29

Distributing Time and Frequency Information

Judah Levine

National Institute of Standard and Technology, United States

29.1 Introduction and Clock Statistics

This chapter describes a number of methods that can be used to distribute time and frequency information with the goal of calibrating or synchronizing a local clock. The primary focus will be on methods that use the signals from global navigation satellites – primarily the signals from the US GPS satellites. However, the basic techniques are quite general and are not limited to any specific navigational satellite constellation.

The two-way method of estimating the delay in the link between the local clock and the remote time reference will also be discussed. Unlike methods that use the signals from global navigation systems, in which both the local and remote systems passively receive the signals from the satellites, both end points actively transmit and receive in all implementations of the two-way method.

The discussion assumes that the user has a local clock or frequency standard, and that the purpose of receiving time and frequency information from a remote source is to improve the time accuracy or the frequency stability of the local device. The goal then is to combine the data from the remote reference with the reading of the local clock to produce a combination that incorporates the best features of both contributions and to realize a combination that is better than either one by itself. From this perspective, any discussion of distribution methods should include some understanding of the characteristics and limitations of real clocks and frequency standards. These concepts will be introduced in the next section.

The statistical methods that are used for characterizing the performance of frequency standards are also useful for describing the statistics of the network connection to the remote reference. In most of the situations that will be discussed in this chapter, the characteristics of the remote reference clock itself are not very important because the remote clock is often much more stable and accurate than the channel delay.

The first step in the synchronization process is to receive the time datum from the GNSS satellite. The methods for doing this are the same as the methods for position or navigation applications. The signals received from the GNSS satellites are processed to extract the pseudorange, with the difference that the position of the receiver is presumed to be known a priori by some method outside of the scope of the synchronization process. Therefore, the signal from a single satellite is sufficient, since the only unknown in the pseudorange determination is the difference between the satellite time and the time of the local system clock.

The pseudorandom code transmitted by each satellite contains an explicit link to the time as derived from the clock onboard the satellite. The parameters in the ephemeris message can link this time to the satellite system time and also to UTC, or Coordinated Universal Time, which is the legal time scale in most countries. For the signals from the GPS satellite constellation, the link to UTC is through UTC(USNO), the time scale of the US Naval Observatory. Other satellite systems have comparable links. These links are generally predictions computed by the ground controllers and uploaded into the satellite periodically. Many timing laboratories and national metrology institutes monitor the time of the constellation and publish the offset between satellite system time and UTC(lab), the UTC time scale of that laboratory. These data are generally not available in real time. From these considerations, it is clear that the stability and accuracy of the clock in the satellite are most important for short averaging times and real-time applications, while the characteristics of the system time and the link to the time scale of the US Naval Observatory are relevant for longer averaging times and non-real-time applications.

It is also possible to compute the pseudorange by using the phase of the carrier. This method supports greater

Position, Navigation, and Timing Technologies in the 21st Century: Integrated Satellite Navigation, Sensor Systems, and Civil Applications, Volume 1, First Edition. Edited by Y. T. Jade Morton, Frank van Diggelen, James J. Spilker Jr., and Bradford W. Parkinson.

resolution, since the frequency of the carrier is 1540 times the frequency of the L1 code. However, the computed pseudorange is ambiguous modulo an unknown number of carrier cycles. The signal-to-noise ratio (SNR) on the code is usually not adequate to determine this integer uniquely, and other methods must be used. A geodetic solution has the same ambiguity, and the methods for resolving the ambiguity are similar. Although the data from a single satellite is adequate for code-based time transfer when the position of the station is known a priori, it can be difficult to resolve the integer ambiguity in a carrier-phase solution with the data from only a single satellite.

A significant problem with carrier-phase pseudoranges is the possibility of a cycle slip, and most carrier-phase algorithms contain a method for detecting and removing these time steps. The cycle-slip detection problem is complicated when the L3 ionosphere-free combination (see discussions in Section 29.3.1) is used to compute the pseudorange, since the magnitude of the cycle slip can be ambiguous in this case. It is especially difficult to detect a simultaneous step in the L1 and L2 contributions to the L3 pseudorange if the steps in the two frequencies are of opposite sign.

Since the carrier-phase measurements are not unambiguously linked to the time of the satellite clock, a carrier-phase analysis typically combines a contribution of the code as well. A common method is to weight the carrier and code contributions in the ratio of 100:1. The code helps determine the correct phase integer. Therefore, a carrier-phase analysis provides increased resolution of the time differences but not necessarily increased accuracy.

The following discussion assumes that the pseudorange has been calculated from the data received from at least one satellite, and the reader is referred to Chapters 2 and 14 in this volume for more details about this. In general, the symbol τ will be used to indicate the elapsed time between data points. It is generally convenient to use a constant value for the time interval, but most of the analysis does not require that this be the case.

29.1.1 Physics of Clocks and Oscillators and Their Parameters

All clocks consist of two components: a device that produces a series of periodic events and a counter that counts the number of events and possibly also interpolates between consecutive events to improve the resolution of the measurement. The count is relative to some origin that is defined by considerations that are independent of the clock itself. As a practical matter, the time origin is generally chosen sufficiently far in the past that most epochs of interest have positive times with respect to the origin.

The choice of any particular time origin is important for everyday timekeeping, but its definition is generally not important for discussing the properties of clocks and oscillators, and this chapter will not discuss the real-world issues associated with the definition of the calendar and related questions. The *time* of a clock or the *frequency* of an oscillator can be considered as the algebraic difference between the device being described and a second, ideal, noiseless device that realizes the definitions of the standards of time or frequency. Both devices generate an output signal at the same nominal frequency. Output signals at 1 Hz and at 5 MHz are commonly used, and these output frequencies are synthesized from the internal frequency of the device, which may be quite different and which usually varies from one type of device to another one. The terms *time difference* and *frequency difference* are often used as synonyms for the *time* and *frequency* defined in the previous sentence to avoid any possible ambiguity in the discussion.

Early clocks used a frequency reference derived from a mechanical oscillation – the swing of a pendulum, for example, or the oscillation of a balance wheel. Starting in the 1920s and 1930s, the mechanical frequency reference was based on the resonance in the vibration of quartz crystals. These crystals were combined with various electronic components to generate a periodic electrical signal derived from the vibration frequency of the crystal.

Although the frequency stability of a quartz-crystal oscillator is far superior to that of a simple pendulum clock, the frequency of a device based on a quartz crystal has poor long-term stability. In addition, because its frequency is based on a physical artifact, it is difficult to reproduce exactly the same frequency from different crystals. Therefore, atomic frequency standards, whose frequencies are referenced to the transition frequency of some atom, were developed starting in about 1955. Although atomic frequency standards are universally used whenever frequency stability and accuracy are important requirements, quartz-crystal oscillators are central to almost all atomic frequency standards, and it is important to understand their statistics and limitations in designing a distribution algorithm.

To introduce the concepts, let us consider the ideal clock as an electrical oscillator that provides an output signal voltage of the form

$$V = V_0 \sin(2\pi f t) \tag{29.1}$$

where V_0 and f are the *constant* amplitude and frequency of the oscillator, respectively, and t is the time measured in seconds. This ideal oscillator generates a "tick" each time the output voltage goes through zero with a positive slope. The times of the ticks, T_n, are given by

$$2\pi f T_n = n2\pi$$

$$T_n = n\frac{1}{f} \tag{29.2}$$

where n is any integer. The output voltage of the second device we are considering is

$$A = (A_0 + \varepsilon(t))\sin(2\pi ft - \varphi) \tag{29.3}$$

Let us assume for now that

$$\frac{\dot{\varepsilon}(t)}{A_0} \ll 1 \tag{29.4}$$

and

$$\frac{\dot{\varphi}}{f} \ll 1 \tag{29.5}$$

so that the amplitude and frequency of the device under test are well-defined quantities, and the amplitude and frequency modulations are small perturbations on the output signal. The ticks of the second device will be at times

$$2\pi f T_n{}' - \varphi = n2\pi$$
$$T_n{}' = n\frac{1}{f} + \frac{\varphi}{2\pi f} \tag{29.6}$$

The time difference with respect to the ideal device is

$$x = T_n{}' - T_n = \frac{\varphi}{2\pi f} \tag{29.7}$$

The frequency difference between the two devices is the evolution of this time difference, which is driven by the time derivative of the phase difference:

$$y = \frac{\dot{\varphi}}{2\pi f} \tag{29.8}$$

If φ is measured in radians and f in hertz, then x is measured in seconds and y is dimensionless. The phase angle, φ, in Eq. 29.3, is ambiguous modulo 2π, and this integer cycle ambiguity results in a corresponding ambiguity of the time in Eq. 29.6 and the time difference in Eq. 29.7. In both cases, the value is ambiguous modulo *1/f*, and every measurement strategy depends on the fact that cycle slips are relatively rare events and can be detected when they occur because the time difference associated with a cycle slip is much larger than the value predicted by Eq. 29.8. In general, a cycle slip is treated as an outlier, and its effect is removed from the data. The time difference in Eq. 29.7 implicitly assumes that we have identified the same integer cycle number in both clocks. This is easy to do in a laboratory setting, where the measurement noise is much smaller than the *1/f* ambiguity, but this is a significant issue when the channel delay is noisy and not very stable. This issue will be discussed in a later section.

As in the discussion above, we will continue to use the symbols *t,* measured in seconds, and *f,* measured in cycles per second or hertz, to describe the physical time and frequency of the clock, where these quantities are expressed in SI units. We also recognize the minute, the hour, and the day, which are exact integer multiples of the time of the clock measured in SI seconds. Thus, the standard day consists of exactly 86 400 SI seconds.

The length of the standard day is modified on an irregular basis by the addition (or subtraction, in principle) of an extra "leap" second. These extra seconds are important for real-world timekeeping, but should not affect the statistics of the clock. For the purposes of the current discussion, we will assume that both the standard device and the device under test implement the leap second in the same way. At worst, this can produce a transient change to the time difference of ±1 s, where the sign of the transient depends on whether the device under test is fast or slow with respect to the standard device. This transient time difference is ignored in estimating the statistical characteristics of the device.

Unfortunately, a number of time-service providers implement the leap second in various non-standard ways, and this can be a source of ambiguity. The leap second is defined as the last second of a specified day, and its official name is 23:59:60. It is the 61st second of that last minute. The following second is 00:00:00 of the next day. Some time-service providers add the extra second to 00:0:00 of the day following the specified day. This has the same long-term characteristics as the definition, but there will be an error of 1 s in the vicinity of the leap second. A more difficult issue is time-service providers who amortize the leap second by a frequency adjustment over some period before the leap second occurs. This method will have both a time and a frequency error during the amortization period. In addition, there is no standard method for implementing this adjustment, so that different realizations may see different errors. This method also has the correct long-term behavior. All of the methods will have a step in the time interval or the frequency if the calculation of either of these parameters crosses a leap second event.

Eqs. 29.7 and 29.8 can be generalized to clocks that generate periodic "tick" events but that do not necessarily generate the ticks from sinusoidal signals. We will define the *time* of a clock as the algebraic difference, measured in seconds, between the clock and the standard device, and the *frequency* of a clock as the evolution of this time difference, measured in seconds per second, where the measurements are made with the output ticks and may not be derived from a corresponding phase angle. The frequency is a dimensionless quantity. These definitions are effectively the same as those in the previous discussion, except that there is no longer any connection between time and phase or between time and the amplitude of the signals.

This frequency has no connection to f, the internal operating frequency of the device, and the two devices may generate ticks using different internal mechanisms.

Although Eq. 29.8 defines an instantaneous frequency difference in principle, the practical realization of the frequency involves the evolution of the measurements of the time difference in Eq. 29.7 over some averaging time, τ, which is generally measured in seconds. (The same implicit averaging time is used even when the time and frequency of the device are not derived from a sinusoidal signal.)

The frequency of a clock often has both deterministic and stochastic variation, and these variations, which are given by the time derivative of y, are expressed in units of seconds per second^2 or 1/s. The time, frequency, and frequency variation of a clock with respect to the standard reference device will be represented by x, y, and d, respectively. It is possible in principle to model the behavior of a clock by adding higher-order derivatives of the time difference, but this is not done in practice for reasons that we will discuss in a following section.

29.1.2 Clock Noise

With the assumption of Eq. 29.4, the amplitudes of the two signals do not enter into the time difference, and the amplitude modulation in Eq. 29.3 generally does not have a significant impact on the time and frequency differences, provided that the tick of each clock is realized by a zero-crossing detector. This is the most common measurement technique, although some systems determine the time difference as the delay needed to maximize the cross-correlation between the signals defined by Eqs. 29.1 and 29.3 (or by some equivalent method that maximizes the cross-correlation between a signal from the reference clock and the same signal generated by the local device). Although this method has the potential advantage that it uses the entire signal rather than just the signal near the instant of the zero crossing, it can be sensitive to harmonic distortion and to amplitude-modulation noise, and the methods for addressing these problems often reduce the theoretical advantage of the cross-correlation method. The cross-correlation method, based on pseudorandom codes, is used in GNSS signals because it requires less peak power in the transmitter. The binary nature of the pseudorandom codes used in GNSS signals minimizes the impact of intermodulation distortion.

An important contribution to the noise budget of a time-difference measurement is the stochastic variation in the phase of the oscillator. This phase noise introduces a corresponding stochastic variation in the time difference through Eq. 29.7. Although we have generalized the concept of time difference to include ticks that are not necessarily related to the zero crossing of a sinusoidal signal, the fluctuations in the time differences of clocks are generally referred to as "phase noise" in the literature even when the output of the generator is not sinusoidal. This chapter will use this designation as well.

In many oscillators, the fluctuations in the phase noise of the clock are rapid compared to the time interval between measurements. The impact of the instantaneous phase fluctuation on any measurement is then nearly independent of the impacts on the previous or following measurements, and we can model the stochastic contribution of the phase noise to the time differences by

$$\langle x(t+\tau)x(t)\rangle = \sigma_x{}^2\delta(\tau) \tag{29.9}$$

where $\delta(\tau) = 1$ if $\tau = 0$ and is 0 otherwise. The variance of the time differences is σ^2.

