Chapter #2

MEASUREMENT HARDWARE

2. OSCILLOSCOPES & CO.

2.1 A Short Look at Analog Oscilloscopes

Most engineers are already familiar with the functions and the operations of traditional analog oscilloscopes, and since those legacy instruments have very limited use in high-speed timing and jitter measurement, we will only have a very cursory look at them, and also cover only those properties and functions that have applications in those types of measurements.

Figure 20 shows a high-level schematic view of a typical analog scope. The input signal is fed into an amplifier (or, for high signal amplitudes, into an attenuator). The normalized signal is split up, with one part going directly into the display system where it is applied to one pair of electrodes of the cathode ray tube to deflect an electron beam vertically, and a second part goes into the trigger system. Of course it is possible that such an oscilloscope has more than just a single channel (two to four is common); in this case the user can select which channel shall provide the trigger.

The trigger system is a slope-sensitive comparator (slope sensitive meaning it can either react to rising edges only, or to falling edges only) with adjustable threshold level. Whenever it detects a trigger edge (i.e. an edge of the appropriate polarity that crosses the threshold), it starts a ramp generator that sweeps the electron beam over the screen horizontally. After the sweep the generator goes back to its initial state and waits for the next signal from the trigger circuit.

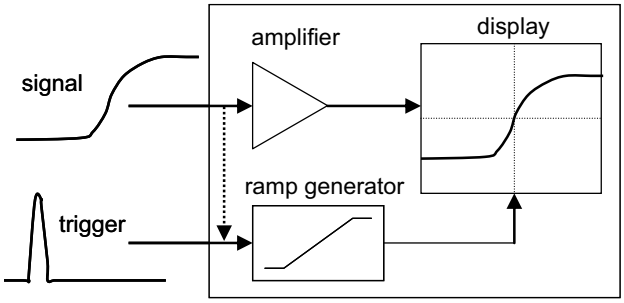


Figure 20: High-level block diagram of a traditional analog oscilloscope.

While the ramp generator sweeps the beam over the screen horizontally at a constant speed, the amplified input signal deflects it vertically in proportion to the signal amplitude at any given instant. Signal variations in time are therefore translated into different vertical positions along the curve. Wherever the electron beam hits the screen, the phosphor coating inside the screen glows for a short time, so the trace can be observed by the user. Normally the glow disappears after a very short time, so in order to obtain a steady image the signal has to be repetitive with a sufficient repetition rate (several Hz at least). The more often the beam hits a certain position on the screen, the brighter it glows. It is fair to say that in an analog scope the storage medium for the measurement is the electron tube's screen, although it is a highly volatile one.

The applicability of such an analog scope is limited by several factors: First, the deflection electrodes in the cathode ray tube have considerable capacity against each other, so do display very high frequencies – where those electrodes have to be charged and discharged very rapidly to follow the signal – very strong and very low impedance drivers are needed. Due to this effect it is very difficult to build an analog scope with a bandwidth exceeding 1 GHz.

Second, the signal information is solely stored in the glowing image on the screen (and that for a very short time), and cannot easily be translated into numbers and processed in a computer, e.g. to apply averaging, or to extract statistics on the edge positions for jitter analysis.

Third, it is difficult to capture very rare events because the screen will be dark most of the time. (One way around is to take a photograph of the screen with very long exposure time, but this requires minimum background glow, and the cycle of exposure – development – analysis is very slow).

Fourth, since the trigger event *starts* the sweep, we cannot observe what happened *before* the trigger – which may be important because it could tell us what led to some peculiar event. A way out of this is to use a passive analog delay line (a fancy word for a long, low-loss cable) to delay the

signal so the trigger comes with a sufficient head start before the signal arrives, but losses in the cable prevent us from making this delay very large (more than a few 10 ns). At the same time it is equally difficult to look at things a long time after the trigger, because the whole interval has to fit on the screen and so large time delays mean our timing resolution will suffer.

For all those reasons analog oscilloscopes have all but disappeared from applications geared towards high-speed, high-accuracy measurements, and today are mostly used in low-range troubleshooting situations where their comparably low price and simple and intuitive operation outweigh their restrictions. One advantage they hold is that as long as the trigger repetition rate is high enough, no high-performance hardware is needed to provide fast screen refresh rates (and so see even elusive glitches), while until recently especially low-end digital scopes often had rather slow screen refresh rates due to poor computing power, so many "old-timers" among the engineers prefer analog scopes since they feel these "really show the signal as it comes".

2.2 Digital Real-Time Sampling Oscilloscopes

With the advent of microprocessors, integrated semiconductor memories, and fast analog-to-digital converters starting in the early seventies, a new type of oscilloscope has over time become the mainstay in most engineering: the digital storage oscilloscope.

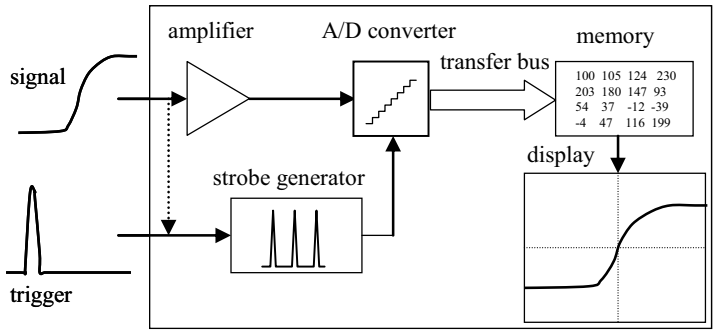


Figure 21: High-level block diagram of a digital sampling oscilloscope.

Looking at the general block diagram of such a scope (Figure 21) we can discern many similarities with an analog scope, but also a number of differences: Just as before, the incoming signal passes through an amplifier/ attenuator that makes its amplitude suitable for the subsequent stages. And

again some part of the signal is fed to the trigger circuitry. But here is where the similarities end, at least what concerns the internals.[25]

The first big difference is that the signal is not fed directly to a cathode ray tube but instead goes into an analog-to-digital converter (ADC). This component *samples* the signal at regular intervals and so translates the incoming (analog) voltage into a stream of binary (digital) numbers that get then stored in a fast acquisition memory. From there the numbers are read by the scope's microprocessor and – possibly after a lot of mathematical manipulation like averaging, scaling or more advanced operations – displayed as a waveform on the screen. Thus the screen no longer has the purpose of information storage – the display is more like a side effect rather than a crucial part of the scope's operation. (If we wanted, we could even read the digital waveform information into an external computer and process or store it there without ever displaying it).

Second, since the data is available in digital format (voltage vs. time in regularly spaced intervals), it is no problem to apply whatever complex mathematical processing to it – averaging, interpolation, edge searches, measurements of frequency, amplitude, rise times and so on, some of which we will discuss later in this chapter.

The humble trigger has also experienced a big jump in complexity. While in the analog scope all it did was to *start* the sweep (acquisition), it now more or less directly *stops* it (or, we could also say, it acts like a filter that decides which sampled data to keep and which to discard), as will become clear with the description below[26]. Figure 22 illustrates the process graphically:

1. The acquisition (sampling) process itself runs without interruption, and data is transferred continuously into the capture memory. When the end of the memory is reached, the storage goes back to the beginning, overwriting the oldest captured data, so we can see the capture memory as a cyclical buffer of a certain length (a few hundred to several million samples is typical).
2. When the trigger event is detected, the acquisition process continues to run for a certain time (see below) and then stops (this is what we meant with "the trigger stops the acquisition"). The captured data – also called a "record" – is transferred to the scope's main memory and the trace displayed on the screen.

