R. W. Hellings, "Relativistic Effects in Astonomical Timing Measurements," *Astron J.* **91**, 650 (1986).