Coordinated Universal Time and the Leap Second

Judah Levine
Time and Frequency Division and JILA
NIST and the University of Colorado
Boulder, Colorado 80305 USA
E-mail: Judah.Levine@nist.gov

Abstract

I will discuss the considerations that were important in the design of the current version of Coordinated Universal Time. The design includes the addition of additional “leap” seconds to keep Coordinated Universal time within ±0.9 s of the UT1 time scale, which is a proxy for the position of the Earth in space. I will describe the advantages and problems associated with the leap-second system, and a number of changes that have been proposed to the realization of the time scale.

1. Introduction

Coordinated Universal Time, generally abbreviated as UTC, is the basis for civil time and frequency in almost all countries. There have been a number of different definitions of the relationship between UTC and International Atomic Time (TAI), but I will discuss only the current realization of UTC, which dates from 1972.

The single UTC time scale is used to realize the definitions of three related quantities: time, time interval, and frequency. It is difficult to design a time scale that can satisfy the requirements of the different applications that depend on these three quantities, and this difficulty is the root of current discussions to change the realization of UTC. In order to appreciate this difficulty, I will present a short historical discussion of the older definitions of time and frequency, because the current definition is an extension of the older definitions of time and time interval.

2. Definitions of Time, Time Interval, and Frequency

The times of events in antiquity were derived from astronomical observations of the positions of the sun, the moon, or a star. The day was defined by local sunrise or sunset, the month by the observation of a new moon, and the year by the spring equinox or by the first observation of a particular star just before sunrise.

Even in antiquity, it was known that the different time scales were not commensurate: that there are not an integral number of solar days in a lunar month or in a solar year. Each society devised a method for dealing with this problem, and the resulting calendars were often quite complex. I will not discuss this complexity, because it is not important for the current definition of UTC, which is based on the length of the second and considers only integer multiples and fractions of this base quantity. For example, a UTC day is exactly 86 400 (24 × 60 × 60) UTC seconds long.

A time interval is the elapsed time between two consecutive “standard” astronomical events. A clock is simply a device that acts as an interpolator. It facilitates the measurement of time intervals between the standard astronomical events.

All clocks comprise two systems: a device that generates periodic events or “ticks,” and a method for counting the ticks to display the elapsed time interval since an origin that is unique to each time scale. When time and time interval were astronomically defined, the interval between the ticks of a clock – or, equivalently the frequency of the tick generator – had to be adjusted so that the time interval displayed by the clock agreed with the astronomical definition. Frequency was therefore a derived quantity that was implicitly defined by astronomical observations. It could not be independently defined. That situation was acceptable until the start of the 20th century, because there were few, if any, applications that depended on frequency and not on time.

3. The Problem with Apparent Solar Time

The interaction between the orbital motion of the Earth with respect to the sun and the spin of the Earth about...
its axis increases the length of the apparent solar day (the interval between two consecutive solar noons, for example) by an average of about four minutes (approximately $24 \times 60/365.25$) relative to the time it takes the spin of the Earth to make a complete 360° revolution with respect to the fixed stars, which is called sidereal time. The elliptical shape of the orbit of the Earth adds an annual variation to this effect. (Kepler’s second law, which is really a statement of conservation of angular momentum, requires that the orbital angular speed is greatest when the Earth is closest to the sun, and decreases as the sun-Earth distance increases.) Since frequency and time interval are derived quantities, this annual variation is pushed into them, as well.

4. Mean Solar Time

The annual variation in the length of the apparent solar day was known in antiquity, but it became more serious when applications that depended on frequency but not on time or time interval were developed. The frequencies assigned to radio stations in the early years of the 20th century are an example of the problem. Artifact frequency standards, initially based on precision inductors and capacitors, and later on quartz crystals, were developed to address the requirement for a standard frequency reference. They had inadequate long-term stability, and required periodic recalibration by astronomical observations. The annual variation in apparent solar time resulted in an unacceptable variation in the calibration of the frequency of these reference devices.