[25] The manufacturers of digital oscilloscopes usually make every effort to hide those complexities and make it look and perform as close to an analog scope as possible as far as the basic handling is concerned.

[26] We should note that the details may vary depending on scope manufacturer and scope model, but the outline below gives a good idea of how it is done in principle.

3. The system is now ready to capture the next trace and continues with 1.

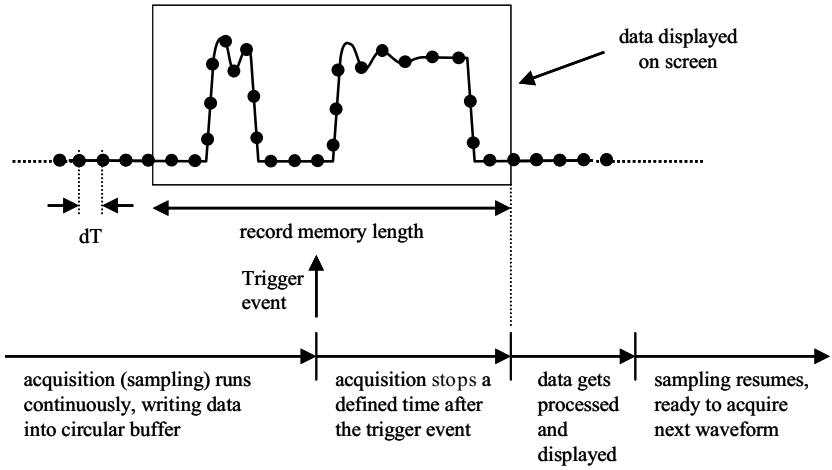


Figure 22: Principle of operation for real-time sampling.

We can select how long the system continues to captures after the trigger by appropriately setting the *trigger position*.[27] For example, if we set the trigger to the very beginning of the record, then the operation is identical to an analog scope – for the user the trigger seems to start the sweep (even though internally the acquisition has already been running before). This setting is also called "100% post-trigger". Usually this is the earliest position that the scope's software allows, even though in principle nothing prevents it from doing e.g. 200% post trigger, i.e. waiting even longer before it stops the sampling (in this case we would not see the trigger time on the waveform).

The other extreme would be to set the trigger position to the end of the record (i.e. "100% pre-trigger"), meaning the scope stops acquisition immediately when the trigger impulse arrives. In this case we see the waveform that happened before and all the way up to the trigger. This is something an analog scope cannot do. Finally, of course every setting between the two extremes is possible, e.g. 50% pre-trigger where the trigger position is in the middle of the record and we see some part of the waveform before as well as some part after the trigger point.

One weakness of this acquisition process is the time required to transfer the data record from the capture memory into the main memory, to process and display it. During this time no data is captured, and this time span may be many times longer than the time required to capture the record. In other words, we may only get a few snapshots of the waveform per second, which

[27] To keep things more comparable to analog scopes, this setting is usually called "horizontal position".

is not always obvious because the screen seems to get updated constantly (but our eyes cannot distinguish between 100 waveforms per second and 10000). If there are rare events (e.g. a glitch) in the signal, chances are high we will never see it at all, while an analog scope would show every single repeat and thus probably produce a faint but visible trace even of this rare event. Some of the highest-end oscilloscope designs seek to improve this situation by having dedicated data transfer and display circuitry (or don't update the screen at all during capture), so the main processor's performance does not get bogged down by those routine tasks. Some are able to capture hundreds of thousands of waveforms per second that way, which is close but not identical to a real analog scope in that respect. Even better, high-end digital scopes have very powerful trigger capabilities that allow triggering on specific features of the signal(s):

A big advantage of real-time scopes[28] as compared to analog scopes is their ability to capture single-time events. For that they also implement very sophisticated triggering schemes – on a good scope we can trigger on pulses that are lower than normal ("runt pulses"), too short ("glitches") or too long, are preceded by a specific bit pattern ("signature triggering") and so on. While important and extremely helpful for troubleshooting digital circuits, those modes are of less importance for standard timing and jitter measurements where the signal itself is usually repetitive, well defined and stable and we aim to characterize its properties with utmost accuracy.

A challenge for any real-time scope is to achieve the highest sampling rate (measured in GSamples/s) possible, so it can accurately capture even fast changing signal (more details to that later in this chapter). Today's leading-edge instruments achieve up to 40 GSamples/s, which by itself is quite an impressive feat, but even that means the interval between samples is 25 ps or longer, a lot of time with high data rates (keep in mind that 10 Gb/s means bit periods of only 100 ps, so even those fast – and expensive – scopes get only four samples per bit). It seems that real-time scopes have a hard time keeping up with those ever-increasing signaling speeds, especially now with the proliferation of ultra-high-data-rate serial transmission schemes.

Since the ADC has to acquire every sample in a very short time, there is not much leeway for it to settle out and obtain a highly accurate reading. Thus the resolution is usually limited to 8 bit (256 values, or a maximum resolution of just over 0.4%), and on some older models as few as 6 bit (64 steps). Even so the fastest scopes need to interleave two or more

[28] Referring to the previous paragraph, it seems that the "real-time" part of "real-time sampling scope" today has come to mean that it can capture a single waveform in a single shot, not that it necessarily processes or displays waveforms in real time as they come (which was the original meaning).

samplers (ADCs) to obtain the highest sample rates. Because of the limited vertical resolution it is important to make best use of it by adjusting the amplifier/attenuator so the signal covers close to full vertical (voltage) range of the sampler. Typically (on almost all scope models) this is the case when the displayed signal fills out the full vertical span on the display. But don't attempt to go further and have the signal exceed this range, no matter how little – this will most likely overdrive the input amplifier so it goes into saturation, and all bets are off with regard to waveform fidelity in this case! (That said, many scopes have designed in a small dynamic reserve of a few percent, but not more).

At the same time, to be able to capture long data streams and look at slow-changing effects even in fast data streams (that need high sample rates), the scope also needs a very large capture memory. The best ones available today thus come with many Megabytes worth of this memory, while lower-end scopes may max out at a few 1000 samples.

2.3 Digital Equivalent-Time Sampling Oscilloscopes

When dealing with very high frequencies (either fast data rates, or fast rise times, or both), normal real-time sampling scopes very soon hit two important barriers: First, they cannot increase their sampling rate beyond a certain limit, because the necessary data transfer rate into the capture memory gets unrealistically large, and even more important the ADC does not have enough time to settle between samples; speed and sampling accuracy/resolution are mutual tradeoffs. Second, as any other amplifier also the scope's input amplifier has some limited bandwidth – the best in class achieve around 12 GHz, i.e. enough to measure accurately signals up to maybe 3 GHz (or 6 Gb/s data rate assuming double-data-rate signaling).

So-called equivalent-time sampling scopes can often be employed for applications where the sampling rate and/or bandwidth of real-time scopes are insufficient. A fundamental difference is that the highest-performance versions do away with the amplifier, so the signal directly hits the sampler (if we ignore the termination resistor that is always present in those scopes to provide good signal integrity). This arrangement gets around the bandwidth restrictions of the amplifier for the price of largely reduced dynamic range because there is no way to scale the voltage range of the sampler. Analog bandwidths exceeding 70 GHz are possible[29], while the typical dynamic range is just around 1 V (normally this can be on top of some selectable offset of maybe +/-1 V).