The Fourier transform of Eq. 29.9 describes the power spectral density of the fluctuations; this transform is a constant at all Fourier frequencies, so that this variation is usually referred to as white phase noise. As we will mention again in subsequent discussions, the integral of a constant Fourier spectral density over all frequencies diverges to infinity, implying infinite total power. This is not a problem in real systems, where the Fourier spectrum in practice is modeled as band-limited at the low-frequency end by $1/T$, the inverse of the total time of the measurement and at the upper-frequency end by $1/2\tau$, twice the reciprocal of the time between measurements. The assumption that the signal is band-limited in this way is particularly important when the data are modeled with a discrete Fourier transform because the sampling interval of the discrete transform aliases power above the high-frequency limit of the model into a lower frequency in the modeled passband. Power at Fourier frequencies below the low-frequency cutoff appear as long-period drifts, which are not well characterized in the Fourier domain.

An implicit assumption of the statistics of the time differences in Eq. 29.9 is that the data are stationary. That is, Eq. 29.9 does not depend on the value of t – the same statistical result would be obtained from a calculation that used any section of the data. This assumption is difficult to verify experimentally because any data set consists of only a finite number of measurements, and it is extremely unlikely, and probably impossible, for *every* section of data to satisfy Eq. 29.9 exactly. In practice, the best that we could hope for is that the deviations from the prediction of Eq. 29.9 for any block of data are consistent with the usual statistics of random variables. This is a self-fulfilling statement in many practical situations, since estimates that do not conform to Eq. 29.9 are treated as outliers and ignored.

The fluctuations in the time of the clock also drive corresponding fluctuations in the frequency, and we will discuss the characteristics of these fluctuations in the next sections. For now, it is important to recognize that these time and frequency differences are related, and that both arise from the stochastic properties of the clock itself.

29.1.3 Measurement Noise

It is useful to consider again the model that was introduced in the previous section. We will consider an oscillator that produces a sinusoidal signal as in Eq. 29.3. The device generates a "tick" each time the output voltage passes through zero with a positive slope, and these ticks are counted to realize the time of the clock. The circuit that detects this zero crossing has some internal noise, $V_n(t)$, and this noise introduces a corresponding variation in the detection of the time of the tick. The slope of the output signal in the vicinity of the time of the zero crossing is given by $2\pi fA$, so that the noise voltage introduces a jitter in the time of the zero crossing, $\delta T_n'$, given approximately by

$$2\pi fA\delta T_n'(t) = V_n(t) \tag{29.10}$$

Or

$$\delta T_n' = \frac{V_n}{2\pi fA} \tag{29.11}$$

This point is illustrated in Figure 29.1.

In this simple model, the noise voltage is a random input that has a mean value of zero and a well-defined standard deviation that does not depend on time. We assume that the stochastic variation in V_n is stationary in the sense of the previous section. We also assume that the noise contribution is not correlated with the signal. Thus

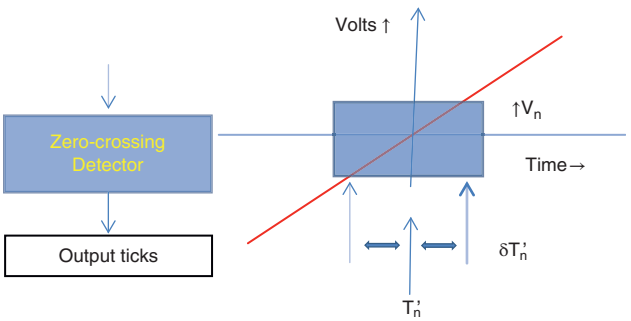

Figure 29.1 Time jitter associated with the hardware the generates an output "tick" each time the input voltage from an oscillator goes through zero with a positive slope. The output voltage in the vicinity of the zero crossing is shown in the red line, and the noise of the zero-crossing detector introduces a time jitter with an amplitude on the order of $\delta T_n'$.

$$< V_n(t+\tau)V_n(t) > \; = \; < V_n^2 > \delta(\tau) \tag{29.12}$$

and, therefore,

$$\langle T_n'(t+\tau)T_n'(t)\rangle = \left\{\frac{1}{2\pi fA}\right\}^2 \langle V_n^2\rangle\delta(\tau) \tag{29.13}$$

where $\delta(\tau) = 1$ if $\tau = 0$ and is zero otherwise.

The power spectra of the noise voltage and time jitter are driven by the Fourier transform of Eq. 29.12. The result is a constant value for all Fourier frequencies, so that this effect contributes to the white phase noise discussed in the previous section. The previous comments with respect to the bandwidth limits discussed in that section apply here as well.

In this simple model, the measured time difference is unbiased and converges to the true time difference, although the convergence may not be uniform. That is, the mean estimated from N measurements will have a smaller formal standard deviation than the mean estimated from $N-1$ measurements, but it is not guaranteed to be more accurate for any specific choice of N. The N-th measurement can deviate from the mean by an arbitrary amount, although large deviations become increasingly unlikely. As a practical matter, most measurement protocols treat a value as an outlier and ignore it if it deviates from the mean by more than 3 or 4 standard deviations. This limit is imposed administratively and is outside of the scope of a strict statistical evaluation of the data.

There is an important difference between estimating the mean value by using all of the data in a single batch process, and estimating it in an iterative fashion as each new data point is acquired. The batch process algorithm computes the mean of N data points in the usual way:

$$\bar{x}_N = \frac{1}{N}\sum_1^N x_i \tag{29.14}$$

The N data points enter the sum symmetrically, and all of them are treated equally. However, if we compute the mean iteratively as each new data point is received, and if we use Eq. 29.14 for $N-1$ points and then again for N points, then the mean after the N-th point is given in terms of the mean value computed from the previous $N-1$ points by

$$\bar{x}_N = \bar{x}_{N-1} + \frac{1}{N}(x_N - \bar{x}_{N-1}) \tag{29.15}$$

Thus, when the calculation is done iteratively, each new point is weighted by the factor $1/N$, and its contribution to the new estimate of the mean depends on how far it deviates from the running estimate. If the mean is calculated in this way, it is important to reject outliers that are acquired early in the measurement process, since they may bias the mean estimate for many subsequent measurement cycles.

Eq. 29.15 is a specific example of a more general process, where the "innovation" – the term in parentheses on the right side of the equation, which describes the difference between the current measurement and the mean value based on previous data, is weighted by a "gain function" $-1/N$ in this case. The weight in Eq. 29.15 is independent of the magnitude of the innovation, but most real algorithms reject a data point whose innovation magnitude exceeds some preset threshold determined by the running estimate of the standard deviation. This effectively sets the weight to zero. Typical thresholds are three or four times the running standard deviation. Most algorithms completely ignore the rejected point, so that the weight assigned to the next value after the rejected one is $N + 1$ rather than $N + 2$.

29.1.4 Analysis for Pure White Phase Noise

If we assume that the time, frequency, and frequency drift parameters of the device under test are strictly constant during the measurements and that the stochastic variation in the estimates of these parameters arises totally from the white phase noise of the measurement process, then it is a simple matter to estimate the state parameters by using conventional least squares. We model the measured time differences by

$$\hat{x}(t) = x_0 + y_0 t + \frac{1}{2} d_0 t^2 \tag{29.16}$$

and we would adjust the constant parameters x_0, y_0, and d_0 to minimize the mean square differences between the modeled time differences and the measurements. The residuals of the least-squares fit are driven by the noise process in Eq. 29.13, and the estimates of the time, frequency, and frequency drift parameters in Eq. 29.16 are unbiased estimates of the true values. An important assumption of this analysis is that the white phase noise arises solely from the measurement process and that there is no associated stochastic frequency variation. In other words, the deterministic parameters of the clock itself are constant values; the white phase noise contribution of Eq. 29.9 is negligible relative to the contribution of Eq. 29.13.

The simple least-squares analysis based on the clock model in Eq. 29.16 is useful in some limited circumstances, but is generally not optimal for three reasons.

1) The parameter estimation method of Eq. 29.16 is a batch process, and all of the previous data must be saved to compute a new estimate each time a new measurement is completed. This becomes increasingly cumbersome as the number of measurements increases, and is particularly difficult in many situations where the measurement process never ends.
2) Each new measurement can potentially modify the previously computed estimates of the parameters of the device, and this may be undesirable if the previously computed parameters have already been used in some application. A batch analysis is not suited to real-time time scale calculations for this reason.
3) The assumption that the time, frequency, and frequency drift of a device are strictly constant is not an accurate description of most real oscillators, and the fundamental assumption of the least-squares analysis in Eq. 29.16 is not adequate to model real devices.

Section 29.1.8 will address these limitations by generalizing the batch model of Eq. 29.16 to an iterative model.

29.1.5 White Frequency Noise

An atomic clock typically consists of a "physics package," which prepares atoms in the lower state of a specified "clock" transition. The atoms are illuminated by the output of an oscillator, which is typically a quartz-crystal oscillator that can be electrically tuned over a small range of frequencies. The frequency of the oscillator is adjusted so as to maximize the number of atoms that make a transition to the upper state of the specified clock transition, and the frequency is electronically locked to that point.

The frequency lock typically uses an error signal derived from the first derivative of the transition rate. The first derivative goes through zero at the peak of the transition rate as a function of frequency. (A number of effects can introduce an offset between the frequency corresponding to the center of the absorption line and the frequency at which the first derivative is zero. For the moment, we will assume that these effects are either constant or else change only very slowly in time.) If the deviation of the frequency from the center of the transition frequency is not too large, the magnitude of the error signal is proportional to the frequency offset, and the sign gives the direction of the frequency error. In principle, the linearity of the relationship between the error signal and the frequency offset is not important in first order, and the frequency of the oscillator reproduces the frequency of the atomic transition (with a possible constant offset). In practice, the zero-crossing detector in the frequency lock loop has some electronic noise, which is completely analogous to the phase noise of the previous section. This electronic noise interacts with the finite slope of the frequency error signal. This noise is in the frequency control loop and therefore introduces short-term random fluctuations in the output frequency of the device. See Figure 29.2.

The white frequency noise generated by the noise in the control loop introduces a random fluctuation into the evolution of the time of the clock. Since the frequency

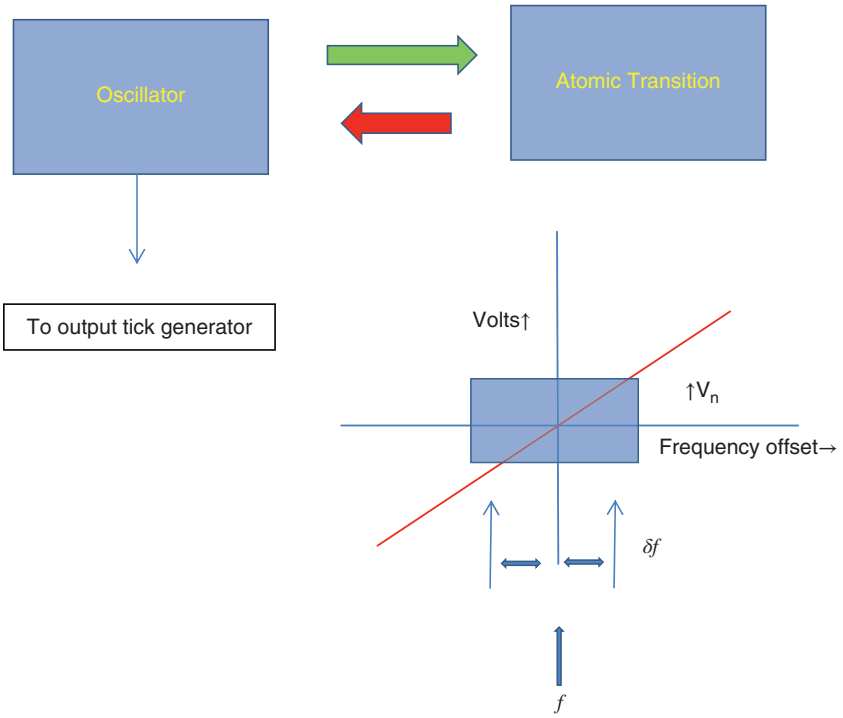

Figure 29.2 An atomic clock typically combines a quartz-crystal oscillator and a "physics package," which consists of atoms that have been prepared in a specific lower energy state. The atoms interact with the output of the oscillator (green arrow), and the frequency of the oscillator is adjusted until the atoms undergo a transition to a specific upper state. The oscillator frequency is then locked to the frequency that maximizes the transition rate (red arrow). There is some noise in the lock loop, analogous to the phase noise source in Figure 29.1, but this noise affects the *frequency* of the oscillator and not its output phase. This gives rise to white frequency noise in a manner analogous to the white phase noise of Figure 29.1.

fluctuations are uncorrelated from one measurement interval to the next one, the *evolution* of the time difference (not the time difference itself as in the previous section) is now randomly distributed about the current value. In other words, the time of the clock is now a random walk with a variable step size. If the frequency noise is a zero-mean process, after any measurement of the time of the clock the time at the next measurement will be randomly distributed around the current value, with a magnitude determined by the time between measurements and the amplitude of the frequency noise. The optimum prediction of the next value is exactly the current value with no averaging. The estimate is optimum in the sense that it is unbiased, and no more accurate estimate is possible. The mean time is no longer a meaningful stationary parameter; the calculated mean time depends on how much data are included in the calculation, and the mean will vary depending on exactly which ensemble of measurements are included in the calculation.

It is clear that the least-squares analysis based on Eq. 29.16 is no longer appropriate because the frequency is now a random variable, and it cannot be treated as a constant value over the entire data set. Since the fluctuations of the frequency are assumed to have a mean of zero, the mean frequency is still a well-defined quantity in this situation, so that the average of consecutive frequency estimates will converge to the true value. If the average frequency is estimated from N equally spaced time differences, then the average is given by

$$\bar{y} = \frac{1}{N}\left\{\frac{x_2 - x_1}{\tau} + \frac{x_3 - x_2}{\tau} + \ldots + \frac{x_N - x_{N-1}}{\tau}\right\} = \frac{x_N - x_1}{N\tau} \tag{29.17}$$

The average frequency over the total time interval is computed from the two end-point values and the total elapsed time. If the frequency has a white spectrum, then this estimate is unbiased and converges to the true frequency of the device. The mean time of the device, computed from any finite ensemble of measurements, exists in a formal sense but is not well defined. The white frequency noise implies that the evolution of the time difference after any measurement is randomly distributed about the current value, so that the optimum estimate of the next value for any time in the future is the current value with no averaging. This estimate is unbiased; the prediction error will have a mean of zero on average, and the variance of the time difference is

driven by the variance of the frequency noise acting over the interval between the measurements. As with any random variable, the variance of the mean frequency decreases as $1/N$ as more frequency estimates are computed. If the frequency is estimated over some time interval $N\tau$, the standard deviation of the prediction error of the time difference over the same interval into the future increases as the square root of the prediction interval.