The first solution was to define mean solar time: a time scale based on the motion of a fictitious sun, moving along the equator at a constant speed that matched the average apparent motion of the real sun moving along the ecliptic. (It is common in astronomy to think of the Earth as stationary, with the sun in orbit around it.) The difference between the apparent and mean suns is the “equation of time,” a periodic function that has an amplitude of approximately 16 minutes. In addition to the variation resulting from conservation of angular momentum, the equation of time has a contribution produced by the periodic difference between the position of the real sun, which moves along the ecliptic, and the fictitious sun, which moves along the equator. This contribution is caused by the tilt of the axis of the Earth with respect to the plane of the apparent annual motion of the sun.

A contemporary realization of mean solar time is Greenwich Mean Time, which is defined in principle as mean solar time as observed on the Greenwich meridian. (In current practice, mean solar time is derived from sidereal time: the apparent motion of the Earth with respect to distant stars.)

5. Limitations of Mean Solar Time

There are some practical difficulties in constructing a clock that realizes mean solar time. There is no astronomical observation that directly realizes mean solar time, so that it is difficult to calibrate a clock without extensive observations. The two possibilities are combining sidereal time with the orbital motion of the Earth, or combining apparent solar time with the equation of time. The length of the sidereal day could be measured with an accuracy of a few milliseconds by observing the times of meridian transit of a bright star at a number of observatories. The observations at each observatory define the time scale UT0. These data are affected by polar motion: the precession and nutation of the axis of rotation of the Earth. The data from the different observatories could be combined to separate the effects of polar motion from the rotation rate of the Earth. This analysis yields the UT1 time scale, which is a proxy for the angle of the Earth in space corrected for polar motion. Contemporary determinations of UT1 and mean solar time are based on very-long-baseline interferometry (VLBI) observations of signals from very distant radio sources.

However, there are more serious problems. In the 1920s, it became apparent that the length of the UT1 day was not a constant as measured by the best pendulum clocks of that era. The length of the UT1 day had an annual variation that could be modeled and removed to produce a time scale called UT2. However, UT2 had a secular variation due to the irregular slow down in the angular velocity of the Earth about its axis of rotation. As I previously discussed, the secular variation in UT2 was pushed into time interval and frequency.

There were a number of changes in the astronomical definition that attempted to preserve time derived from astronomy as the fundamental unit, but none of them was completely successful. For example, one attempt to define the second was to use the length of the year 1900. Since the length of the year 1900 was not an observable for practical metrology, the practical standards of time, time interval, and frequency had to be based on some method of extrapolating the definition and realizing it in some contemporary observation or physical device. The extrapolated reference device became the de facto standard quantity, since there was no effective way of linking this reference back to the fundamental definition. (This result was not unique to time or time interval. The length of the meter was defined in principle in terms of a portion of the circumference of the Earth, but the length in practice was determined by the artifact standard that was constructed from the fundamental definition. The fundamental definition was irrelevant from the perspective of practical metrology.) However, a more fundamental change was already on the horizon.
6. Frequency and Quantum Mechanics

The considerations that I have discussed above were fundamentally changed by quantum mechanics. The frequency associated with a transition between quantum states was proportional to the energy difference between them. This energy difference could be calculated by combining the principles of quantum mechanics with the properties of atomic particles, such as the masses and charges of electrons and protons, the speed of light, and other similar quantities. The point was that frequency was a fundamental parameter from the quantum-mechanical perspective, and the frequency associated with an atomic transition was a fundamental invariant property of nature. The natural extension of this idea was to make frequency the fundamental unit, and have both time and time interval be derived quantities.

7. The Cesium Second

Although the quantum-mechanical principles that I discussed in the previous section are true for any atom in principle, various engineering considerations favored the use of a hyperfine transition in the ground state of cesium as the definition of frequency. (These considerations are less important today, and it is likely that a different atomic transition will be chosen as the definition of frequency in the foreseeable future. This change is unlikely to have a significant impact on the realizations of the standards of time and frequency.)

In order to make the transition between a system based on astronomical time as the fundamental definition to one based on the cesium frequency, Essen and Perry, at the National Physical Laboratory in the UK, and Markowitz, at the US Naval Observatory in Washington, USA, measured the frequency of the cesium transition in astronomical time units. After several years of observations ending in 1958, they concluded that the length of the second should be defined as 9 192 631 770 cycles of the cesium hyperfine transition in the ground state. This value was accepted as the definition of the length of the second in 1965. The intention was to minimize any discontinuity in the length of the second.