[29] The highest-bandwidth commercially available sampling head that the author is aware of exceeds 100 GHz.

Of course physical limitations for the sampling rate still hold true, so those oscilloscopes do not even attempt to sample the signal in one sweep. Instead they rely on the assumption that the signal to be measured is repetitive, and they acquire only one sample per repeat and put them together to reconstruct the original waveform.[30] This process is illustrated in Figure 23.

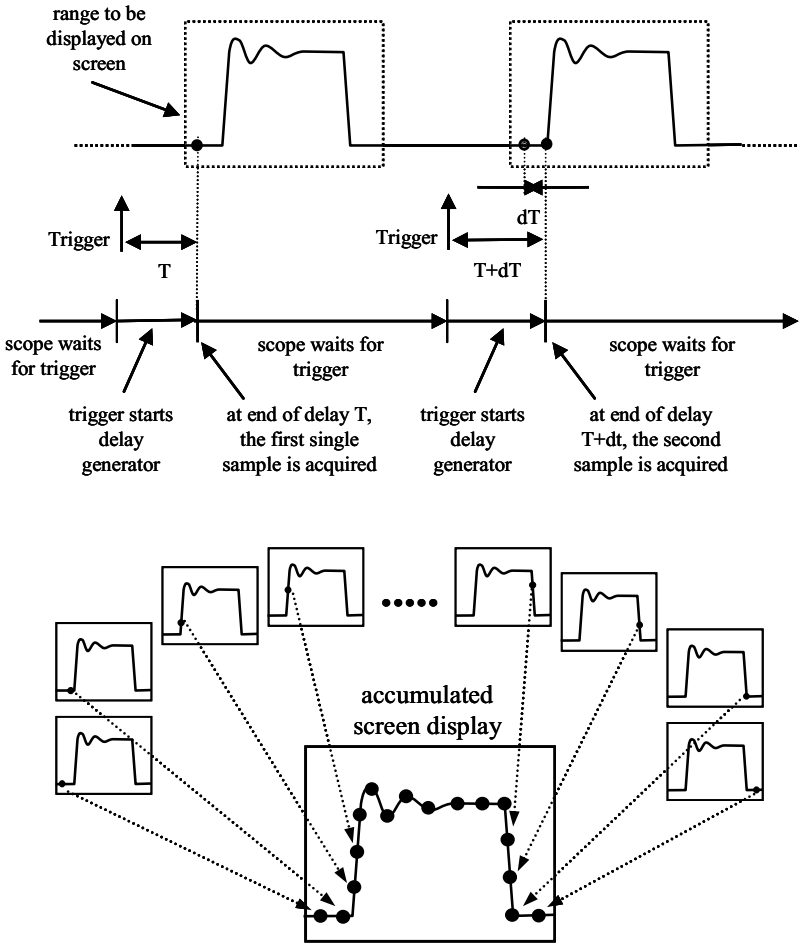


Figure 23: Principle of operation for equivalent-time sampling.

When the trigger arrives, the scope needs a short time (around 20 ns is a typical number) to set up the sampler which then acquires a single data

[30] Note that it is not absolutely necessary that the signal is *periodic*; it only has to repeat with a fixed relation between trigger and signal.

point. Then the scope waits until the next trigger arrives, the trigger is delayed a tiny bit longer and another sample is taken. Repeating this process many times yields a series of samples at increasing *equivalent time*s along the waveform, which is then displayed on the screen.

Since only a single sample is taken at once, all effort can be put into this, and so the sampler can achieve much higher resolution (up to 14 bits today compared to 8 for real-time scopes), which compensates somewhat for the lack of an amplifier for small signals, and provides unsurpassed resolution (and low noise) for larger signals.[31] The absence of the input amplifier also means that zooming into the waveform (increasing the voltage resolution) merely changes the *displayed* voltage range, but not the *actual* resolution (that is fixed by the total dynamic range and the number of sampler bits). So – unlike on real-time scope – the fact that the waveform exceeds the displayed range does not necessarily mean we are overdriving and saturating the input.

The *equivalent* sampling rate – i.e. the difference in delay from the trigger from one sample to the next – is only limited by the accuracy with which the scope can generate this delay. Resolution well below 1 ps is standard on today's scope models (compared to at least 25 ps and more even on the fastest real-time sampling scopes). On the other hand the *true* sampling rate – the number of samples acquired per second – is rather low, even the fastest such scopes don't exceed a few MSamples/s (most are in the kSamples/sec range). This becomes very visible if many waveforms have to be acquired (e.g. for averaging) or if the trigger does not comes very often (e.g. a frame trigger in a very long data pattern) – the lower of the maximum sample rate and the trigger rate determines the number of samples taken per second! Out of that reason the record length on equivalent-time sampling scopes is usually limited to a few 1000 points at best.

Another detail is that the input impedance into an equivalent-time sampling scope (and thus the load put onto the circuit under test) is virtually always 50 Ohm. True, one could use an active scope probe in front, but that – the probe being a bandwidth-limited amplifier circuit – would destroy the major advantage of such a scope – large bandwidth. The only exception may be when one needs to probe a differential signal since not every such scope on the market offers true differential inputs.

There is really no good technical reason to use an equivalent-time sampling scope for applications where the timing resolution, bandwidth, and timing accuracy of a real-time scope are sufficient, but equivalent-time sampling scopes are among the very few instruments suitable for today's highest data rate signals (above maybe 6 Gb/s) and for maximum-accuracy

[31] We can easily reduce the amplitude of signals that are too large with a passive high-bandwidth attenuator.

measurements. Among the tradeoffs are acquisition speed and flexibility. The fact that it takes some time for the scope to strobe the sampler after the trigger has been received means that just like an analog scope (and unlike a digital real-time scope) an equivalent-time sampling scope cannot directly show what happens at or before the trigger instant, unless a delay line is used to delay the data signal by more than the minimum data delay – but such a delay line will degrade the path bandwidth and thus negate at least part of the reason for using an equivalent-time sampling scope in the first place.

Because the market for such ultra-high bandwidth and accuracy instruments is still rather limited, and because of all the difficulties of designing such an instrument that exceeds the performance of the best real-time scopes, there are only very few test equipment vendors active in this area.

One final reason for using an equivalent-time sampling scope is that they are relatively inexpensive – they tend to cost only half or less for the same (or higher) bandwidth compared to a real-time sampling scope or a BERT box.

2.4 Time Stampers

Time stampers are fundamentally different from oscilloscopes in one important respect: Oscilloscopes are basically fast voltmeters, i.e. they provide the instantaneous voltage at given instants in time, or in other words, they measure the voltage when the time reaches a certain value. On the other hand time stampers do exactly the opposite: They measure the timing whenever the input voltage crosses a certain user-adjustable threshold. While the oscilloscope stores a series of such voltage-versus-timing measurements (either directly on the screen for an analog scope, or in memory in the case of digital scopes), the time stamper stores the sequence of timing numbers in its memory for subsequent analysis.