29.1.6 Frequency Drift

All quartz oscillators and many atomic frequency standards also have both deterministic and stochastic frequency drift. The frequency of a quartz-crystal oscillator depends on a number of environmental parameters such as temperature and humidity, and the transition frequency of the atoms in the physics package is affected by extraneous electric and magnetic fields (the Stark effect and the Zeeman effect, respectively), by collisions between the atoms themselves and between the atoms and the container in which they are trapped, and all of these effects often change slowly with time. These effects complicate the analysis of the previous section, because the fluctuations in the output frequency of the device are no longer a simple random variable. As with all other parameters, the deterministic variation is much less troublesome than the stochastic variation, because the deterministic variation in the frequency can be estimated or modeled and removed.

For example, it is straightforward to estimate a deterministic, constant frequency drift in the presence of pure white frequency noise. In this simple situation, we could model the frequency at any time, t, in terms of the assumed constant frequency drift, d, the time that has elapsed since the measurements were started at time t_0 and the frequency at that time, y_0.

$$y(t) = y_0 + d(t - t_0) \tag{29.18}$$

and we would estimate the two parameters, y_0 and d by using standard least squares. The residuals of the fit would be driven by the white frequency noise, and the estimates of the initial frequency and the frequency drift would be unbiased. Although this procedure is simple in principle, it often has a practical difficulty because the stochastic fluctuations in the frequency caused by the white frequency noise are larger than the impact of the deterministic frequency drift, so that the drift determination has a significant uncertainty unless very long measurement times are used. In other words, the frequency drift, which is the slope parameter in Eq. 29.18, is not well constrained by the least-squares process because the scatter of the frequency estimates about the least-squares straight line is too large. This problem will appear again when discussing the iterative model of the clock parameters.

This problem becomes more serious if higher-order derivatives are used in the model. For example, if we were to assume that the frequency drift also had a constant deterministic variation, Δd, then we could estimate it by adding a quadratic term to the right side of Eq. 29.18:

$$y(t) = y_0 + d_0(t - t_0) + \frac{1}{2}(\Delta d)(t - t_0)^2 \tag{29.19}$$

where y_0, d_0, and (Δd) are the values of the frequency and its linear and quadratic variability, and all of them are assumed to be constants for the entire data set. The impact of the Δd parameter on the frequency increases as the square of the elapsed time, so that it is very important in the long run. However, separating its impact from the fluctuations in the frequency noise may be difficult in real measurements which span a relatively short time interval. In many practical devices, the elapsed time needed to estimate the drift parameters in the presence of white frequency noise is so long that the assumption that the parameters are constant is no longer accurate.

29.1.7 The Fourier Picture and Flicker Noise

All of the white processes discussed in the previous sections are characterized by an autocorrelation function similar to the relationships in Eq. 29.9 or Eq. 29.13, and therefore all of them have a power spectral density that is constant for all Fourier frequencies as discussed above (with the understanding that every real process must be band-limited in some way and that the infinity implied by a constant power spectral density is not a real-world problem).

However, during any measurement interval, the evolution of the frequency has two components: a random component due to the white frequency noise and the *integral* of the frequency drift fluctuations. Likewise, the evolution of the time has three components: a random component due to the white phase noise, the *integral* of the frequency fluctuations over the interval between measurements, and the double integration of the deterministic and stochastic contributions of the frequency drift. Each of these time-domain integrations multiplies the input power spectral density by a factor proportional to $1/F^2$, where F is the Fourier frequency of the noise contribution. The evolution of the time difference, which is typically the only parameter that is observed, therefore has a noise spectrum that is very far from a white process – the spectrum of the time differences typically has a very strong divergence at low Fourier frequencies, which is the result of the integrations discussed earlier. (This can be a serious problem if a discrete Fourier transform is computed, since a large fraction of the power

may be concentrated in the first one or two frequency estimates. This is an important justification for the use of the time-domain statistics to be discussed in the next sections.)

In addition to the even-power noise spectra discussed above, real data also exhibit a noise contribution to each parameter whose power spectral density varies as 1/F. There is no simple physical explanation for these so called "flicker" processes, although flicker of phase, for example, can be modeled as intermediate between white phase noise, where the estimate of the time difference continues to improve without limit as more data are acquired, and white frequency noise, where the best estimate of the time difference is the current value with no averaging. From this perspective, a flicker process has a finite optimum averaging time; the data look coherent over some time interval, but that coherence is not sustained as more data are acquired because the underlying driving term is a random process. The integral of the power spectral density of a flicker process diverges logarithmically, so that a real process must be band-limited as discussed above for white processes.

There is no guaranteed optimum strategy for predicting the state parameters of a clock in the presence of flicker noise as there was for the even-powered noise processes discussed in the previous sections, since measurements driven by a flicker process are not a Gaussian random variable with a well-defined mean and standard deviation. If t is the current time and $x(t)$ is the time difference at that instant, then, if the system noise is pure flicker frequency fluctuations, an approximate prediction τ seconds into the future is given by

$$\hat{x}(t+\tau) = 2x(t) - x(t-\tau) + \left(\Delta^2 x\right)_{avg} \tag{29.20}$$

The first two terms on the right side of Eq. 29.20 are simply the prediction based on the assumption of a linear extrapolation based on a constant frequency offset. The third term is the average value of the second difference of the data over the previous measurements, where the second difference is also computed with an averaging time of τ.

29.1.8 The Iterative Model of Clock Behavior

As mentioned at the outset, the implicit purpose of receiving time information from a remote reference is to characterize a local clock so that the local clock reading can be used in some application. (Although it is possible in principle to dispense with the local clock and to directly use the time data received from the remote reference, this is almost never an optimum strategy because the stability of the local clock provides useful information, as discussed below. Its time is also used in "holdover" mode if the connection to the remote reference fails.)

Based on the discussion of the previous sections, the time of the local clock has a complicated noise spectrum, and the calibration algorithm must be designed with these noise contributions in mind. Since neither the frequency nor the frequency drift can be regarded as simple constants over extended time intervals, and since any batch analysis algorithm has practical difficulties as discussed above, we replace the simple model of Eq. 29.16 with a more realistic model that supports the evolution of the clock parameters that are needed for real-world devices.

As before, we characterize the state of the clock with three parameters: the time, the frequency, and the frequency drift, but these parameters have stochastic contributions driven by noise inputs, are no longer constant but evolve with time. The parameters at time t_k are estimated based on the previous values at time t_{k-1}:

$$\begin{aligned}
\hat{x}(t_k) &= x(t_{k-1}) + y(t_{k-1})\tau + \frac{1}{2}d(t_{k-1})\tau^2 + \xi \\
\hat{y}(t_k) &= y(t_{k-1}) + d(t_{k-1})\tau + \eta \\
\hat{d}(t_k) &= d(t_{k-1}) + \zeta \\
\tau &= t_k - t_{k-1}
\end{aligned} \tag{29.21}$$

It is convenient to work with a constant time interval between measurements, but that is not a requirement of the model.

Each of the three state parameters has an associated noise contribution, which is assumed to be a random process with no correlation to the other noise contributions and with no dependence on the time, t. All of the noise terms obey relationships of the form

$$\begin{aligned}
\langle \xi(t+\tau)\xi(t) \rangle &= \sigma^2 \delta(\tau) \\
\langle \xi(t)\eta(t) \rangle &= 0
\end{aligned} \tag{29.22}$$

The goal of the measurement process is then to estimate the deterministic parameters in these equations so that the predicted state of the clock for some interval into the future is as accurate as possible. In general, the calibration process consists of measuring the time difference between the device under test and the reference device. We assume for now that the data from the reference device is correct and that the communications channel is noise-free. We will return to the real world in Section 29.3.

The fundamental difficulty with this process is that there is a single measurement – usually the time difference, but the variance of this measurement must be distributed among three deterministic (x, y, d) and three stochastic (ξ, η, ζ) parameters, none of which are known a priori in general. At best, the only things known about the stochastic parameters are their variances. Furthermore, there is no term that models the flicker contribution in any of the equations.

In addition, the equations themselves have an implicit inconsistency. For example, the first equation assumes that the frequency y_{k-1} is a constant for the entire time interval between t_{k-1} and t_k, but the second equation explicitly contradicts this assumption. The equations are a reasonable approximation only if the variation in the frequency between measurements is small enough to be ignored. That is,

$$\{d(t_{k-1})\tau + \eta\}\tau \ll y(t_{k-1}) \tag{29.23}$$

Since the deterministic and stochastic parameters are characteristics of the device and are generally not adjustable, this equation implicitly limits the maximum time between measurements, which started out as a free parameter. On the other hand, if the goal of the measurement process is to estimate the deterministic model parameters of the clock, then the measurement interval must be long enough so that the measurements are not dominated by the white phase noise of the measurement process. That is,

$$y(t_{k-1})\tau + \frac{1}{2}d(t_{k-1})\tau^2 \gg \xi \tag{29.24}$$

These two equations define the range of the time interval between measurements, although there are strategies that can sometimes reduce the requirement of Eq. 29.24. To see this, suppose that we choose an interval between measurements that reverses the inequality in Eq. 29.24. That is, we choose an interval between measurements short enough such that

$$y(t_{k-1})\tau + \frac{1}{2}d(t_{k-1})\tau^2 \ll \xi \tag{29.25}$$

In this case, the first Eq. 29.21 shows that the measured times are limited by pure white phase noise,

$$\hat{x}(t_k) = x(t_{k-1}) + \xi \tag{29.26}$$

and the mean of N of these rapid measurements, made over some small time interval, dt, will have a standard deviation of $\xi/\sqrt{\mathrm{N}}$.

$$\hat{x}(t_k) = \frac{1}{N}\sum_{\propto\, = \,-dt/2}^{\alpha\, = \,dt/2} x(t_{k-1} + \alpha) \tag{29.27}$$

$$\sigma\left(\hat{x}(t_k) = \xi/\sqrt{N}\right. \tag{29.28}$$

We can now use Eq. 29.21, replacing the time state, $x(t_{k-1})$, by the estimated mean value of the time state computed by using Eq. 29.27. The phase noise contribution has been reduced by the factor $1/\sqrt{N}$, and this reduction can be used to modify the requirement of Eq. 29.24 and thereby allow a shorter time interval between measurements.

In summary, this measurement strategy consists of groups of measurements made very rapidly so as to be dominated by white phase noise. The groups are separated by a longer time interval as defined by Eqs. 29.23 and 29.24. The primary advantage of this shorter time interval between measurements is to provide more rapid detection of outliers and other errors. In practice, Eq. 29.27 would be modified to reject outliers – data points that did not satisfy a consistency comparison with the running standard deviation. Another technique would compare the running standard deviation, computed over the previous points, with a value computed from a longer time series. This method can be used to detect a change in the statistical characteristics of the device under test that is not limited to a single outlier point. As discussed before, rejection of outliers is an administrative decision that is outside of the realm of statistics.

This example illustrates a more general requirement. In order to design any measurement algorithm, we must first estimate the contributions of the stochastic parameters in Eq. 29.21, recognizing that the process is under-determined because there are three noise parameters but measurements of only a single quantity – the time difference between the device under test and the reference device.

An important result of this section is that the averaging time between measurements is constrained by the magnitude of the various noise processes. The two-sample Allan variance provides a mechanism for estimating the contribution of each noise term to the variance of the measurement for any specified averaging time, and statistic will be introduced in the next section.

29.1.9 Time-Domain Statistics, the Simple Two-Sample Allan Variance and Its Fancier Relatives

To introduce the concept, suppose we wish to determine the frequency difference between two clocks that have no intrinsic noise of any type. We estimate the frequency difference by two measurements of the time difference separated by a time interval, τ. See Figure 29.3. Each of the time-difference measurements has an uncertainty, ε, which is illustrated by the error bars on the two measurements. This noise is a property of the measurement process – *not* the clocks. Although the true frequency difference, shown by the red line, is constant, the noise in the measurement of the time differences introduces an uncertainty in the estimate of the frequency difference as shown by the two yellow lines. The magnitude of this uncertainty is approximately $\delta y(\tau) = \sqrt{2}\varepsilon/\tau$, and we would expect that repeated estimates of the frequency difference between the two devices would scatter around the true value by this amount. We can turn this result around and assert that if

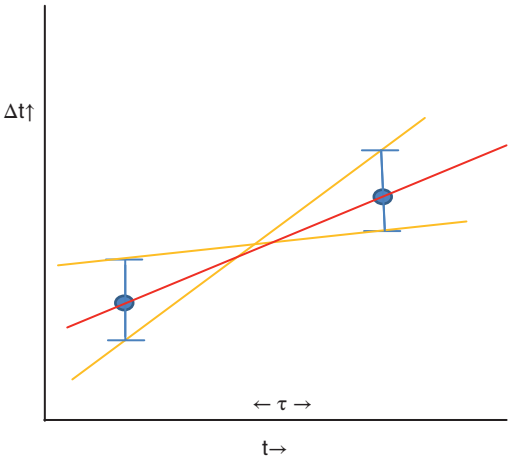

Figure 29.3 Estimating the frequency difference between two noiseless clocks. The estimate is based on two time differences separated by a time τ. The noise in the two time-difference measurements is shown by the two error bars and has a magnitude ε. The noise in the time differences translates into an uncertainty in the estimate of the frequency difference, as discussed in the text.

we measure the frequency difference between two devices by two time differences separated by a time interval τ, and if these frequency estimates exhibit a standard deviation, $\delta y(\tau)$, that varies as the reciprocal of the time between measurements, then the noise type of the measurements is characterized as white phase noise that has a standard deviation on the order of $\tau\delta y(\tau)/\sqrt{2}$. A log-log plot of $\delta y(\tau)$ as a function of τ would have a constant slope of -1, and the intercept would provide information on the amplitude of the white phase noise in the measurement process, ε.

It is important to keep in mind that the variance of the frequency estimates we have computed has nothing to do with a variation in the actual frequency of the clock itself. It arises purely from the phase noise of the measurement process, and using these data to steer the frequency of the clock would be a mistake. The clock will be more stable if left alone. Likewise, trying to improve the measured frequency stability by improving the clock will have no effect. If we are not satisfied with the noise in the data, then it is the measurement process that needs improvement. We will use similar considerations in the next sections to obtain additional insight into the optimal strategy for synchronizing a local clock based on data received from an imperfect remote clock over a noisy communications channel.