The length of the cesium second implicitly defined the length of the cesium day as 86 400 cesium seconds. The minute and the hour were implicitly defined in the same way as exact integer multiples of the cesium second. The lunar month and the solar year were now derived quantities that had to be measured in units of the cesium second.

From the start, it was clear that the value adopted for the length of the cesium second resulted in the length of the day that was too small relative to astronomical observations. The fractional frequency difference was about $3 \times 10^{-8}$, which produced a time dispersion of about 1 s/year. The source of the discrepancy was that the value used for the length of the UT1 day was based on relatively old observations, and the rotation rate of the Earth had slowed down by the time of the comparison experiments in the 1950s.

A somewhat larger value for the number of cycles in a cesium second would have solved the immediate problem. However, the irregularities in the rotation rate of the Earth combined with the secular increase in the length of the astronomical day guaranteed that no choice for the number of cesium cycles in a second could permanently remove the discrepancy, which would have secular and irregular variations no matter what value was chosen. The choice of frequency as the fundamental parameter was consistent with quantum mechanics and with the assumption that the frequencies of atomic transitions, which are calculable in principle in terms of the fundamental constants of nature, should be invariant. However, it did nothing to remove the variability in the astronomical time and time intervals. These quantities now had the secular and deterministic variations that were previously associated with frequency when time was the fundamental parameter.

8. The Conflict Between Time and Frequency

The definition of the length of the second in terms of the transition frequency in cesium, and the increasing availability of commercial cesium standards in the 1960s, divided the user community into three distinct groups:

Group 1 was the scientific community, which regarded frequency as a fundamental parameter derived from atomic properties by the use of quantum-mechanical principles. This group considered frequency (or wavelength) as conceptually equivalent to the other fundamental constants, such as the charge and mass of the electron and proton and similar parameters. Any difference between a time scale derived from the cesium frequency and a time scale related to astronomy could be handled by defining an offset parameter that would be published as needed and administratively applied. The offset could not be predicted algorithmically because of the irregular variation in the rotation rate of the Earth.

Group 2 was the astronomical community, which regarded time as a proxy for the angular position of the Earth in space. Although the offset between astronomical time and atomic time could be tabulated and administratively applied, there were applications that were designed with the premise that the offset would be a small quantity, so that the official UTC time scale could be used as a proxy for the orientation of the Earth in space without additional parameters or corrections.

Group 3 was the engineering community, which required a stable and easily constructed artifact frequency...
standard that could be used for accurate and relatively rapid calibrations. There was nothing wrong with a standard that was derived from the properties of atoms and the values of other fundamental constants, but that was a secondary consideration. In addition to its variability, astronomical time scales did not provide an easily accessible frequency standard that could be used for routine calibrations.

Based on current observations (in 2016), the difference between the length of the day defined astronomically and the length of a day defined by cesium seconds would diverge at a rate of somewhat less than one second per year, or on the order of one minute per century. This difference would not be observable in everyday timekeeping for a long time, since it was much smaller than the width of a time zone, or the offset in apparent solar time introduced by a change to daylight-saving time. Some extrapolations predicted a more rapid divergence, perhaps as large as several minutes per year.

9. The 1972 Solution

In 1972, the standards community attempted to design a time scale that would satisfy all of the previous conditions. The result was the current version of UTC.

1. The frequency of UTC would be the same as the frequency of International Atomic Time (TAI), a time scale that was designed to realize the SI second on the rotating geoid as closely as possible. The length of the International Atomic Time second was defined based on the hyperfine transition frequency in the ground state of cesium as realized on the rotating geoid. The length of the UTC second was fixed at the previously accepted value of \( 9,192,631,770 \) cycles of the cesium transition. UTC time signals and data could thus be used as a source of the standard reference frequency and the standard reference time interval almost all of the time (except for intervals that crossed a leap second, as I will discuss in the next section). The UTC time scale would be disseminated by timing laboratories and national metrology institutes based on a local realization of the cesium second.