Figure 24 depicts a typical block diagram of such a time stamper (as usual, the architecture of a specific implementation may differ, but the example is intended to show the relevant principles). It consists of two slope-sensitive comparators (so one can choose to look at rising or falling edges only, respectively), a fast timing system consisting of a highly stable master clock and an interpolator, and storage memory to keep the results. The master clock – running maybe at 10 or 100 MHz – provides a stable time base, while the interpolator enables fine timing resolution (down to ps).

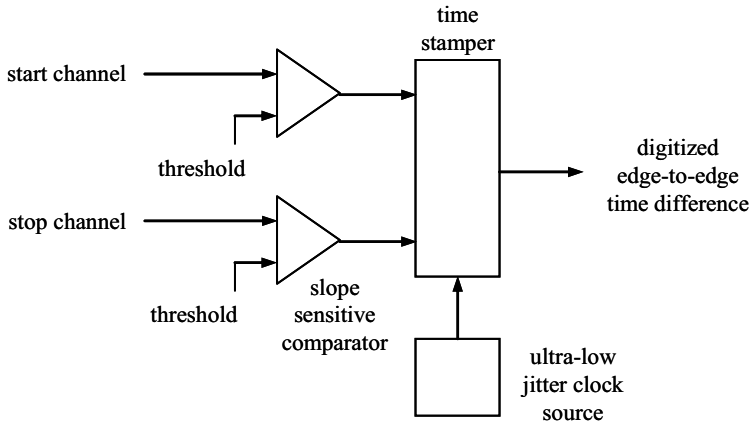


Figure 24: Block diagram for a typical time stamper.

An obvious question that arises is why *two* identical comparators are necessary: The comparators employed as well as the electronics to transfer the results into memory have only limited re-fire rates, i.e. after the comparator triggers it takes a while before it can trigger again. Those dead-times are often of the order of a few ns or more, much too long to be able to capture two subsequent edges of a fast data stream. Having two comparators allows having the second one waiting until the first has fired, so one can capture arbitrarily small delays between events.

There is a variety of timing numbers that we can get this way:

- Signal period: One comparator triggers on a certain edge polarity (rising or falling edge, respectively), the second triggers on the following edge of the same polarity.
- Pulse width: One comparator triggers on a certain edge polarity (e.g. rising edge), the second triggers on the following edge of the opposite polarity (falling in this case).
- Rise times: One comparator triggers on a certain edge polarity at the low threshold (e.g. rising edge, 10% of swing), the second triggers on the same edge but at a different threshold (e.g. 90%, so the timing difference gives the 10/90 rise time).
- Duty cycle: This is simply a combination of period and pulse width measurements.
- Phase noise (N-period jitter): Similar to signal period measurement, but now we measure the jitter between an edge and one several (or many) cycles later. A sweep of the number of cycles (e.g. ranging from one

cycle up to hundreds or even thousands), for each setting acquiring the N-period jitter statistics, gives the jitter trends versus the time delay or, in frequency domain, the phase jitter vs. frequency (inverse of time delay). Unlike an equivalent-time sampling scope a time stamper can handle jitter exceeding a bit period because it does not acquire after a specified time, but rather after a specified number of *edges* and so is immune to large *cumulative* edge movements.

Statistics over many subsequent such measurements can give values for timing jitter, as we will discuss later in this book.

As long as we are only concerned with *timing* measurements, time stampers have the advantage of acquiring exactly the data that we want and nothing more, while scopes always gather the full two-dimensional waveform but in principle we are only interested in a single point (or a few) on this curve, namely the threshold crossings. What's more, in order to achieve reasonable measurement resolution oscilloscopes must have very high sample rates, which complicates their design and increases the amount of (for pure timing measurements unnecessary) data acquired, slowing down data acquisition and processing. Of course scopes can do many things time stampers can't (e.g. signal noise, ringing, and overshoot measurements), because they provide more information (two-dimensional curves vs. one-dimensional timing values).

It seems thus fair to say that for troubleshooting and for comprehensive measurements scopes can't be beaten, but when it comes to speed of acquisition for pure timing measurements time stampers have a well-established niche. For that reason one finds them often as part of a large-scale digital production tester where speed of test and automation of the acquisition process takes priority. If they are standalone units, they are often called *time interval analyzers* (TIAs) but the functionality is largely the same. Leading-edge time stampers can acquire up to maybe 100000 timing values per second with minimal overhead, and they also offer true differential inputs, so measurements can be made on differential signals as well.

Apart from the fact that they do not measure voltage but timing, the design of time stampers puts an additional limitation on the capture of subsequent edges in a fast data stream: Due to the relatively slow refire rate a time stamper with two comparators can only acquire two edge timings, then there is a rather long break (at least a few ns, if not μs) before more edges can be captured. This runs the risk of missing any medium and long range effects. Some recent models of time stampers seek to improve the situation by using up to 10 comparators, but even this limits us to 10 subsequent edges. In that respect they are well inferior to real-time sampling

scopes (but not equivalent-time sampling scopes), which in a single shot can continuously capture waveforms containing thousands if not millions of edges. As a consequence time stampers have to make certain assumptions and some modeling if they want to determine jitter numbers decompose jitter into its components – in the next chapter, about jitter and jitter measurements, we will go into some more details about that.

A final limitation is that even the most recent time stamper models max out at an analog bandwidth of around 3 GHz. Remembering that to do meaningful measurements our bandwidth must be *at least* three time higher than the highest frequency of interest, this makes them usable for data rates of at best 2 Gb/s (for double-data-rate signaling) or clocks running at 1 GHz – and today's fast serial busses already exceed this range. Time stamper cards integrated into production testers normally don't even reach those bandwidths; they rarely even attain 1 GHz bandwidth.

2.5 Bit Error Rate Testers

In principle, when transmitting *digital* signals, all we *really* care about is if the digital information (the ones and zeros) makes it from the sender to the receiver without error, where an error would be a one received as a zero, or a zero received as a one. Maintaining signal integrity and clean waveforms is just a means to an end – the receiver shall be able to distinguish zeros from ones. No transmission is absolutely perfect and error free (see the discussion about random jitter later in this book), so usually all we can do is guarantee a certain maximum *bit error rate (BER)*. This BER is the number of "broken" bits (bits that get received incorrectly) to the total number of bits transmitted. This is where bit error rate testers (BERTs) come in.

BERTs basically consist of a fast data source as well as a fast receiver (level comparator, often with programmable threshold). The principle BERT setup is shown in Figure 25. Since BERTs originate from serial data transmission schemes, most of them have only serial drivers and receivers, but in principle nothing prevents us from building a BERT box with a parallel data source. What a BERT does is create a data stream, send it to the system under test, receive the transmitted data stream, compare it to the original data and count the number of failing bits – which gives the BER. Very often the data stream is some sort of pseudo-random bit stream (PRBS, more details to that in the next chapter) that assures all different possible bit sequences up to a certain maximum length are present in the data stream, but depending on the application the stream may also be any type of user-defined bit sequence, for which purpose most BERTs have a large built-in linear pattern memory.