The two-sample Allan variance, $\sigma^2_y(\tau)$, is a more general version of this idea. It uses the dependence of the variance of the measured frequency of a device, $(\Delta y)^2$, as a function of the time interval between measurements to infer the underlying noise type:

$$\sigma_y^2(\tau) = \frac{1}{2}\left\langle (\Delta y)^2 \right\rangle = \frac{1}{2\tau^2}\left\langle \left(\Delta^2 x\right)^2 \right\rangle \tag{29.29}$$

where the variance in the first-order difference of the frequency measurements is typically realized by a second-order difference in the measured time differences as shown by the second form of the definition. The Allan variance is often estimated by means of a series of time differences, x_i, with a constant time interval between measurements, τ. If the first measured time difference is x_0 at time t_0, the n-th measurement is at time $t_0 + (n-1)\tau$. If there are N measurements with indices $0, 1, \ldots, N-1$, the two-sample Allan variance for an averaging time $n\tau$ is then computed by

$$\sigma_y^2(n\tau) = \frac{1}{2(n\tau)^2(N-2n)} \sum_{i=0}^{N-2n-1} (x_{i+2n} - 2x_{i+n} + x_i)^2 \tag{29.30}$$

The Allan variance assumes that the data are stationary, and therefore the estimate does not depend on t_0. The usefulness of the Allan variance derives from the fact that a log-log plot of the variance as a function of the averaging time has a slope that serves as an indication of the underlying noise type, and it is very often the case that only one noise type dominates the variance at any averaging time. The relationship between the slope of the Allan variance and the underlying noise type is given in Table 29.1.

The simple two-sample Allan variance is widely used to characterize all types of clocks and oscillators, but it has a number of intrinsic characteristics that follow from its definition and that limit its usefulness in some situations.

1) The Allan variance is a measure of frequency stability and not frequency or time accuracy. It is not sensitive to a constant time difference or a constant frequency

Table 29.1 Slope of the log-log plot of the Allan variance and Allan Deviation (square root of the variance) as a function of the averaging time for the five common noise types

	Slope	
Noise type	**Avar**	**Adev**
White phase noise	−2	−1
Flicker phase noise	−2	−1
White frequency noise	−1	−0.5
Flicker frequency noise	0	0
Random walk frequency noise	+1	+0.5

The simple Allan variance cannot distinguish between white phase noise and flicker phase noise.

difference, even though those parameters might be important in some applications.

2) A clock that has a *constant* frequency drift is just as stable and predictable as a clock that has a constant frequency and no frequency drift, but the Allan variance treats the two situations very differently – the constant frequency is ignored, but the constant drift is not. In fact, a clock whose frequency variation can be accurately modeled in *any* algorithmic way is just as predictable as a clock whose frequency is constant. Again, the two-sample Allan variance treats these two situations quite differently.
3) The simple Allan variance cannot distinguish between white phase noise and flicker phase noise (see Table 29.1). This ambiguity is removed by the modified Allan variance, which computes the original recipe over blocks of measurements rather than individual ones:

$$mod.\sigma_y^2(n\tau) = \frac{1}{2\tau^2 n^2 (N-3n+1)} \sum_{j=1}^{N-3n+1} \left(\sum_{i=j}^{n+j-1} (x_{i+2n} - 2x_{i+n} + x_i) \right)^2 \tag{29.31}$$

The slope of the log-log plot of the modified Allan variance is −3 for white phase noise and −2 for flicker phase noise, which removes the ambiguity. See Table 29.2.

The normalization constants for both variances are defined so that the estimate for white phase noise agrees with the estimate based on the conventional variance, which is optimal for a white process.

The time variance, $\sigma_x^2(\tau)$, estimates the time dispersion associated with the frequency fluctuations estimated by the modified Allan variance. It is defined as

$$\sigma_x^2(\tau) = \frac{\tau^2}{3} mod\sigma_y^2(\tau) \tag{29.32}$$

Table 29.2 Slope of the log-log plot of the Modified Allan variance and Modified Allan Deviation (square root of the variance) as a function of the averaging time for the five common noise types

	Slope	
Noise type	**Mod Avar**	**Mod Adev**
White phase noise	−3	−3/2
Flicker phase noise	−2	−1
White frequency noise	−1	−0.5
Flicker frequency noise	0	0
Random walk frequency noise	+1	+0.5

Note: The modified Allan variance distinguishes between white phase noise and flicker phase noise.

so that the slope of a log-log plot of the time variance as a function of the averaging time is +2 relative to the slope of the modified Allan variance for any noise type.

The slopes of log-log plots of the modified Allan variance or the time variance as functions of the averaging time provide estimates of the noise type, but the significance of the magnitude of these statistics at any averaging time assumes that the data will be used with the same averaging algorithm that is used in the definition of the statistic. This is often not the case, and a more conservative statistic of the time dispersion at any averaging time is often used. This statistic estimates the time dispersion over an averaging time of τ as $\tau \times \sigma_y(\tau)$.

29.1.10 Statistics in the Frequency Domain

In many cases, the characteristics of the clock or the measurement channel are characterized in the Fourier frequency domain rather than in the time domain, as in the previous discussion. In this method, the stochastic performance is characterized by a spectral density, $S_y(F)$, as a function of the Fourier frequency, F. (The Fourier frequency is the independent variable of the power spectral density, and is not related to the relative frequency of the device with respect to some reference standard, which is identified as y or to the SI frequency of the device, which is identified as f.) If the spectral density function is known, the two-sample Allan variance can be calculated as

$$\sigma_y^2(\tau) = \int_0^\infty 2 \left[\frac{sin^4(\pi F\tau)}{(\pi F\tau)^2} \right] S_y(F) df \tag{29.33}$$

The quantity in brackets in the integrand is the weighting function that maps the contribution of the spectral density at some Fourier frequency to the two-sample Allan variance. At Fourier frequencies such that $F << 1/\tau$, the weighting function tends to zero as F^2. This dependence results from definition of the variance in terms of the first difference of the frequency. At very high Fourier frequencies such that $F >> 1/\tau$, the weighting function also tends to zero as $1/F^2$ because the frequencies estimates are computed by averages of the evolution of the time difference during the time interval τ. The corresponding relationship between the spectral density and the modified two-sample Allan variance is given by

$$Mod\ \sigma_y^2(\tau) = \int_0^\infty 2 \left[\frac{sin^3(\pi f\tau)}{(n\pi f\tau)\sin(\pi f\tau_0)} \right]^2 S_y(f) df \tag{29.34}$$

The asymptotic behavior of Eq. 29.34 is similar to that of Eq. 29.33. Eqs. 29.33 and 29.34 do not have inverse relationships in general, so that it is usually not possible to compute

the spectral density from either the two-sample Allan variance or the modified version. As a special case, if $S_y(F) \sim F^{\alpha}$ and $\sigma_y^2(\tau) \sim \tau^{\mu}$, where α and μ are small integers, then $\alpha = -\mu - 1$. Thus, if the noise type can be characterized by an integer slope of the log-log plot of the two-sample Allan variance, it can also be characterized by an integer slope of the log-log plot of the power spectral density as a function of the Fourier frequency.

An important second special case is when there is a "bright line" in the spectrum of the fluctuations in the phase or the time difference. That is, the difference has a periodic contribution with a Fourier frequency F_0:

$$x(t) = A\sin(2\pi F_0 t) \tag{29.35}$$

The power spectral density is a delta function in F, and the contribution of this variation to the two-sample Allan variance is given by

$$\sigma_y(\tau) = \frac{2A}{\tau}\sin^2(\pi F_0 \tau) \tag{29.36}$$

which has a peak at an averaging time that is somewhat less than one-half of the period of the frequency F_0 superimposed on an envelope proportional to $1/\tau$.

29.1.11 Summary of Clock Statistics

The two-sample Allan variance machinery is most useful in characterizing noise processes, especially those processes that have a power spectral density that diverges at low Fourier frequencies. These processes are commonly found in time and frequency applications. Although a standard Fourier analysis expands the observational data in a complete set and can always represent the same information in principle, it can be more difficult to understand the characteristics of the oscillator because so much of the power in a Fourier analysis is concentrated in the lowest frequency estimates for this type of data. On the other hand, the two-sample Allan variance is less useful in modeling data that have deterministic contributions – a deterministic frequency drift, for example, or a noise contribution that is well characterized by a bright-line spectrum. It is almost always a better strategy to pre-whiten the data in these situations and apply the two-sample Allan variance to the residuals after the deterministic effects have been estimated and removed. This method is not perfect, of course, since the deterministic and stochastic contributions are not perfectly independent of each other. It is also difficult to pre-whiten data when a bright-line contribution to the variance is smeared out in frequency space by amplitude modulation from an unknown source, or when the bright-line spectrum is not sinusoidal so that the effect in frequency space is a series of harmonically related bright lines.

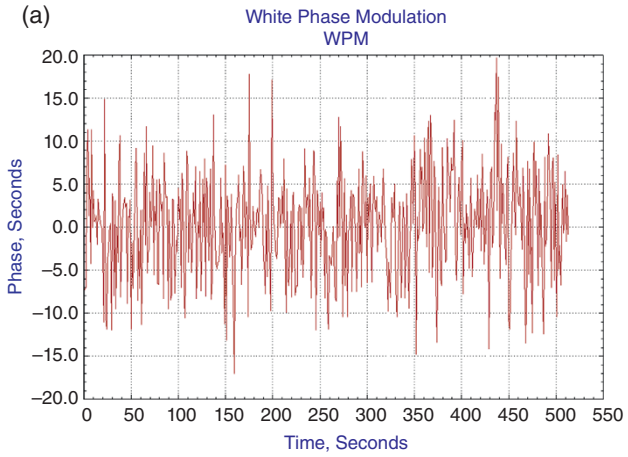

Figure 29.4a Simulated data with pure white phase noise.

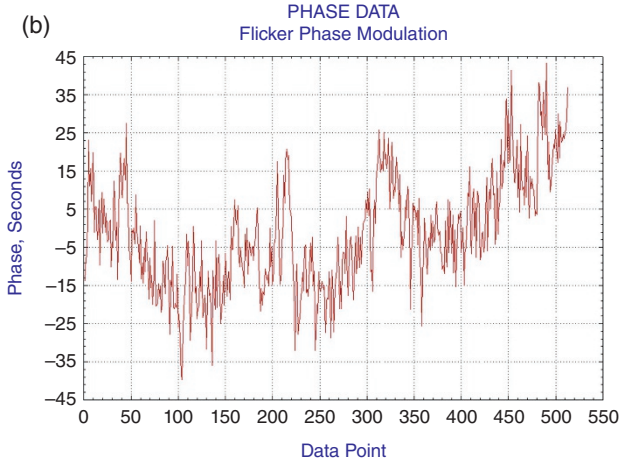

Figure 29.4b Simulated data with pure flicker phase noise.

Estimating and removing each of these contributions is possible in principle, but can be difficult in practice, especially when the Fourier estimation is complicated by the presence of a sloping noise background.

Figures 29.4a–e show simulated 1 s time-difference data computed from each of the five noise types. In each case, the time series are constructed to have the same two-sample Allan variance for an averaging time of one sample. From Eq. 29.21, we can see that the frequency drift is implicitly integrated to compute effective frequency, and the frequency is integrated to compute the time difference. These integrations scale the power spectral density of the input fluctuations by a factor proportional to the inverse of the Fourier frequency squared. This scaling "reddens" the power spectrum, and the data appear smoother in the short term as a result. However, the appearance is misleading – all of these data sets are pure noise processes. It is clear from

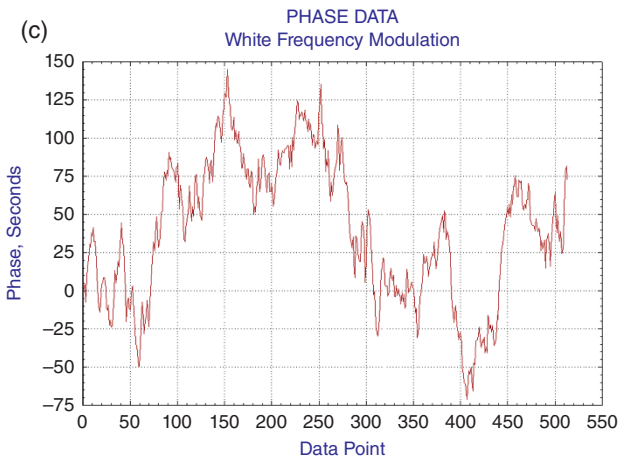

Figure 29.4c Simulated data with pure white frequency noise.

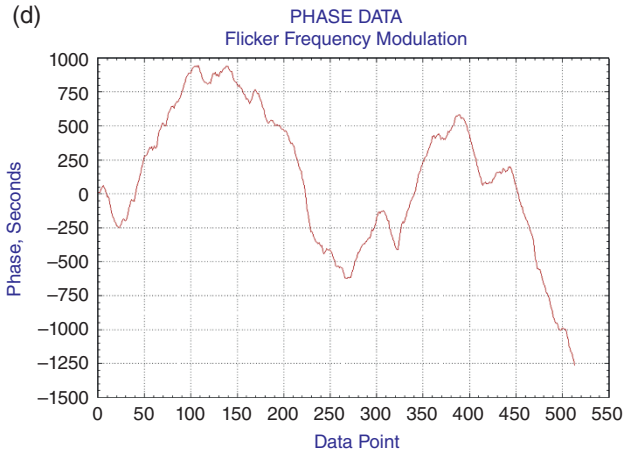

Figure 29.4d Simulated data with flicker frequency noise.