2. The Bureau International de l’Heure was initially charged with monitoring the difference between atomic time and UT1, a time scale based on the rotation of the Earth. The job was transferred to the International Earth Rotation and Reference Service (IERS) in 1988, at the same time as the tasks of computing International Atomic Time and UTC were passed to the International Bureau of Weights and Measures (BIPM). Since the length of the cesium second was too short to begin with, and since the discrepancy was expected to increase with time, the length of the day determined by counting cesium seconds would be too short. When the discrepancy approached 0.9 s, a leap second would be added to UTC so that the discrepancy would not exceed 0.9 s. The effect of the leap second would be to allow UT1 to catch up to cesium-based UTC. The leap second would be added after the last second of the last minute of the last day of a month. In other words, the leap second would be added following 23:59:59 UTC. The months of June and December were preferred, and all of the leap seconds to date have been added at the end of one of these months.

3. The name of the leap second would be 23:59:60, and the following second would be 00:00:00 of the next day. In a month when a leap second was scheduled, the last minute of the last day of the month would thus have 61 seconds.

From the astronomical perspective, the leap second was not really an extra second. It was inserted to correct for the fact that all of the previous seconds since the last leap second were somewhat too short, and the leap second should not count in the integer number of seconds that have elapsed between any two epochs. From the astronomical perspective, the leap second would not have been needed if all of the previous seconds had had the correct length. In other words, the leap second would not have been necessary if civil time were based on UT1. (In this sense, it is analogous to the leap day, which is inserted to correct for the fact that a 365-day year is also somewhat too short relative to the solar year. As with the leap second, the leap day is not used in many applications that compute time intervals.)

Although the use of cesium seconds between leap-second events would result in a short-term discrepancy between the UTC time scale and the angular position of the Earth, there would be no long-term divergence between these two quantities. The difference between the two time scales, which was called dUT1, would be transmitted with a resolution of 0.1 s by radio time services. This resolution was considered adequate for most astronomical applications that required higher accuracy in the angular position of the Earth in space than was provided by the uncorrected cesium time scale, which could have an offset from UT1 that could be as large as \( \pm 0.9 \) s. It was implicitly assumed that applications that required a resolution greater than 0.1 s in dUT1 would already be administratively inserting the value of this parameter.

10. The Problems

The definition of UTC had two fundamental problems. These were recognized in 1972, but were not regarded as very serious at that time. Both of them are now much more serious, which is one of the reasons why a change to the definition of UTC is being considered.

10.1 Problem 1

Unlike the astronomical perspective, which regarded cesium seconds as fundamentally too short, the engineering
and scientific communities saw them as having the correct duration. From this perspective, the leap second was an extra second that introduced a step in time intervals or frequencies measured across a leap second that was not consistent with the evolution of a real time process. Measurements of velocity or of time of flight would be affected if the measurement interval included a leap second. Radio navigation systems, such as GPS, therefore defined a private time scale that did not include leap seconds beyond those that were already defined when that scale was initialized.

The decision to include the leap seconds in the initial value of the GPS system time increases the confusion. There had been 19 leap seconds added to UTC when the GPS time scale was initialized in 1980, so GPS system time was equal to UTC at that instant, but was 19 seconds behind TAI. The GPS system time does not include subsequent leap seconds, so that there are now two time scales: UTC, and GPS system time, which differ by a different number of seconds from TAI.

Galileo, the European global navigation satellite system, defined its own system time. When the system time was initialized in 1999, it was set to 13 seconds ahead of UTC at that instant. There had been 32 leap seconds added to UTC at that time, so that Galileo system time was 19 seconds \((-32+13)\) behind TAI. GPS and Galileo systems times thus had the same integer second.

Beidou, the Chinese navigation system, also defined its own system time. It was set to UTC on January 1, 2006. It therefore included the 33 leap seconds that had been inserted into UTC at that instant, but would not include any future leap seconds. It therefore would have a constant integer-second offset with respect to the Galileo and GPS system times, and an increasing integer-second offset with respect to UTC as additional leap seconds were added to UTC after January, 2006.