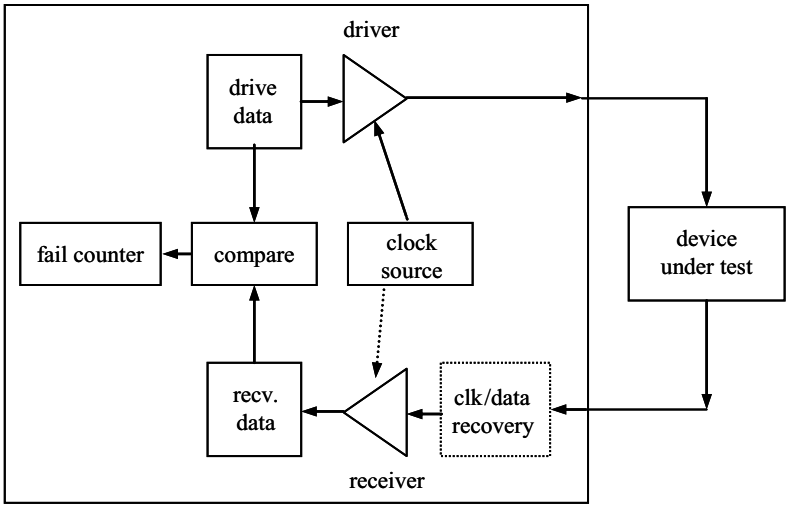


Figure 25: Block diagram of a bit error rate tester (BERT).

In the final application the receiver's strobe is likely to be placed at the center of the bit period (also called the *unit interval, UI*), but a BERT box allows us to move the strobe around. Without going into too much detail (we will dig into it more in a later when we look at jitter) it is clear that if we move the strobe closer and closer to the ideal position of the transitions, due to timing jitter the signal will more and more likely transition at the wrong side of the strobe, so the BER will increase. In that way a plot of the measured BER versus the respective strobe positions yields statistical information about the timing jitter of the signal averaged over many transitions, and thus – indirectly – edge timing information.

One big advantage of a BERT is that by design it captures pass-fail information from *every* bit even of a long data stream, so unlike e.g. time stampers or equivalent-time sampling scopes it has no dead times where it may miss failures, and it does so very fast (with the transmission data rate). So even if some effect only appears very rarely, e.g. for a very specific combination of bits, it will capture it as long as the bit combination is contained in the pattern.

On the other hand, a BERT only gives statistical jitter information, but no direct timing data. E.g. if a certain bit failed, we don't know how the failure looked like, not even if it just barely missed the right timing or if it was completely off. Also, while a single measurement (running the pattern once for a specific strobe position) may be relatively fast, the acquisition time for a full scan of the strobe timing over a bit period soon becomes excessively long if we demand good timing resolution (i.e. small increments of the strobe timing position). In addition, the receiver being a simple single-

threshold comparator, a BERT cannot give us much information in the voltage domain.

2.6 Digital Testers

Digital testers, be it for large-scale production test or for bench-top characterization work, are geared towards the *functional* test of devices with a large number of digital data pins. The basic emphasis in those machines is thus overall on sending and receiving digital bits rather than measuring analog waveforms and timings. Since for digital tests it is sufficient to determine if a signal is higher or lower than some threshold, but the absolute value is of no further importance at least as far as functional test is concerned, they employ so-called *comparators* to measure the incoming signals – identical to a BERT box with the difference that the latter rarely has more than one or maybe a few channels.

From a user's point of view one can see those comparators as simple switches that will yield "high" whenever the incoming signal exceeds the threshold and low whenever the signal is below the threshold (in many testers the comparator – from the user's point of view – seems to have *two* thresholds, where "low" means "lower than threshold 1" and high means "higher than threshold 2", but internally this is realized as just two single-threshold comparators in parallel). The threshold itself is user selectable (programmable), although very often only when no pattern is running. Also programmable are the timing during a test pattern run when the comparator shall *strobe* (i.e. transfer the comparison result to the digital capture memory). This is usually possible once within a device cycle period, e.g. every 500 ps when the test data rate is 2 Gb/s, and the results of those comparisons are captured in real time.

Such an input circuit is simple compared to the sophisticated sampling hardware that a scope requires, so it is possible (financially as well as complexity-wise) to have one comparator on each of the many channels. The comparators can be single-ended or differential depending on the design of the tester's pin electronics.

If we want to relate comparators to scopes (since they, too, measure voltages at specific instances in time), we can regard them as 1-bit A/D converters with adjustable threshold. A single bit is enough to distinguish between "higher than threshold" and "lower than threshold", but not more. The real-time strobe rate is sufficient to capture every single bit in a data stream, but much too slow to accurately resolve transition timings or detailed waveforms (at least in real time), last but not least because digital testers normally lack the sophisticated interpolation algorithms of oscilloscopes. But in the last chapter of this book we will see how one can nevertheless

accurately trace analog waveforms (i.e. emulate the functionality of an oscilloscope) and measure edge timing and jitter.

Since the input side of a comparator is just an analog electronic circuit (which does not care if its output gets sampled with 1 bit resolution or with 16 bit), it is of course subject to the same considerations regarding analog bandwidth and rise time as any oscilloscope. Because the main emphasis in the comparator design is usually on small, simple, inexpensive circuitry (after all, the tester designer needs to integrate hundreds if not thousands of those into his tester), not optimum waveform fidelity, the achievable accuracy is limited.

2.7 Spectrum Analyzers

A spectrum analyzers is quite a departure from all other instruments that we have looked ad, insofar as it does not measure the signal in the time domain (i.e. signal amplitude vs. time) but rather in the frequency domain (i.e. signal amplitude vs. frequency). Another way to see this is that it does a decomposition of the signal into a Fourier spectrum of sine waves of different frequencies and phases, and shows the amplitude (power, to be exact) of each of those sine components. Of course this spectrum is related to the signal in time through Fourier transformation, so in principle a spectrum analyzer can yield the same fundamental information as an oscilloscope.

Traditional spectrum analyzers are built as shown in Figure 26: The signal passes a very narrow-band bandpass filter[32] that filters out everything except a small band around its center frequency. Those filters are multi-pole filters (4 poles is common) and thus have a much steeper roll-off away from their center than a simple Gaussian filter. The pass-band can be as narrow as just a few kHz or less. The center frequency can be swept over a certain range with high resolution, and the spectrum analyzer shows the amount of signal that goes through for each frequency – resulting in the frequency (Fourier) spectrum of the signal.

[32] As shown in Figure 26, the actual design employs a mixer/downconverter that mixes the incoming signal with a locally generated swept-frequency signal. The mixing product contains both sum and difference components of difference and local frequency. A constant-frequency bandpass filter extracts the difference frequency. While technically very different, this whole contraption is functionally equivalent to a swept-frequency bandpass filter, but makes it easier to achieve the required filter characteristics because all subsequent stages only have to deal with constant-frequency signals.

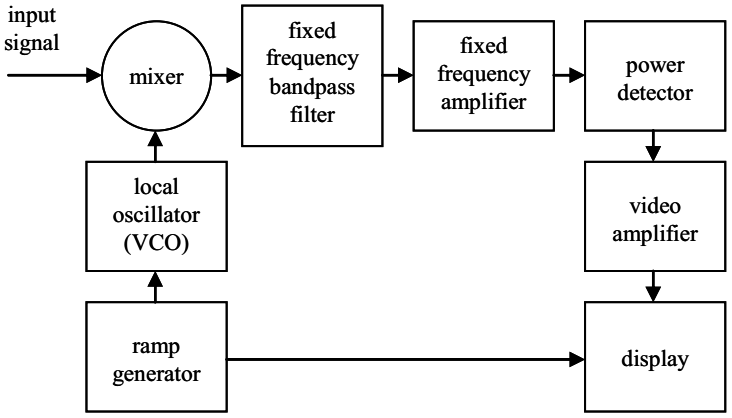


Figure 26: Simplified block diagram of a spectrum analyzer.