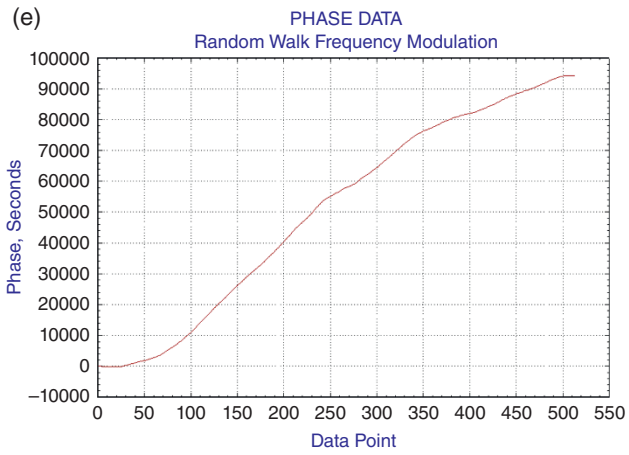

Figure 29.4e Simulated data with random walk frequency noise.

the figures that the classical variance is not stationary and depends on how much data are used to calculate it.

29.2 Accuracy and Traceability

In the previous sections, we developed the machinery that is used to characterize the performance of a local clock. The discussion was focused on stability in general and on techniques for quantifying stochastic variations and identifying the source of the noise, especially noise contributions that are often found in clock data and that cannot be adequately addressed by standard Gaussian statistics.

Characterizing the stability may be enough in some applications, but accuracy is often an equally important characteristic, and the discussion of the previous section does not provide any insight into this parameter. By "accuracy" I mean how closely does the time or frequency of the device under test conform to accepted standards – typically the standards of time and frequency maintained by a national metrology institute or timing laboratory. The standards maintained by these institutions are in turn linked to the international definitions maintained by the International Bureau of Weights and Measures (the BIPM in French).

Not all applications, however, depend on accuracy with respect to an international or national standard of time and frequency. In some applications, the accuracy requirement is that all of the members of a group of clocks agree with some common definition of time or frequency that may be only loosely related to the international definitions. For example, the *navigational* accuracy of the GPS system (as opposed to the *timing* accuracy) depends on the fact that the difference between the time signals transmitted by all of the satellites and the GPS system time be as accurately estimated and predicted as possible. The relationship between GPS system time and the international definition of time, which is Coordinated Universal Time (UTC), is not important from the navigational perspective. Since the parameters broadcast by each satellite are predictions, the accuracy requirement is driven by the frequency stability of the satellite clocks and the resulting time dispersion over the time interval between uploads. The melting pot version of common-view, to be discussed below, has a similar requirement – the method depends on the accuracy with which the time transmitted by every satellite can be related to the GPS system time. There are many other examples where the operation of a complex system depends on the time synchronization of its independent components and only indirectly on the connection of the time to national or international standards.

Long-term stability is a necessary prerequisite for accuracy, but it is not sufficient. A well-designed measurement process will be limited by white phase noise, which is amenable to averaging. The mean converges to the unbiased value and does not degrade the timing accuracy. This assumes that the clock and the measurement hardware are sufficiently stable to support whatever averaging time is required; that the averaging is confined to the white phase noise domain and that the application does not require a real-time output with a delay shorter than the required averaging time. In the same way, the frequency accuracy of a clock will not be degraded by white frequency noise, since the average frequency will converge to the true underlying frequency with no bias, provided that the averaging is confined to the white frequency noise domain. However, the time of the clock does not have a meaningful average value in this situation.

Every accuracy claim assumes a comparison with some standard, where the standard may be the time scale of a timing laboratory or the ensemble average of a number of clocks. The accuracy claim should be supported by traceability documentation, which I will describe in the next section.

29.2.1 Technical Traceability

In the technical sense, a calibration result is traceable if there is an unbroken chain of measurements between the end-user device and a recognized national metrology institute or standards laboratory. Each link in the chain of measurements must be characterized by an uncertainty. Generally, the calibrations that establish traceability must be repeated periodically because there are often long-term changes in the system that affect its calibration. For example, the links to the time scales of most national metrology institutes and the International Bureau of Weights and Measures are calibrated every two years. In the United States, technical traceability to the standards of time and frequency is satisfied if the chain of measurements is linked to the time scales at the US Naval Observatory or the National Institute of Standards and Technology.

Traceability of a frequency application can usually be satisfied by documenting the stability of the link delay over the averaging time of the frequency measurement, since the channel delay does not matter provided that it can be shown to be constant. The long-period stability of the channel delay is usually not important. This is not adequate traceability for an application that requires accurate time, since the channel delay enters in first order at every averaging time.

Full technical traceability can be very difficult to realize in many configurations because the last link in the measurement chain is often not calibrated or certified. For example, the signal in space from a navigation satellite, such as a satellite in the GPS constellation, is traceable because it is monitored by the US Naval Observatory, and the offset between the transmitted time and the time scale maintained by the Observatory is transmitted by the satellite as part of the navigation message. (The situation is actually somewhat more complicated, because the satellite transmits a prediction of this offset based on previous measurements. The actual offset between the time broadcast by the satellites and the time scale of the US Naval Observatory is published after some delay. Similar data are acquired and published by other national metrology institutes and timing laboratories. Therefore, full traceability is realized only after the fact by including these ancillary data.)

However, this traceability does not extend to the user equipment in general because it has neither been calibrated nor monitored in real time. The same limitation applies to downstream applications that use the time signals from a satellite receiver. This limitation will become increasingly important as jamming and spoofing of satellite signals becomes more common. Jamming is a denial-of-service attack, and the user can usually detect that it is happening, but spoofing, in which a rogue transmitter transmits signals that appear to originate from a real satellite, can be extremely difficult to detect. Furthermore, there are a number of local corrections, such as the additional delay introduced by the troposphere, that change with time and that cannot be estimated a priori.

Although the focus of this discussion of traceability is on signals from navigation satellites, the lack of calibration and the generally inadequate characterization of the last link in many time distribution systems is a more general problem. In most configurations, the last link in the measurement chain is not under the control of the source of the time signals, and so it is the responsibility of the end user to ensure that any traceability requirement is satisfied. Unfortunately, the end user often does not understand the subtleties of traceability and has neither the expertise nor the equipment to calibrate the end-user system. This problem is certain to become more serious as the requirements for traceable time become more accurate.

29.2.2 Legal Traceability

A number of users at financial and commercial institutions are legally required to maintain a reference time source that is traceable to national standards. In addition to the requirements of technical traceability discussed in the previous section, there are additional documentation requirements. Some of these documentation requirements are

imposed by the regulators, and others may be needed to document traceability that might be important in a potential future adversary proceeding in which the time accuracy of a financial transaction is a factor. For example, insider trading is fundamentally a time-based crime, and the accuracy of the time stamp applied to a transaction is a determining factor. The requirements are more legal than technical and are outside of the scope of this description. However, a very common problem is a reporting system that reports only failures and problems. A system that is operating properly may have no entries in the log file in this case, and it may be difficult to distinguish at some later time between a system was operating properly and a system that was totally inoperative.

29.3 Determining the Time Delay Through the Distribution Channel

In general, the accuracy of any process that is used to distribute time and frequency information is limited by the accuracy of the estimate of the delay through the calibration channel. While it is true that an application that depends on the distribution of frequency does not require a determination of the absolute channel delay, it does depend on the fact that the delay is a constant during the transmission process, and determining that the delay is, in fact, adequately invariant is often not substantially easier than going the full way and measuring it. There are a number of methods that are used to evaluate the channel delay or to remove its effect, and they will be discussed in the following sections.

29.3.1 The Two-Way Method

A widely used method for estimating the delay is the two-way method, in which the one-way delay between the reference clock and the local device is estimated as one-half of the measured round-trip value. There are several different implementations of the two-way method; we will describe the version that is generally used to estimate the delay in a digital network, but the general principles are the same for other implementations.

The two-way conversation makes use of two messages sent in opposite directions along a single channel. (The subtle implications of a "single channel" will be discussed below.) System 1 transmits a time packet at time t_{1s} as measured by its internal clock. The message is received at system 2 at time t_{2r} as measured by the clock on system 2. System 2 sends a reply back to system 1. The reply is sent at time t_{2s} as measured by the clock on system 2 and is received back at system 1 at time t_{1r}.

The round-trip network path delay is given by

$$\Delta = (t_{1r} - t_{1s}) - (t_{2s} - t_{2r}) \tag{29.37}$$

The first term on the right side of Eq. 29.37 is the total elapsed time for the round-trip exchange as measured by the clock on system 1, and the second term is the latency between when system 2 received the request and when it replied. Most digital-network implementations use the query-response model in which one of the systems initiates the conversation and the second one replies only when it receives the request. Both terms will be non-negative in this configuration.

It is also possible to design a system in which both ends transmit continuously and asynchronously. Without loss of generality, this configuration may be analyzed in the same way as above by pairing transmissions in both directions so that both terms in Eq. 29.37 are always positive.

From the perspective of the clock on system 1, the signal that was transmitted at time t_{1s} reached system 2 at time $t_{1s}+d$, where d is the one-way path delay. This message reached system 2 at time t_{2r}. The time difference between systems 1 and 2 is estimated by

$$\delta_{12} = (t_{1s} + d) - t_{2r} \tag{29.38}$$

If the inbound and outbound delays are equal, then $d = \Delta/2$. With this assumption, the one-way delay in Eq. 29.38 is estimated as one-half of the round-trip estimate in Eq. 29.37. The time difference in Eq. 29.38 becomes

$$\delta_{12} = \frac{t_{1s} + t_{1r}}{2} - \frac{t_{2r} + t_{2s}}{2} \tag{29.39}$$

The magnitude of the round-trip delay is not important. The accuracy of the estimate of the one-way delay is limited by the assumption that the round-trip delay is symmetric: that the inbound and outbound contributions are equal. If the true inbound and outbound delays are not equal, then the computed time difference will be incorrect.

Suppose that the outbound delay is not $\Delta/2$ but $k\Delta$, where the parameter k can have any value between 0 and 1. The value $k = 0$ implies that the outbound delay is negligible relative to the delay in the opposite direction, while $k = 1$ implies the reverse. The assumption of perfect path delay symmetry will introduce a time error of

$$\epsilon = (k - 0.5)\Delta \tag{29.40}$$

so that the maximum possible error is one-half of the round-trip delay, and the asymmetry problem can be minimized by minimizing the round-trip delay.

Eq. 29.40 shows that the two-way method attenuates an asymmetry in the loop delay by a factor of 2. On the other hand, a delay that is outside of the two-way measurement loop is not attenuated at all. Therefore, it is advantageous to

include as many of the delay contributions as possible into the two-way measurement loop. In other words, the local end point of the measurement loop should coincide as closely as possible with the application that will use the time datum.

The channel that links the two systems must be able to transmit messages in both directions. Real-world two-way systems can be implemented as two one-way channels that are assumed to have the same characteristics, as a single half-duplex channel where the message direction can be reversed, or as a full duplex channel that supports transmissions in both directions simultaneously.

Packet-switched networks can transmit data in both directions, but the messages in the two directions generally cannot be transmitted simultaneously. Therefore, the two-way cancellation depends on the fact that the characteristics of the channel do not change in the short interval between the transmissions in the two directions. A residual asymmetry on the order of a few percent of the round-trip delay is quite common in these networks. Since a typical round-trip delay is on the order of 100 ms, the residual asymmetry, which limits the accuracy, is on the order of a few milliseconds.

Two-way message exchanges can also use a microwave link through a communications satellite. Many timing laboratories use this configuration to compare clocks and time scales at the different laboratories. Although the path delay to a geosynchronous communications satellite is much longer than for a packet-switched network, the asymmetry is much smaller, and it possible to realize subnanosecond time comparison in this way. The accuracy of two-way satellite time transfer comes at the price of significant complexity and expense of the ground-station hardware. Maintaining the symmetry of the inbound and outbound delays is complicated by the fact that the two directions do not use the same hardware. The situation can be further complicated by the difference in the environmental parameters at the two stations, since the delay through the ground-station hardware may have an asymmetric sensitivity to the ambient temperature.

29.3.1.1 Limitations of the Two-Way Method

The accuracy of the two-way method is limited by a static asymmetry, and there is generally no way of detecting or removing this type of asymmetry unless the true time difference between the two clocks is known from some other method. It is often possible to estimate and remove the fluctuations in the asymmetry, and I will describe how this can be done below.

In addition, the two-way method requires that both the local and remote stations cooperate in establishing a measurement schedule and protocol; anonymous collaborations are not possible. Both the local and remote stations must maintain state variables so as to compute the round-trip delay by the use of Eq. 29.37 and the time difference by Eq. 29.39. This can become a significant issue when a remote calibration station is communicating with many stations simultaneously.

The two-way method is not possible if the data are received from a navigational satellite, since the transmissions are only in one direction, and other methods must be used.

29.3.2 Estimating the Delay from Ancillary Parameters

A common way of estimating the transmission delay is to use parameters transmitted by a navigation satellite or determined by ancillary measurements made on the ground. For example, all navigation satellites transmit orbital parameters that can be used to compute an estimate of the position of the satellite at any time. These parameters can be combined with the known position of the receiver (determined by some method outside of the time and frequency application) to compute the delay due to the geometrical time of flight. This computed delay might be augmented by estimates of the ionospheric delay derived from ionospheric models or by a measurement of the two-frequency dispersion – the apparent difference in the transit time of signals transmitted by the satellite at two different frequencies. Finally, the additional delay due to the refractivity of the troposphere can be estimated from ground-based measurements of atmospheric temperature and pressure. However, single-point measurements of these parameters, acquired near the receiving antenna, may not be representative of the values along the path, especially for elevation angles far from the zenith.

The uncertainty of each of these contributions is not directly related to their magnitude. For example, the geometrical path delay from a navigation satellite to a receiver on the ground is approximately 65 ms, but the uncertainty in this delay is determined by the error in the satellite ephemeris and any error in the position of the receiver. These two contributions might result in a timing error on the order of 10–20 ns, several orders of magnitude smaller than the geometrical delay itself.

The same attenuation can be realized in the contribution due to the refractivity of the ionosphere. This refractivity adds approximately 65 ns to the geometrical delay, but this additional delay can be estimated by using the dispersive nature of the ionosphere, and the uncertainty of the estimate is an order of magnitude (or more) smaller than the contribution itself. However, there are other considerations that will be discussed below.

The refractivity of the ionosphere is proportional to $1/f^2$, where f is the frequency of the carrier of the signal transmitted by the navigation satellite. The proportionality constant depends on the free-electron density of the ionosphere, which generally has a strong diurnal variation. The details of this constant, or its diurnal variation, are not important because the refractivity can be expressed as a product of a function that depends only on the properties of the ionosphere and a function that depends only on the carrier frequency.