There are a number of other satellite navigation systems, and almost all of them have similar definitions and offsets. For example, the origin epoch for IRNSS, the Indian Satellite Navigation system, is August 22, 1999. At that instant, the INRSS system time was 00:00:00 22 August 1999; the corresponding UTC time was 21 August 1999 23:59:47. IRNSS system time was thus set 13 seconds ahead of UTC, so that it has the same integer second as GPS and Galileo. (There is no fixed relationship between the seconds fractions of the various system times.)

Finally, Glonass, the Russian navigation system, currently uses the UTC time scale as system time.

The orbital speed of all of the navigation satellites is about 4 km/s, so that a time step of 1 s can introduce a significant offset in the determination of position or time of flight. The proliferation of private time scales is undesirable in principle, and is a potential source of errors and confusion in practice. The global navigation satellite systems transmit the difference between UTC and satellite system time, but there have been many examples of receivers not processing this information correctly. These errors reappear with disturbing frequency each time a new receiver is developed.

10.2 Problem 2

The official name of the leap second was 23:59:60, but almost all clocks cannot represent that time. This is especially true for all digital systems, which keep time as the number of seconds since some origin time. These systems convert this count to the more conventional year-month-day hour-minute-second format when the value is output. This system has no provision for identifying the extra leap second. The usual implementation in these systems is therefore to effectively stop the clock for 1 s and transmit a time corresponding to 23:59:59 twice: once when that time arrives, and the second time during the leap second.

Assigning the same time tag to two consecutive seconds introduces an ambiguity in determining the time-order of events, since 23:59:59.2 during the second second actually occurred after 23:59:59.5 during the first second. To further complicate this issue, the leap second is defined with respect to UTC (and not local time), so that it occurs in the morning of the following day in Asia and Australia, and late in the afternoon in California and Hawaii. This is potentially disruptive for commercial and financial applications that use UTC to apply time tags to transactions.

Unfortunately, some implementations of digital time systems insert the extra leap second by transmitting the time equivalent to 00:00:00 of the next day twice. This has the same long-term behavior as the official version, but it puts the leap second in the wrong day, and has a time error of 1 s with respect to UTC during the leap-second insertion.

A more serious issue is the “leap smear” method, which amortizes the additional leap second over some longer time interval by adjusting the effective frequency of the clock oscillator to account for the extra second. This has the obvious advantage that the time is monotonic and time stamps cannot violate causality. However, it has an error both in time and in frequency with respect to the definition of UTC over the interval of the smear. In addition, the parameters of the “smear” are not defined in any standard, so that there is no assurance that different implementations of the method will agree among themselves on the time during the adjustment period. These considerations may not be important for a casual user of the time services, but they may be very important for users who are required to use time stamps that are traceable to national and international standards. This is especially the case for commercial and financial transactions in Europe, which are currently required to maintain a sub-second time accuracy that is traceable to national time standards. It is likely that these sub-second accuracy requirements will be implemented in the United States as well in the foreseeable future.
The Network Time protocol is often used in a hierarchical client-server model, where many systems act as clients to systems closer to a standard time reference and simultaneously act as servers to other systems. In general, most clients are configured to query multiple servers to facilitate error detection. The leap-smeared technique can be troublesome in this configuration, especially if a client system queries some servers that realize the smearing technique and some that don’t. The two queries will return time stamps that differ by a time of order 0.5 s; a discrepancy of this magnitude will be treated as an error by the client, but it is not clear which of the two time stamps will be accepted.

Finally, there are some computer systems that simply ignore leap seconds altogether. These systems will have a 1 s error from the instant that a leap second is inserted until the system is resynchronized after the insertion. This is a problem in principle, but is much less serious in practice because the clocks in these systems are not intended to be accurate at the level of a few seconds anyway, so that the additional offset due to the leap second does not make a significant difference. The only concern with this implementation is that users should understand the design assumptions and not rely on the system time for time stamps that are accurate at the level of 1 s.

The Internet time servers operated by NIST receive approximately 140 000 requests per second for time in standard network formats. The Network Time Protocol (NTP) has a provision for announcing a future leap second, but has no provision for identifying the leap second when it occurs. The protocol therefore transmits a binary time equivalent to 23:59:59 a second time during the leap second, as I discussed above. During a leap-second event, the NIST time servers will receive approximately 280 000 requests with a binary time equivalent to 23:59:59, and the users will have no simple way of knowing which of these is the actual time and which is the leap second.