This architecture allows us to achieve huge frequency ranges (good high-end spectrum analyzers can cover a range from just a few Hz all way up to over 100 GHz) with amazing selectivity (filter bandwidth). This is difficult to match with an oscilloscope: On the low end a frequency of a few Hz corresponds to acquisition sets that must cover close to a second in time. High maximum frequency means the scope would have to do this with a very high sample rate. Both together result in unrealistically large data sets – no currently available high-end real-time scope has a capture memory of more than a few 10 MSamples, which limits the time span at maximum sampling rate (about 40 GSamples/sec) to less than a millisecond. At the same time the fact that a spectrum analyzer basically does a steady-state (DC) measurement of an amplitude enables unmatched signal-to-noise ratios (well over 100 dB is common) and thus measurement accuracy.

Unfortunately the loss of phase information (as only the amplitude is retained when passing through the filter chain) prevents us from fully reconstructing the original signal in the time domain, as already illustrated in Figure 1 (section 1.1). There is one class of spectrum analyzers that acquires the signal like a real-time sampling scope and then applies real-time FFT (Fast Fourier Transformation) to it. This preserves the phase, but on the other hand the acquisition architecture is the same as for an oscilloscope and thus holds no advantage over them regarding maximum or minimum frequency. Also their signal-to-noise ratio (SNR) is limited to maybe 40 – 60 dB, compared to 100 to 130 dB for "real" spectrum analyzers. Actually many of the higher-end real-time sampling oscilloscopes offer real-time FFT as a possible display mode.

Spectrum analyzers are great tools to track down and characterize periodic features of the signal (e.g. periodic jitter), and can also be useful and very accurate (because of their superior signal-to-noise ratio) in cases

where the phase information is not absolutely needed – one case being random jitter measurements. We will see more about that later in this book.

3. KEY INSTRUMENT PARAMETERS

3.1 Analog Bandwidth

When looking at a scope (or similar measurement device), we see that the first step in acquiring the signal is to deliver it to the digitizer (analog-to-digital converter, ADC). The part before the ADC is considered the *analog* portion of the signal path, while everything in the signal chain behind the ADC is considered the *digital* portion.

Being purely analog, any distortion on the signal will enter into the displayed result; as usual, we are mostly interested into the low-pass filter behavior that will limit the maximum frequency (minimum rise time) that can be delivered to the sampler – if the signal (or better: its high-frequency components) does not reach the sampler, then the best ADC is of no help – the signal is irreversibly lost.[33]

In the previous chapter we have learned that rise time and bandwidth are connected by the simple formula

$$T_r = \frac{k}{BW_{-3dB}}, \quad k \approx 0.33 \ldots 0.6 \tag{24}$$

Small k's indicate a rather smooth drop-off of the response with frequency, while higher numbers correspond to a faster, "brick-wall"-like drop-off. Examples for the former are Gaussian and simple exponential (single-pole) filters, while filter types like Chebyshev and Butterworth are representative of the latter category. Figure 27 shows some typical filter response curves.

Many commonly used formulas and rules of thumb assume Gaussian filters – one example being the addition of rise times (or inversely the calculation of the true input rise time when the system rise time is known). Unfortunately, modern digital oscilloscopes rarely behave Gaussian, but instead have a steeper drop-off beyond the 3 dB point. One reason is that in

[33] As long as at least some portion of these components is preserved one can theoretically re-create the full signal numerically through digital signal processing (see the last chapter in this book), but even this method reaches its limits rather soon with increasing attenuation.

order to avoid aliasing[34] the sampler must not get any frequency components that exceed half the sample rate, so the analog front-end strives to filter those components out. The second reason is that a Gaussian response – with its gradual drop-off not only beyond but also before the 3 dB point – would mean that we have considerable attenuation even well below the 3 dB bandwidth limit, which is definitely not what we want because it introduces large amplitude errors even for moderate rise times.

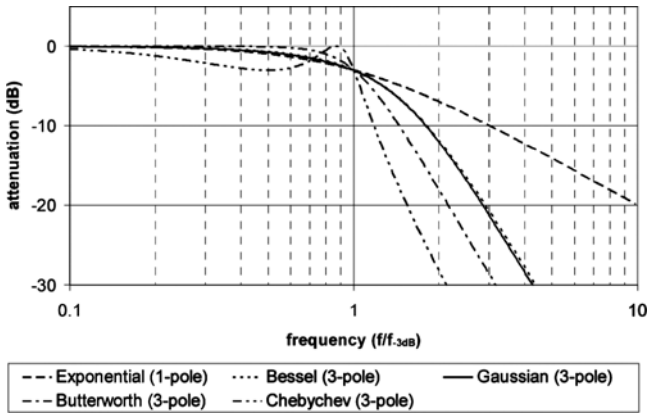


Figure 27: Filter responses (transmission loss vs. frequency) for several different commonly used filter types.

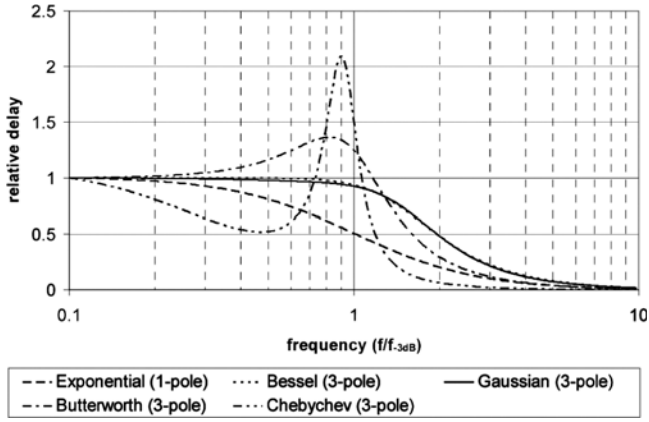


Figure 28: Filter delay vs. frequency (i.e. dispersion) for several different commonly used filter types.

[34] Aliasing means high frequency components of the signal get "folded back" into the range below the sampling bandwidth, causing distortion and artifacts in the sampled waveform.

At first glance the solution to both problems seems to be to implement one of the "brick-wall" filter types. Unfortunately, while they look tempting in frequency domain (where they have indeed wide application), such filters don't behave nicely in time domain: they exhibit large ringing that takes a long time to settle, as shown in Figure 29. They also have a fair amount of dispersion, meaning signals of different frequencies experience different delays through the filter (see Figure 28). An edge (consisting of a broad spectrum of frequency components) will get "washed out" because some components take longer than others to traverse the filter; this will also cause edges with different rise times to experience different delays, causing timing errors. Thus the scope designers have to make some tradeoffs between high-frequency filtering, flat attenuation curve below the 3 dB bandwidth, minimum dispersion (timing delay change with frequency), and clean time-domain response. The end result is a filter with k of around 0.4 – a bit closer to a Gaussian filter than to Chebyshev of Butterworth.