With this separation of the refractivity into a product of two functions, it is possible to construct an "ionosphere-free" combination of the signals transmitted by GPS (and other) navigation satellites at the two different frequencies. The transmitted frequencies are called L1 and L2, and this combination, which is often called L3, cancels the additional deterministic component of the contribution of the ionosphere to the path delay. (The calculation of L3 implicitly assumes that the L1 and L2 signals travel along the same path and therefore sample the same refractivity. This assumption neglects the differential bending of the two signals at the boundary between media having different indices of refraction as calculated from Snell's law.)

The stochastic component of the signals at the two frequencies are usually not correlated and do not cancel in the computation of the L3 signal. A detailed analysis shows that the L3 signal is approximately three times noisier than either of the two transmitted frequencies that were used to construct it, assuming that the stochastic contributions to the L1 and L2 signals are approximately equal. This price is worth paying in the current context because the improvement in removing the refractivity of the ionosphere is worth the threefold increase in the noise. However, we will reconsider this point in the next section, where the conclusion is not so clear.

There is also a much smaller additional delay as the signal travels through the troposphere because its refractivity (the difference between the index of refraction of the medium and the vacuum value of exactly 1) is small, but not zero. The refractivity of the troposphere is not dispersive at radio frequencies, so that the two-frequency method that is used to estimate the refractivity of the ionosphere cannot be used.

The refractivity of the troposphere at the microwave frequencies used by the navigation satellites (on the order of 1.5 GHz) is about 300 ppm, and the scale height of the atmosphere is about 7 km. The refractivity results in an optical path that is approximately 2 m longer than the geometrical distance, so that the additional delay is typically about 6 ns when the satellite is at the zenith. If we assume a simple model of the refractivity where it is homogeneous and isotropic, then the additional delay due to the refractivity increases as the elevation decreases roughly as the reciprocal of the sine of the elevation angle. If this model was really accurate and if there were no other sources of noise, then we could observe a satellite as its elevation angle changes, fit the $1/sin(elev.\ angle)$ model to these time-difference data, and extract the delay of the troposphere at the zenith. We could use the same method to observe several satellites simultaneously at different elevation angles and extract the zenith delay of the troposphere. More sophisticated mapping functions could also be used in the same way.

These methods of estimating the contribution of the troposphere to the transit time might work in some ideal situations, but they are not really of much use in practice because there are too many other noise sources that have the same signature in time or in comparing the data from several satellites. They might be better than nothing at all, but they are often too simplistic to accurately model the true refractivity, especially at a location with a topography like that of Boulder, Colorado, which has mountains to the West and plains to the East so that the refractivity of the troposphere is unlikely to be either homogeneous in elevation or isotropic in azimuth. In addition, although any combination of multiple signals may attenuate a correlated contribution such as the refractivity of the troposphere, the uncorrelated contributions do not cancel, and the stochastic magnitude of the combination is almost always worse than the stochastic magnitudes of any one of the contributing signals taken singly. This is the same effect described above with respect to the calculation of the ionospheric-free signal L3 from L1 and L2. The improvement is usually worth the price in the case of the ionosphere, but the situation in the case of the troposphere is less clear.

In summary, the tropospheric delay often makes a larger contribution to the overall error budget of the measurement, even though its overall contribution to the total delay is much smaller than the other contributions I have discussed. There are also a number of smaller contributions to the calculation of the delay such as the motion of the station due to the Earth tides, which have a total diurnal and semi-diurnal amplitude on the order of 0.3 m in position or 1 ns in time, and the polar motion of the entire Earth. These effects are included when the very highest accuracy is required, but are normally ignored in more routine applications. Although the one-way method that estimates the propagation delay is often used in routine applications, it is not adequate when nanosecond accuracy is required, and other methods are used for these applications. These methods will be discussed in the next sections.

29.3.3 Physical Common View

In a physical common-view measurement, two or more receivers observe a single physical source at the same time. Each receiver measures the arrival time of the signal from the source with respect to its local clock, and the two measurements are exchanged over a separate channel and subtracted. If the path delays between the source and the two receivers are equal, then both the delay and the characteristics of the source cancel in the difference. Common-view observations of the signals from navigation satellites will be discussed in this chapter, but the method is more general than that, and many signals have been used for common-view measurements.

An important advantage of the common-view method is that the transmitter need not "know" that it is being used for this purpose. (Compare the two-way method described above, which cannot support anonymous calibrations.) For example, signals from TV transmitters and the LORAN navigation system have been used for common-view measurements. The only important requirement is that the delays from the source to each receiver be as equal as possible and that it is possible to estimate any residual difference in the delays of the two paths.

Although the common-view method is simple in principle, there are a number of subtle considerations. If the delays along the two paths from the satellite to the receivers on the ground are exactly equal, then the delays cancel in the common-view difference, and the time of the satellite clock cancels as well. However, it is very difficult to realize this exact cancellation, and it is more usual for the paths to be only approximately equal. In this case, the differential delay makes a first-order contribution to the time difference, and this differential delay must be determined. This difference is typically determined from the known position of the receiver and the position of the satellite computed from the ephemeris. (The *differential* contributions of the ionosphere, the troposphere and the other effects that were discussed in the previous section must also be included.)

When the path delays are not equal, however, the signals that arrive simultaneously at the two receivers were not emitted at a single time by the satellite, so that the evolution of the satellite clock and the satellite position during this time difference must be computed. Conversely, if the common-view method makes use of a single signal from the satellite, then it does not arrive at the same time at both receivers, and the evolution of the receiver clocks during this time interval must be evaluated. Most common-view algorithms adopt this latter configuration and apply a time tag derived from the time of reception. The time of transmission of the signal received by each receiver must then be computed iteratively by computing the position of the satellite at the instant of reception, calculating the transit time, subtracting this time from the time of reception, and using the adjusted time to re-compute the position of the satellite. The satellite orbital velocity is about 4 km/s, so that the satellite moves about $4 \times 10^3 \times 0.065 = 260$ m during the transit time of the signal from the satellite to the receiver. The diurnal rotation of Earth implies a tangential speed of about 440 m/s at the equator, so that a receiver on the equator moves by about $440 \times 0.065 = 28$ m during the transit time of the signal from the satellite to the receiver. The exact change in the range depends on the details of the geometry. This iteration normally converges after one or two cycles.

A second consideration is determining which signal to use for computing the common-view difference. It was natural to use the ionosphere-free L3 signal in the previous one-way discussion because the L3 signal canceled the additional delay caused by the refractivity of the ionosphere, but it is not so clear that that is the correct strategy in the common-view method. As mentioned in the previous section, the L3 signal has more noise than the L1 and L2 signals from which it was computed. If the receiving stations are not too far apart, the use of L1 may be a better choice because the signals to both stations have the same ionospheric delay, which is going to cancel in the common-view difference, and there is no point in paying the price of constructing L3 when the effect of the ionosphere is going to be cancelled by common-view anyway. The same sort of argument applies to the troposphere, although the troposphere generally has shorter-wavelength variation so that it is less likely that the signals to the two receivers will have the same tropospheric delay unless the path length between the two receivers is quite short.

The choice between common view based on L1 and the common view based on L3 depends on the distance between the receiving stations compared to the wavelength of the variation of the delay through the ionosphere. This choice was not available to the first generation of GPS receivers, since they could receive only L1 signals. However, more modern multi-channel receivers can process both L1 and L2 signals simultaneously, and the choice between common view based on L1 and common view based on L3 can be made after the fact.

29.3.4 The "Melting Pot" Method of Common View

The common-view method discussed in the previous section depends on the fact that the receivers that are participating in the measurement can receive the signals from the same physical transmitter. When the signals from navigation satellites are used in common view, stations that are

sufficiently far apart cannot see any common satellite, and so the simple version of common view cannot be used.

However, the time signals transmitted by all of the navigation satellites in the same constellation are related to the system time of the constellation. A prediction of the offset between the satellite time and the system time is transmitted as part of the navigation message; this offset can be combined with a measurement of the physical time difference between the signal received from the satellite and the time of the local clock to compute a common-view time difference with respect to the system time rather than the physical time transmitted by the satellite. The result is a "logical" common view between the user clock and the system time, which is a common view with respect to a paper time scale that is not realized by any real clock in the constellation.

The usefulness of the melting pot method depends on both the accuracy of the calculation of the transmission delay from the satellite to the receiver and on the accuracy of the link between the physical time signals transmitted by the satellite and the system time scale. The time of the satellite clock with respect to the system time scale is computed by the ground controllers and is uploaded to the satellite periodically. Therefore, this relationship is generally a prediction rather than a measurement.

The stations that are participating in the melting pot method are generally not receiving signals from the same satellites, so that all of the path delay estimates discussed in a previous section must be computed by each station, and there is no cancellation of errors in these computations as there would be in simple common view of a single signal from a single physical source. The usefulness of the melting pot method is made possible by the increased accuracy of the position and clock solutions of post-processed ephemerides, but this has the disadvantage that these ephemerides cannot support real-time applications.

29.3.5 Physical Common View and Melting Pot Compared

A comparison between the two common-view methods depends on the distance between the two receiving stations on the ground. When the distance is very short, the two stations receive signals from the same satellites in the constellation, and so the two methods are effectively equivalent.

The physical common-view method computes the time difference satellite by satellite for all of the satellites that are in common view and then averages the resulting time differences. (Some outlier detection is often included in this average computation.) The melting pot method computes the time difference between the time at the receiver and the satellite system time and then averages those differences. (Again, some outlier detection is generally included.) Since the same satellites are used in the calculation by both stations, the offset between the satellite clock of each satellite and the system time cancels in the difference.

As the length of the baseline between the two receivers increases, the comparison between the two methods is a trade-off of two effects. The common-view method based on a common physical signal from a single source is less sensitive to the details of the transmitter and the characteristics of the path delay because a significant fraction of these parameters are common to the signals received by both receivers and cancel in the difference. On the other hand, receivers that use the melting pot method can make use of signals from more satellites in general, so that measurement noise will be attenuated provided that the orbital parameters and the relationships between the clocks on the satellites and the system time are known with sufficiently small uncertainties. For sufficiently long baselines, the melting pot method is the only choice.

The increased accuracy of post-processed ephemerides generally favors the melting pot method for all applications except those that require real-time calibrations. Neither of the common-view methods provides any attenuation of local effects – the calibration of the receiver delay, for example, and any error in the position of the antenna. The full advantage of either of the common-view methods is generally not realized in practice because these local effects dominate the error budget in both cases. A very important local effect is caused by multipath reflections, and will be discussed in the next section.

29.3.6 Multipath Reflections

The multipath effect is the result of the sum of the direct signal from the satellite and a copy of the signal that reaches the antenna after reflection from some surface that is generally close to the antenna. The reflected signals always travel a longer path than the direct version, and therefore always arrive afterward. The receiver combines the direct and reflected signals with the result that the measured time difference is systematically too large.

The offset caused by the multipath reflection always has the same sign, so that it cannot be removed by averaging the measurements. The magnitude is a complicated function of the geometrical relationship between the satellite, the reflectors, and the antenna; and it therefore varies with time as the satellite moves through the sky.

Figure 29.5 shows the importance of the multipath contribution. The data in the figure are the time differences, computed every second, between two receivers observing a single satellite for the entire time that the satellite was

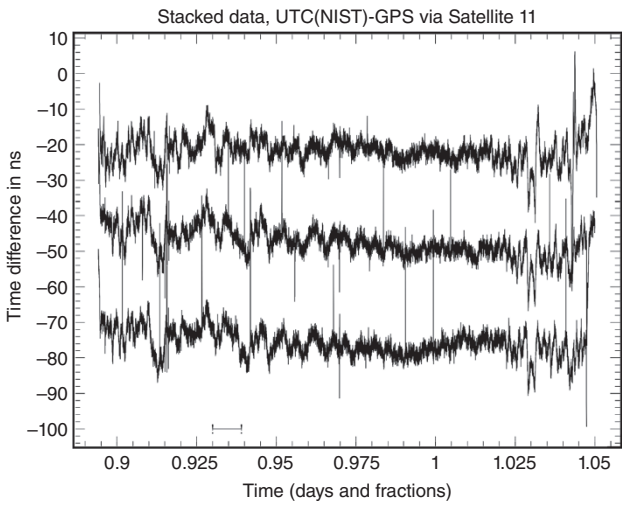

Figure 29.5 The short-baseline common-view time difference between two GPS receivers that were observing a single satellite and were connected to the same reference clock. The three traces were acquired on consecutive days and have been displaced vertically for clarity. In each case, the satellite is tracked from horizon to horizon. The scale of the vertical axis has not been altered. The time of the points in successive traces have been shifted left by 4 min as discussed in the text. The horizontal bar near the x-axis shows the length of the 13 min averaging time that is typically used to compute common-view time differences.

in view. The two receivers were connected to the same clock and the antennas of the receivers were a few meters apart on the roof of the NIST laboratory building in Boulder, Colorado. The three traces were obtained on consecutive days. The traces are offset vertically for clarity, but the vertical scale is not changed. The time of each trace is shifted by 4 min earlier with respect to the previous trace for reasons discussed below. The bar near the x-axis shows the duration of the13 min averaging time that is typically used by national metrology institutes and timing laboratories. The variation of the short-baseline time difference during the 13 min averaging time is significant.

The very strong correlation between the "noise" on consecutive days after the time tags of the data have been shifted by 4 min is a very strong indication that the fluctuations are due to multipath and are not a stochastic contribution at all. The significant increase in the magnitude of the variation near the ends of the traces, when the satellites are low in the sky, is also a characteristic of multipath reflections.

The p-p multipath contribution is several nanoseconds for the entire data set, and is significantly larger than this value near the start time and end time of the observing period, when the elevation of the satellite is relatively low in the sky.

It is clear that the effect of multipath is not canceled in common view, even when the antennas of the two receivers are close together at the same location. This is important for the relative calibration of a receiver, which is often realized by short-baseline common view as discussed below.

A "choke ring" antenna is an active-element antenna surrounded by a series of passive concentric rings that have a geometry that attenuates signals reaching the antenna from low elevations. The assumption is that a low-elevation signal probably is a reflection from a nearby object and should be attenuated. The dimensions of the rings in the usual choke ring antenna are designed to attenuate signals at the L1 frequency, but dual-wavelength choke rings, which are cut with an internal step near the bottom of the rings, will also attenuate low-elevation L2 frequency signals. The choke ring antenna is also mounted on a ground plane to block signals reaching the active element of the antenna from below. These signals, which are reflected from the material directly below the antenna, can be quite large, especially when the satellite is near the zenith.