NIST operates a single time server that transmits UT1 time in the standard Network Time Protocol format. The accuracy of this service is limited by the stability and reciprocity of the network time delay between the user and the time server in Colorado. This is typically on the order of a few milliseconds. This accuracy is significantly higher than the 0.1 s accuracy of radio transmissions of the dUT1 parameter.

11. The Future

The rate of increase in the length of the day is irregular, so that its short-term behavior cannot be modeled. However, there is every expectation that the length of the day will continue to increase, so that leap seconds are likely to be required more frequently in the foreseeable future. This will exacerbate the difficulties of applying them, and may increase the pressure to change the current definition of UTC. A number of changes have been discussed for the last 15+ years, but none of them has been implemented as of now (December, 2016). The proposals can be divided into two broad categories: (1) proposals that maintain a connection between UT1 and UTC, and (2) proposals that do not.

The simplest version in the first category is to do nothing: to simply continue the system as it is currently implemented. Other proposals in this category increase the limit on the dUT1 parameter so that leap seconds are required less often. However, multiple leap seconds could be needed in this solution. One version of this proposal is to add leap seconds on February 29 every four years. The insertion date would be fixed, but the number of leap seconds would vary. A similar solution would use leap minutes instead of leap seconds. Leap minutes would probably be needed only once or twice per century if the current slowdown rate continues unchanged. It is not clear that increasing the interval between leap-second events and increasing the number of leap seconds at each event is an improvement over the current system. It certainly exacerbates the problems associated with stopping the clock during the leap-second(s) event. The magnitude of the dUT1 parameter would exceed 1 s, which would break many of the radio time services that transmit this parameter.

The simplest version of the proposals in the second category is simply to stop adding leap seconds to UTC but to make no other changes. The difference between UT1 and UTC would increase, and users who needed UT1 time would have to administratively add the value of dUT1. Time services that transmitted UT1 time, similar to the existing NIST Internet time server, might be developed, although this would be another source of confusion. The format of many time transmissions would have to be modified to accommodate values of dUT1 that would exceed 1 s. The formats of the messages from the NIST radio stations WWV and WWVB would have this problem. The integer-second time offsets between UTC and the various time scales of the global navigation systems would remain but would not get any larger.

A related proposal would be to stop adding leap seconds, and rename the resulting time scale to something other than UTC. This would emphasize the fact that the implementation of the time scale had changed. (There is a precedent for not doing this. The fundamental definition of the meter has been changed several times without renaming the quantity, and there has been no change in the name when the realization of the volt was changed.)

The most extreme version of proposals in this category would be to use International Atomic Time (TAI) instead of UTC. This would require a large time step to remove all of the leap seconds that have been inserted into UTC since 1972. The time offsets of the various global navigation system time scales with respect to TAI would presumably remain. A change in the name of the legal time scale to anything other than UTC would have a significant impact on the legal definition of time in many countries.
There has been some discussion that there is (or should be) a difference between the metrological definition of UTC as a realization of the SI second on the rotating geoid, and the details of the format used to transmit the UTC time and the dUT1 offset in radio signals. The metrological definition would presumably be the province of the standards community in general and the International Bureau of Weights and Measures (BIPM) in particular, while the transmission format would be the province of the International Telecommunications Union (ITU) and the World Radio Conference. These discussions are in the very early stages. This question will probably be discussed at the next meeting of the BIPM Consultative Committee for Time and Frequency (CCTF), which will meet next year (June 2017). The World Radio Conference (WRC-15) of the International Telecommunications Union decided to postpone a final decision to the next full meeting in 2023, and it is almost certain that the current leap second system will remain in place at least until then.

12. Summary

The process of adding leap seconds to UTC is designed to maintain a close link between UTC and UT1, a time scale related to the rotation of the Earth. The leap-second process introduces difficulties for applications that use UTC for frequency or time interval, and for applications that must apply time tags to events that happen during a leap second. There is no simple solution that can satisfy all of these requirements. Whatever solution is finally adopted is going to introduce some level of difficulty for some class of applications. Where you stand on this question is largely determined by where you sit.

13. Bibliography