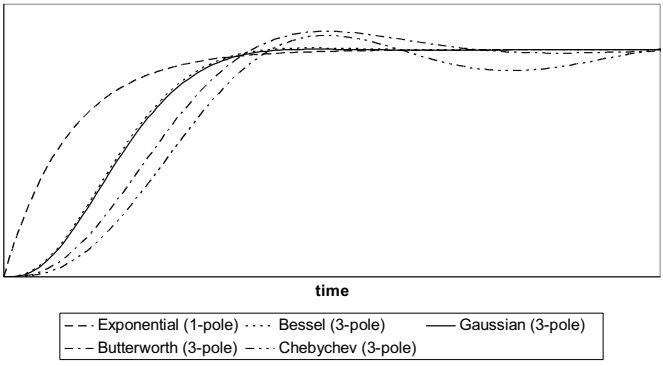


Figure 29: Filter responses in the time domain for several different commonly used filter types.

This deviation from Gaussian behavior means we have to be somewhat careful when applying our rules of thumb; they are *approximately* valid and work as long as the corrections are not too large. E.g. let's consider a scope with 1 GHz analog bandwidth, which measures a signal edge; assume that the 10/90 rise time we see on the screen is 1.1 ns. What is the true rise time of the signal going into the scope?

We know that rise times add up as squares (exactly valid only as long as all edges and filters are Gaussian), thus

$$T_{r,true} \approx \sqrt{T_{r,meas}^2 - T_{r,scope}^2} \approx \sqrt{T_{r,meas}^2 - \left(\frac{0.4}{BW_{r,scope}}\right)^2} \qquad (25)$$

Putting in the numbers we get

$$T_{r,true} \approx \sqrt{(1.1\,\text{ns})^2 - \left(\frac{0.4}{1\,\text{GHz}}\right)^2} \approx 1.025\,\text{ns}\,.$$

In other words the calculated correction is of the order of 7%, which qualifies as "small", and we can assume that even with the non-Gaussian behavior of the scope the result is reasonably accurate. But don't try this in cases where the scope rise time is of the order of (or larger than) the signal rise time!

3.2 Digital Bandwidth; Nyquist Theorem

A second limiting factor in the acquisition process is the sample rate. Keep in mind that instrument "bandwidth" (or "rise time") is simply a measure of how fast the instrument's acquisition system can react to a signal level that changes in time. It does not make any restrictions as to *what* exactly causes this limitation in reaction time.

If a signal makes a sudden jump between one sample and the next (e.g. at one sampling point it is still low, and at the next sampling instant it is already high), then the scope has no way of knowing what exactly happened between the two sampling instants – it only knows the original and the final state. Any signal that rises faster that the sample interval will result in the same sampling result, no matter if it rises within $1/10^{\text{th}}$ of the interval or if it takes the full interval to rise. So the scope has to make some assumption about the interval between them, and the simplest approach is to assume that the signal just rose linearly between those two points. This is illustrated in Figure 30(a). There we can also see that there is an even worse case where it looks to the scope as if the signal actually needed *two* intervals to rise. So in other words, from Figure 30(b) we can deduce that without any further data processing (interpolation) the limited sample rate is equivalent to an effective rise time (or bandwidth) limitation somewhere between 0.8 and 1.6 times the sample interval:

$$T_{r,dig} = (0.8 \ldots 1.6) \times \Delta T_{sample}\,, \tag{26}$$

or to a so-called *digital bandwidth* of approximately

$$BW_{dig} \approx \frac{0.4}{T_{r,dig}}\,. \tag{27}$$

This is a serious limitation for any real-time sampling scope where we cannot increase the sample rate above a certain limit (40 GSamples/sec for the best scopes available today). Another conclusion from above formula is that our sample rate must be at least twice as high as the highest frequency component of interest. All this is of much less concern on equivalent-time sampling scopes because there the (equivalent) sample rate can be made almost arbitrarily small (well below 1 ps on the best available scopes, corresponding to a *digital* bandwidth of over 400 GHz, which is one of the main reasons to use this type of instrument).

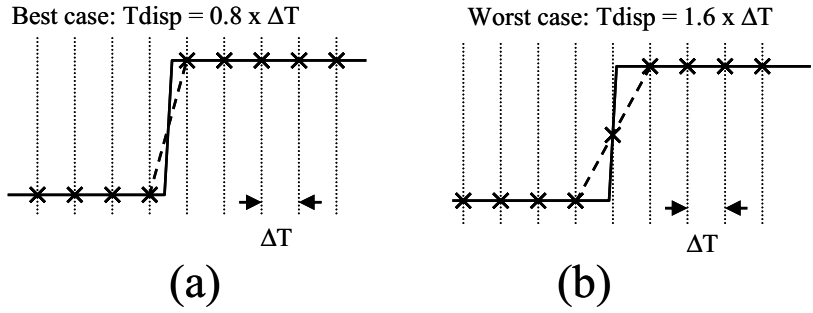


Figure 30: Rise time limitations caused by limitations in the sample rate (digital bandwidth): (a) best case, (b) worst case.

As a matter of fact above considerations were a bit "cavalier" on the impact of frequency components in the order of (or higher) than the sample rate. Mathematical theory of sampling – the well-known Nyquist theorem – gives a bit more stringent conditions: According to this theorem the sample rate must be *more than twice* the highest frequency component in the signal (the *Nyquist* frequency).

The theorem stems from an effect that is called *aliasing*. This effect is easily illustrated in Figure 31: It shows a sine wave of 1 GHz, sampled at a rate of 1 GSamples/sec and 1.5 GSamples/sec, respectively. Evidently the sample points at 1 GSamples/sec always fall on the same spot on the sine wave, and the reconstructed signal has constant level. The slightly faster sample rate (1.5 GSamples/sec) falls on different places on the curve, so the reconstructed signal is not a steady DC signal but varies in time, but with an apparent frequency of only 0.5 GHz. Those two cases illustrate a general principle that is dealt with in detail in the mathematical theory of Fourier transformation: Let's assume a signal with some frequency spectrum (harmonics) that is sampled at some sample rate. If we now reconstruct the signal from the sampled points (as done in the upper half of Figure 31), we will find that the signal has been distorted (through the low-pass filter effect). But what's even worse, the frequency components exceeding half the

sample rate are not simply gone, they get "folded back" (mirrored) into the frequency spectrum, the mirror being the Nyquist frequency (equal to half the sample rate): In our case the Nyquist frequency for 1 GSamples/sec is 0.5 GHz, so the 1 GHz signal mirrored into a 0 GHz signal, i.e. a DC level, just as we see in Figure 31. For the case of the 1.5 GSamples (Nyquist frequency 0.75 GHz), the 1 GHz signal is mirrored back into a 0.5 GHz signal. This is also indicated in Figure 31. What makes this effect so devilish is that after sampling, if we take the 1.5 GSample case, a mirrored (aliased) 1 GHz signal becomes absolutely and perfectly indistinguishable from a "true" 0.5 GHz signal – they both look like 0.5 GHz. No amount of digital signal processing can reverse this!