An impedance mismatch at the connection of the antenna cable to the receiver results in an effect that is similar to multipath. The impedance mismatch causes a reflected signal to go back up the cable toward the antenna, where it may be reflected again and arrive back at the receiver delayed by twice the travel time through the antenna cable. It is very difficult to match the impedance of the cable to the effective input impedance of the receiver, since the front-end is often an active device with a complex effective impedance. This problem can be addressed by inserting an attenuator in the antenna cable. The direct signal passes through the antenna once, but the reflected multipath-like signal passes through the attenuator two additional times and is therefore differentially attenuated much more strongly.

One way of addressing the multipath problem caused by external reflectors near the antenna is to adopt an observation schedule for every satellite that advances by 236 s (roughly 4 min) every day as was done in the discussion of the data in Figure 29.5 above. This advance exploits the fact that the satellite returns to the same point in the sky with respect to the receiver with a sidereal-day period, so that the geometry that gives rise to the multipath effects also repeats with this period. This strategy converts the multipath effect at every point in the track to a constant offset that varies from point to point and is unique to each satellite that is observed. The International Bureau of Weights and Measures (BIPM) adopted this strategy in the design of the tracking schedules for the common-view time comparisons between timing laboratories that use signals from the GPS navigation satellites.

The 4 min advance of the track schedule for each satellite is a mixed blessing. It removes the variability of the

multipath effect, but it converts it to a systematic offset that is constant over short periods for each track but changes slowly as the long-period variation of the satellite orbit modulates the multipath correction. This slow variation in the apparent time difference between the two stations can be hard to separate from the long-period random walk of the frequency, as discussed in the previous section.

The sidereal-day advance that converts the multipath effect to a constant offset can be exploited in a different way. If the time differences between the receiver and the satellite clock are used to estimate the frequency difference between the clock in the satellite and the clock in the receiver with an averaging time of exactly one sidereal day, then the multipath contribution to the time difference, which repeats every sidereal day, will cancel in this frequency estimate. This estimate can be computed point by point as long as the satellite is in view with the hope that the multipath contribution will cancel in these differences even when the magnitude of its contribution is not well known. This technique is not very useful if the time difference is required, since the various frequency calculations have an unknown and varying multipath time offset. This method depends on the fact that the clock at the receiving station is sufficiently stable so that an estimate of its frequency over a sidereal day is a meaningful operation. For example, if the uncertainty of the time-difference measurements is on the order of a few nanoseconds, the local clock must have a stability on the order of 10^{-14} for an averaging time of one sidereal day to be meaningful. In a later section, we will show that this requirement restricts the method to a high-performance cesium clock or a hydrogen maser.

The multipath effect is sensitive to the details of the satellite position and the magnitude of the multipath contribution is different for different satellites observed at the same time. The magnitude of the contribution can be difficult to detect in methods that average the time-difference data derived from signals received from all of the satellites in view at any time for this reason. Both the standard common-view and the melting pot method typically report these averages, and the temporal variation of the multipath contribution to the average is generally attenuated by the averaging process and not easily detectable. However, the multipath contribution to the time differences is always a systematic effect, and a residual, site-dependent bias almost always remains. Since the periods of the satellites are not exactly one sidereal day, the multipath contribution of each satellite changes slowly with a period of several months, and this is likely to introduce a long-period variation in the effective calibration of the receiver, especially if the calibration is performed with a short-baseline common-view comparison as will be described in the next section.

29.4 Determining the Time Delay Through the Antenna and the Receiver

The time delay through the antenna and receiver is attenuated in the two-way method, and the delays at the two end stations will cancel exactly if they are equal. There is no need to know the magnitude of the delay, and this is a significant advantage for the two-way method. However, these delays enter directly when one-way data from a global navigation system satellite are used.

There are two techniques that are used to calibrate the delay through the receiving system. The two methods produce somewhat different calibration values. The difference is generally on the order of 1–2 ns, which will be significant only for the highest-accuracy applications, which use some form of the common-view method as described above.

29.4.1 Short-Baseline Common View

In this calibration technique, two receivers are connected to a common clock, and the antennas for the two systems are placed close together. Both receivers are used to measure the difference between the signals received from all of the satellites in view and the local clock. The difference between the two data sets is the relative calibration delay of the two receivers. The calibration obtained in this way is a relative calibration, but is adequate if the two receivers will be used in the common-view method, which depends only on the difference in the delays of the two receivers and not on the absolute value of either delay.

The receivers of the national metrology institutes and timing laboratories are calibrated with this technique, where the short-baseline common view is measured with respect to a “standard” traveling receiver. The result is the differential calibration between the receiver at each laboratory and the traveling receiver, which acts as a transfer standard. The absolute delay of the traveling receiver is not important, provided only that it is constant for the duration of the calibration campaign.

The main advantage of this system is that it calibrates the combination of the receiver, the cable and the antenna in a way that is as close as possible to how the receiver will actually be used. Although the antennas from the two receivers are placed close to each other, the multipath effect can be quite different as shown above. The multipath effect varies from one site to another, so that the accuracy of the differential calibration may be compromised if the multipath effect is significant. It can be difficult and challenging to realize a calibration whose uncertainty is on the order of nanoseconds for this reason.

29.4.2 Hardware Calibration

A second calibration technique uses a GNSS signal simulator to calibrate the response of the receiver. The cable can be calibrated at the same time, or its delay can be determined separately by time-domain reflectometry or an equivalent method. The delay through the antenna is usually determined by placing the antenna in an anechoic chamber and measuring its transmission characteristics by using a standard antenna driven from a simulator. An important parameter in this calibration is the stability of the delay as a function of the position of the transmitting antenna. This variation is often specified as the stability of the electrical phase center of the antenna as a function of the direction to the source.

This method has greater resolution because there is no noise in the measurement, and potentially greater accuracy because there is a known, accurate relationship between the signal that is used to test the hardware and the timing reference that drives the simulator, and there is no problem with local effects such as multipath. However, the real accuracy is uncertain because the system is not tested in the same way as it is going to be used. Most timing laboratories do not have a GNSS simulator, so that this method is not generally used, and most receivers are calibrated by the short-baseline common-view method described in the previous section.

29.5 Synchronization Strategies

We can combine all of the previous discussion to derive a number of synchronization strategies. We will first discuss the considerations that define each of the strategies, and then give a number of examples based on real data.

29.5.1 Do No Harm

The general goal of synchronizing a local clock by using data received from a remote GNSS system (or any other technique, for that matter) is to improve the performance of the local clock. In terms of the language of the preceding sections, we would use the data from the remote GNSS system if, and only if, the two-sample Allan variance of the received data was smaller than the corresponding free-running variance of the local clock for some averaging time. Since neither the statistics of the local clock nor the statistics of the data received from the GNSS constellation are adjustable, the primary output of any strategy is the optimum averaging time and the statistics of the local clock that can be expected. We will present a simplistic, somewhat artificial strategy to illustrate the idea in the next section. There may also be other considerations that we will discuss in the following sections.

29.5.2 A Simple Strategy

Suppose that the timing signal from the GNSS constellation can be characterized as white phase noise over a very wide range of averaging times. This is a simplifying assumption because it is quite unusual for the data to be this stable. With this assumption, the two-sample Allan deviation can be written in the form

$$\sigma_y(\tau) = \frac{R}{\tau} \tag{29.41}$$

where R is a proportionality constant. Next, assume that the time of the local clock can be measured by a local process with negligible white phase noise. With this assumption, the statistics of the local clock are driven by only the white frequency noise of its oscillator. Again, this is an optimistic assumption because it assumes that the local clock has no deterministic or stochastic frequency drift. With this assumption, the two-sample Allan deviation of the local clock is given by

$$\sigma_y(\tau) = \frac{L}{\tau^{1/2}} \tag{29.42}$$

where L is another proportionality constant. Eqs. 29.41 and 29.42 have different dependencies on the averaging time. Eq. 29.41 increases more rapidly than Eq. 29.42 at short averaging times, which means that the remote clock seen through the communications link is less stable than the local clock at these averaging times, since the time of the local clock can be measured at a local device. Conversely, at longer averaging times, the two-sample Allan deviation defined by Eq. 29.41 will be smaller than the corresponding deviation for the local clock, so that the remote clock is more stable. The two will be equal when

$$\tau = \left(\frac{R}{L}\right)^2 \tag{29.43}$$

and this specifies the minimum averaging time at which the data from the remote clock will begin to improve the statistics of the local clock. The optimum averaging time increases as the statistics of the remote clock seen through the channel become larger. The local clock is more stable than the remote clock for averaging times less than the value given in Eq. 29.43, and using the remote data to adjust the time of the local clock will degrade its statistics by adding channel noise. The local clock would be more stable if it were left alone.

29.5.3 Rapid Error Correction

An important consideration in some applications is the requirement to minimize the time during which an error in the local clock remains undetected. Since time errors are generally outside of the statistical framework, it is generally not possible to evaluate the optimal response to this concern in any quantitative way. It is certainly possible to use a time interval between measurements that is shorter than the optimal time interval calculated based on the considerations outlined above, but the statistical arguments suggest that adjusting the clock based on these measurements is counterproductive, since the local clock would have been more stable *on the average* if it were left alone. The strategy used in the NIST time service is to use a measurement interval that is about 30% of the optimal interval calculated by means of the considerations outlined in the previous section, but to not adjust the clock based on these data unless the time error is at least three times the two-sample Allan deviation of the clock and the measurement channel evaluated at this shorter averaging time. This is an administrative decision made more on the basis of experience than on hard statistical evidence. This method detects a possible time error approximately 0.1% of the time – about one measurement in a thousand. Classifying these measurements as time errors, as opposed to low-probability measurements that are consistent with the statistical characteristics of the data, is an administrative decision based more on experience than on a rigorous statistical justification.

All of these methods are less successful at detecting frequency errors because the time dispersion they produce can be difficult to separate from the measurement noise at short averaging times. The short averaging times, which are often used to detect time errors quickly, make the detection of frequency steps more difficult. This is especially troubling in real-time applications. The effect of the frequency error has already compromised the previous data by the time it has been detected. A change in the frequency drift, which is often seen in rubidium devices and hydrogen masers, is even more difficult to detect in real-time applications for the same reason.

29.5.4 Multiple Satellite Error Detection

Every one of the methods described so far requires the data from only a single satellite to compute a time-difference value. Since there are often many satellites in view at the location of the receiver, it is possible to compute multiple nominally independent estimates of the time difference. This technique is called a "Redundant Array of Independent Time Measurements" (T-RAIM) analysis. All of these calculations should give results that differ from each other by an amount that is consistent with the statistical estimate of the measurement noise. Any computation that differs from the mean of the others by a statistically significant amount is an indication of a problem, although it is not always easy to identify the source of the trouble.

The T-RAIM algorithm is most useful for methods that are based on the simple one-way time transfer or that use the melting pot version of common view, since these techniques are sensitive in first order to errors in the ephemeris or the clock model of each satellite. The simple common-view method attenuates these effects in the common-view subtraction, so that they are both more difficult to detect and less harmful to the calculation with this method.

29.6 Illustrative Data

Concepts discussed in the previous sections will be illustrated with examples of typical atomic clocks and time-difference data acquired at NIST in Boulder. As discussed above, the basis for all synchronization strategies is the free-running stability of the local clock oscillator.

29.6.1 Statistics of Atomic Clocks

Figures 29.6 and 29.7 show the Allan deviation and the approximate time dispersion for three atomic clocks: a rubidium standard, a conventional cesium standard, and a "high-performance" cesium standard. All of the devices are commercial instruments, and the values in the figures are derived from the manufacturer's literature. The time dispersion is calculated as $\tau \times \sigma_y(\tau)$. In general, actual

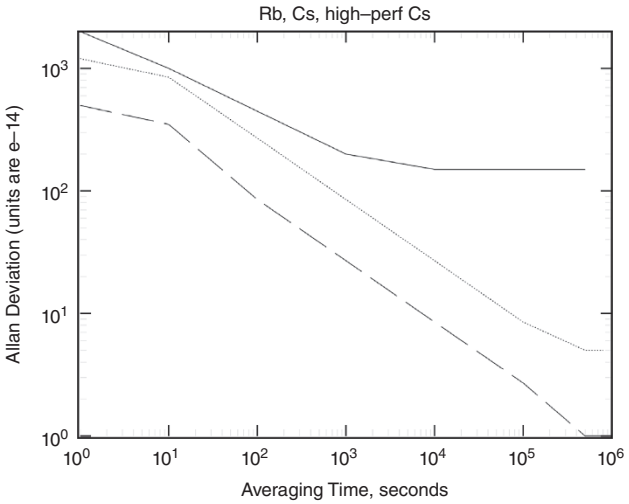

Figure 29.6 Two-sample Allan deviation of rubidium oscillator (solid line), cesium oscillator (dotted line), high-performance cesium (dashed line).

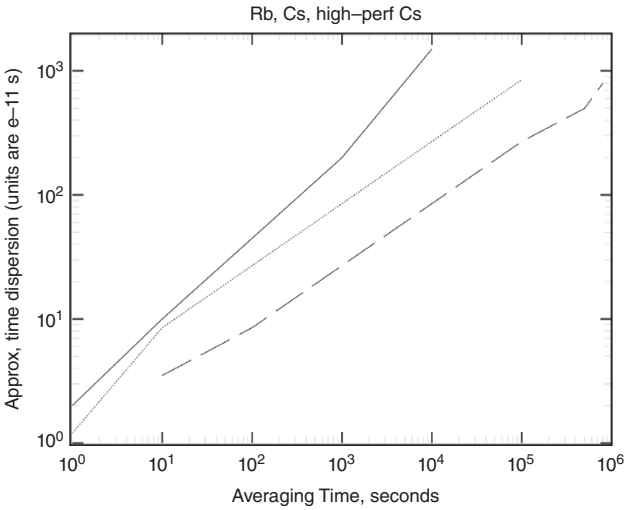

Figure 29.7 Time deviation of rubidium oscillator (solid line), cesium oscillator (dotted line), high-performance cesium (dashed line).

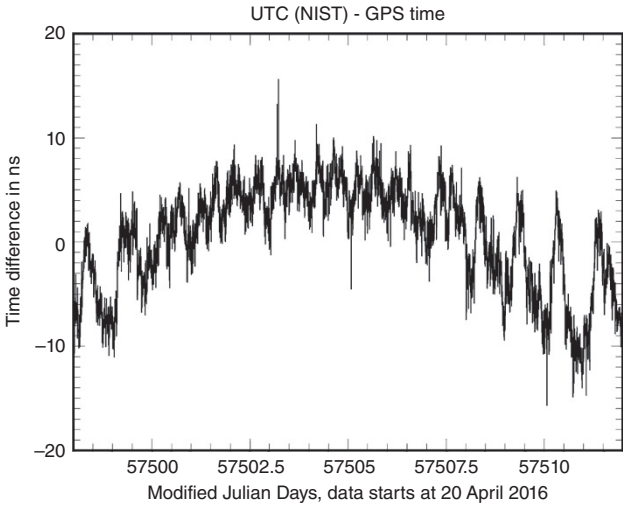

Figure 29.8 UTC(NIST) – GPS time, 5 min averages of the time difference estimated by using data from all of the satellites in view at each time.

devices are often more stable than the published specifications by a factor of about two or three.

The two-sample Allan deviation of both cesium devices varies as the reciprocal of the square root of the averaging time over almost all of the values in the figure, so that the stochastic noise of these devices can be characterized as white frequency noise at almost all averaging times (see Table 29.1). There is some evidence of the onset of flicker frequency noise at the longest averaging times of several days.

The statistics of the rubidium standard are much less favorable. The two-sample Allan deviation is larger, and the device has reached the flicker floor at an averaging time of several hours. The time stability is approximately 10 ns at that averaging time, which is adequate for many applications. The time stability of a cesium standard is a few nanoseconds even at averaging times of a few days.

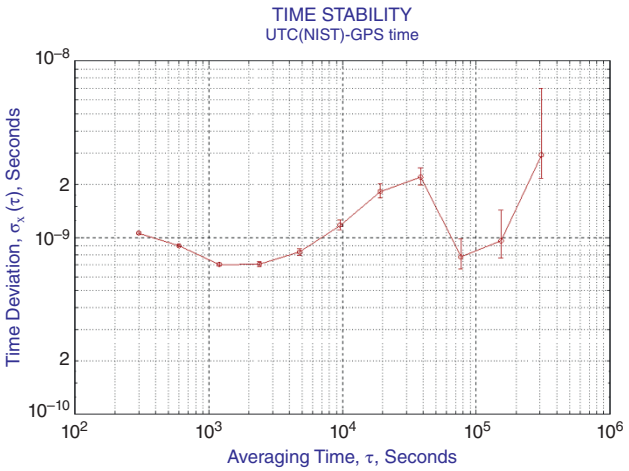

Figure 29.9 Time deviation of the data shown in the previous figure. The peaks near 20 000 s and 40 000 s suggest semi-diurnal and diurnal "bright lines in the spectrum (Eq. 29.36)".

29.6.2 Statistics of GPS Time

Figure 29.8 shows the average time difference between UTC(NIST) and GPS time. Each 5 min data point is the average time difference computed using the code data from all of the satellites that were in view at that time. The short-term fluctuations, which have an amplitude of about 5 ns p-p, are approximately flicker phase noise (compare to Figure 29.4a and Figure 29.4b) and are mostly due to the measurement noise of the GPS receiver. Figure 29.9 shows the time deviation of the data (defined in Eq. 29.32). The data in Figure 29.8 show clear diurnal and semi-diurnal periodic effects, and these are confirmed in the time deviation plot, which shows peaks at roughly one-half of the periods of these contributions (Eq. 29.36). The longer-period variation may be some combination of contributions from errors in the broadcast ephemeris or the difference between GPS system time and UTC(NIST).

To use these data to synchronize a local clock, we would compare the time deviation of the received data in Figure 29.9 with the time deviatioin of the free-running clock as shown in Figure 29.7, and the interval between calibration cycles would be chosen so that the stability of the GPS data as received is better than the stability of the local clock. For example, the cross-over would be at a period of a few hundred seconds for the rubidum device whose characteristics are shown in Figure 29.7. This algorithm would regard the diurnal and semi-diurnal variations as "signal"

rather than "noise," and the control loop would steer the local clock to remove them. This is a mistake, as will be shown below, but this cannot be recognized from these data.

The optimal algorithm for synchronizing a high-performance cesium standard is more complicated. There is an intial cross-over between Figure 29.9 and Figure 29.7 at a period of about 20 000 s, but the GPS data has a diurnal and semi-diurnal variation that makes its stability worse at periods from several hours to about one day. As with the algorithm to synchronize a rubidium standard, there is no way of identifying this variation as noise from these data, although the stability of the local device would support this assumption as reasonable, but not proven.

Figure 29.10 shows the common-view code time difference between cesium clocks at the NIST Boulder Laboratory and the NIST radio station WWV, which is located approximately 100 km northwest from Boulder. The figure shows the average common-view time difference computed satellite by satellite for all of the GPS satellites in view at each time. Figure 29.11 shows the time stability of the data in Figure 29.10.

These data confirm that a large part of the variation shown in Figures 29.8 and 29.9 is not due to the fluctuations in the local clocks or to other local effects such as multipath, since the variations are very significantly attenuated by the common-view subtraction. There is a residual peak in the time deviation at one-half of the diurnal period, but the magnitude of the time deviation is less than 1 ns for all averaging times greater than a few hundred seconds. Based on other data, this peak is most likely due to multipath reflections at NIST.

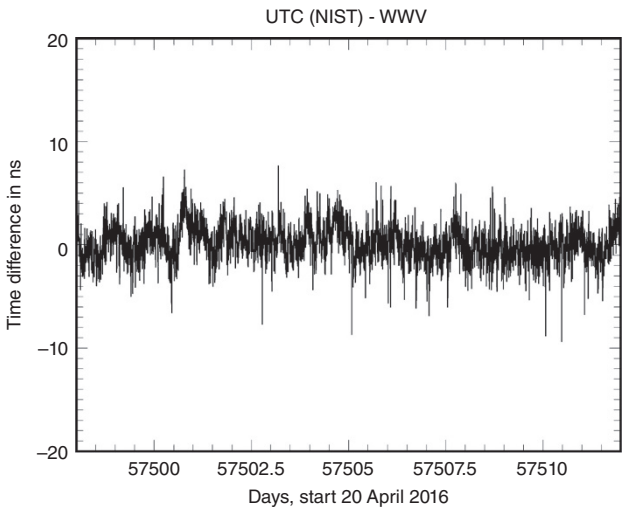

Figure 29.10 UTC(NIST) − cesium clocks at WWV 5 min averages of the time difference estimated by using data from all of the satellites in common view at each time. Note that the common-view method has significantly improved the variance of the data.

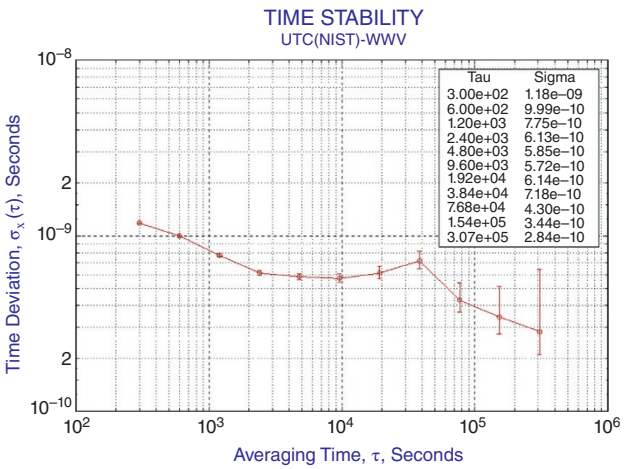

Figure 29.11 Time deviation of the data shown in the previous figure. The peak near 40 000 s suggests a diurnal "bright line (Eq. 29.36)". Note how the common-view method has improved the stability relative to the data in the previous figures.

The example of the common-view time difference between WWV and NIST, Boulder, illustrates the power of the common-view subtraction, but may be more favorable than average because the separation between the stations is only 100 km. However, most common-view station pairs can support synchronizing a high-performance cesium device after an averaging time of a day or two, and these common-view time differences are still used to compare the time scales of some national metrology institutes and timing laboratories. These time scales are generally realized as ensembles of high-performance cesium devices and hydrogen masers.

29.7 Adjustment Methods

In the previous section, the concept of the optimum averaging time was introduced, which was derived from a consideration of the two-sample Allan variance of the local clock and the two-sample Allan variance of the remote clock as seen through the communications channel. If we have decided that the local clock should be adjusted by a time T_a seconds based on the results of the calibration process, how should we apply the adjustment? There is no "best" answer to this question – the choice depends on a number of administrative considerations, which will be discussed in the next sections.

29.7.1 Correcting the Clock with a Single Time Step

This is the easiest solution to implement if the hardware supports a step adjustment. However, most algorithms use this type of adjustment only following a cold-start of

the control loop. There are three reasons for this. If the adjustment moves the time backward, then causality may be violated by a process that uses the clock time because moving the clock backward can reverse the time ordering of two events if the event that occurred later is assigned an earlier time stamp. If the time adjustment is positive, then some time values will not exist, and an application that is waiting for an explicit time to perform some action may wait forever because the time may never arrive. Finally, an estimate of the frequency of the clock based on the evolution of the time difference over some measurement interval will not give the correct value if the time interval crosses the adjustment instant.

29.7.2 Slewing Adjustment Rate

A slewing adjustment amortizes the time error by adjusting the frequency of the local clock. The frequency is raised if the time adjustment is positive and lowered if it is negative. The magnitude of the frequency adjustment is a compromise between time accuracy, which would favor a large frequency offset so as to amortize the time error over the smallest possible interval, and frequency stability, which would favor a small frequency offset to minimize the degradation of the frequency stability of the local clock. Frequency adjustments are normally applied as a single step in the frequency, but it is also possible to slew the frequency over some averaging time, which is equivalent to adjusting the frequency drift parameter. In either case, the length of the adjustment is calculated to produce the desired modification to the time of the clock

A frequency adjustment may be realized in hardware or in software, depending on the capabilities of the system. In the software system, which is used to discipline the clock in most computer systems, a physical oscillator generates ticks at a known rate. The operating system adds a constant to a register which then holds the accumulated time since the origin used by that computer operating system. The effective frequency of the clock can be adjusted by modifying the constant value added on each physical tick – increasing the value raises the effective frequency of the software clock and causes the time of the clock to become more positive relative to the reference time. Decreasing the constant value has the opposite effect. The change in this parameter is typical about 0.3%, so that it takes approximately 5 min to amortize a time error of 1 s. This adjustment rate, which is often fixed as an internal parameter in the operating software, is quite large, since it changes the effective frequency of the clock by a factor of perhaps 1000 relative to its free-running stability. This rapid adjustment rate implicitly assumes that the frequency stability of the clock oscillator is not as important as the time accuracy of the system.

At the other end of the scale, a national metrology institute generally steers its realization of UTC to agree with the paper time scale of the International Bureau of Weights and Measures. The frequency stability of such a time scale is important, and the *maximum* steering rate used at NIST is typically 0.5 ns/day, which is a fractional frequency adjustment of about 5×10^{-15}. A typical steering adjustment would be a fraction of this value, and the intent would be to amortize the time error over a period of about ten days. The frequency offset implied by these parameters is about the same as the free-running stability of the time scale, so that the steering correction generally is not observable by most users.

29.8 Summary

The stochastic variations in the time and frequency of clocks and oscillators are not well modeled by the usual statistical methods or by the standard statistical parameters such as the mean and the standard deviation. Instead, the two-sample Allan variance provides insight into the variation of the time and frequency and the source of this variation. The two-sample Allan variance can also provide an indication of the presence of a "bright line" in the spectrum of the variation. The power spectral density provides the same information as the two-sample Allan variance in principle, but the time-domain statistic is usually easier to interpret because most of the power in the Fourier computation is usually concentrated in the lowest few Fourier frequency bands.

A statistically optimal technique for synchronizing a clock by using data from a navigation satellite is to compare the two-sample Allan variance of the received data with the corresponding parameter for the free-running stability of the local clock. The statistical performance of the combination will include the best features of each contribution and so will generally be better than either contribution by itself. This is because the free-running stability of the local clock typically exceeds the stability of the data received from the satellite system at sufficiently short averaging times, while the opposite is true at longer times. These considerations emphasize the conclusion that it is almost always better to synchronize a local clock to a remote time source rather than to use the signals from the remote time source directly in the local application.

An algorithm that is based on statistics does not provide any estimate of accuracy. In addition, its estimate provides a root-mean-square error and has no information about worst-case performance. Therefore, statistical estimators are normally combined with some method of detecting

and rejecting outliers. The method used for outlier detection assumes that these data points are non-conforming errors rather than conforming errors, but very low probability statistical events. The detection of an outlier is therefore outside of a statistical discussion by definition.

References

The detailed definition of the two-sample Allan variance and a discussion of its statistical properties and other estimators derived from it can be found in the reprint volume:

Sullivan, D.B., Allan, D.W., Howe, D.A., and Walls, F.L., "Characterization of clocks and oscillators," NIST Technical Note 1337, 1990. Available from the publication database on the NIST Web page at tf.nist.gov as publication 868.

See also:

Howe, D.A. and Tasset, T.N., "Theo1: Characterization of long-term frequency stability," *Proc. 2004 European Time and Frequency Conference*, p. 5, 2004. Also available from the publication database on the NIST Web page at tf.nist.gov as publication 1990.

A bibliography of publications on the general subject of time and frequency measurement is

Hackman, C. and Sullivan, D.B., "Resource letter TFM-1, time and frequency measurement," *Am. J. Phys.*, Vol. **63**, pp. 306–317, 1995. Also available from the publication database on the NIST Web page at tf.nist.gov as publication 616.

The synchronization principles based on the Allan variance discussed in the text are also discussed in

Levine, J., "An Algorithm for Synchronizing a Clock when the data are received over a network with an unstable delay," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, Vol. **63**, pp. 561–570, 2016. Also available from the publication database on the NIST Web page at tf.nist.gov as publication 2769.

Levine, J., "An algorithm to synchronize the time of a computer to universal time," *IEEE/ACM Trans. Networking*, Vol. **3**, pp. 42–50, 1995. Also available from the publication database on the NIST Web page at tf.nist.gov as publication 1064.