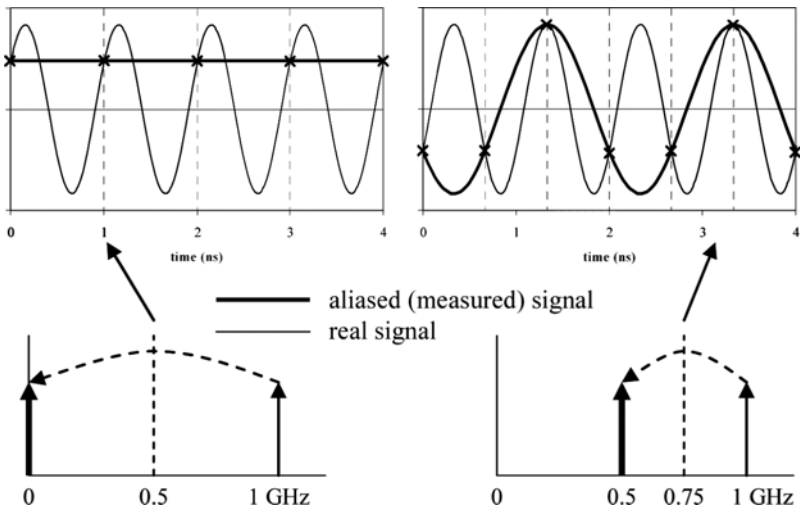


Figure 31: Aliasing caused by undersampling (insufficient sampling rate): The signal frequency component gets mirrored at the Nyquist frequency (identical to half the sampling rate). Upper half: time domain representation, lower half: frequency domain representation.

As a consequence, in order to avoid these effects one must thus make sure that no frequency component higher that half the sample rate gets to the sampler in the first place. In other words, some *analog* signal conditioning is indispensable before the conversion of the signal into digital numbers. Oscilloscope designs do that by matching the analog bandwidth to the available digital bandwidth, so the analog portion provides the necessary filtering. It should also be noted that the limit of 2 (sample rate equal to twice the Nyquist frequency) is not attainable in practice. Realistic numbers for good oscilloscopes are in the range of 2.5 to 3. And indeed, if we take a look at different oscilloscopes, we will usually find out that the sample rate

is approximately three times the analog bandwidth or more – and now we know why![35]

3.3 Time Interval Errors, Time Base Stability

Whenever we talk about timing, we mean the position in time *relative to some reference*. There just isn't anything like an absolute time (like "noon" or 3:25pm) – our references tend to be transitions on some trigger or clock channel, or maybe some specific edge in our data stream. As a result, what we really measure and display are really time *intervals* between two events (reference event and event of interest). Thus the important figure of merit for any instrument is how accurate it can measure those intervals.

Every oscilloscope, BERT box, time interval analyzer etc. acquires the signal based on some internal timing reference (often called the *time base*). Any inaccuracy in this time base will show up directly as an inaccuracy in the acquired timing information. From a high-level point of view we can separate those inaccuracies into two major types of errors: random and deterministic. Random errors will be different every time we repeat the measurement. Deterministic errors will always be the same as long as the test conditions did not change, and in the present context they are commonly referred to as *time base nonlinearity*[36]. We can compare this to measuring distances with a ruler that has slightly inaccurate markings. Those inaccuracies will show up as deterministic measurement errors, e.g. when we measure the length of an object that in reality is 1 m long, we may always get 1.01 because the 1 m marker is off. But at the same time our limitations in being able to see the exact marking may result in readings of 1.011, 1.008, 1.013 etc. if we repeat the measurement several times – these are random errors.

Achieving maximum time base accuracy is of course of utmost important to us, but it virtually always become more difficult the longer the intervals become. This is why most manufacturers specify the time base of their instruments as something like $\text{time base error} = \text{A} + \text{B} \times \text{interval}$. Apart from choosing an instrument with good specifications for time base accuracy;

[35] One exception are lower-end oscilloscopes that sometimes offer excessive sampling rate compared to their analog bandwidth. This is because in the range they are working digital performance is cheap and oversampling eases the requirements on signal processing (interpolation), making for a cheaper overall design.

[36] We should note that it can depend on the specific measurement setup and the specific conditions if a certain deterministic *instrument* error shows up as (pseudo-)random *measurement* error or as a deterministic (constant for constant conditions) error. Also, in jitter analysis "deterministic" is usually assumed to stand for "non-Gaussian statistical distribution", while here we use it to denote the fact that the error will not change if we repeat the measurement.

special measurement methods can reduce the time base error even further. In the last chapter of this book we will look into this more closely (for the case of equivalent-time sampling scopes).

4. PROBES

4.1 The Ideal Voltage Probe

When dealing with high-speed digital signals, virtually the only probe types in use are voltage probes[37]. The probe enables us to connect our measurement instrument to the system under test. In this function we would of course like it to be as close to invisible as possible to our signal, while enabling reliable access to the probe points in our system. In other words, an ideal probe should have the following characteristics:

- It has zero rise time (i.e. infinite bandwidth).
- It does not distort the measured signal, i.e. it has zero (or at least constant) losses over the whole frequency range, and its time delay (group delay) is constant over the whole frequency range.
- It has infinite amplitude (voltage) range.
- It does not add noise to the signal.
- It puts zero load on the system under test, i.e. we won't influence the system behavior when probing it.
- Its size matches the available probing possibility (which for high-speed applications and modern fast devices usually means it has to be very small).
- It is mechanically rugged and reliable.
- Last but not least, it should cost as little as possible.

[37] We deliberately limit ourselves here to electrical signals. It is true that some of the highest-speed transmission systems use optical (light) signals, but no oscilloscope or tester can deal with such optical signals anyway. They all have to get translated into an analog voltage signal by the input stage first (in principle just a fast photo diode or photo transistor). Actually there are a few higher-speed current probes available, but the achievable bandwidth is much lower than for voltage probes – a few 100 MHz at best for the former compared with over 10 GHz for the latter. What's more, on a controlled-impedance transmission line (e.g. 50 Ω) in absence of reflections (i.e. matched termination at the end) there is always a direct correspondence between voltage and current, so measuring the voltage automatically provides the current profile as well.

Unfortunately, no single real-world probe can fulfill all those requirements perfectly at the same time, but it is often possible to get very close to ideal for a few of them. Every probe type we are going to look at in the following sections represents a compromise between all those competing goals.

The first major choice that we face is between active and passive probes.

4.2 Passive Probes

Like the name implies, a passive probe consists entirely of passive elements (resistors, capacitors, inductors, cables, etc.). As such, lacking any means of amplification, it can never provide more than the original signal amplitude to the measurement instrument. Since all the energy has to come from the original signal, the main tradeoff here is between measurement signal size and the load on the system under test[38]. On the upside, there is no hard limit for the maximum voltage such a probe can handle, the circuitry is simple, and it does not need any power supply, all this making it a very inexpensive solution. Being purely passive, it also will not add any random jitter or noise to the signal (at least as long as it does not pick up external fields, but we will discuss that later).

A passive probe can be as simple as a piece of cable. The effective load on the system depends strongly on how the path is terminated in the oscilloscope[39]: Until several years ago, when the maximum frequencies in digital systems were low (a few MHz at best) and many device drivers were not designed to drive a 50 Ω load, the scope input was high impedance, and everybody assumed the load to be the purely capacitive (using the total capacitance of the cable). But as we already know this completely neglects the distributed characteristics of this cable capacitance as well as its inductance, and as a consequence delay, reflections and ringing. Once the signal bandwidth becomes high enough, transient effects (reflections at the unterminated end) would completely dominate the picture. Overall, at today's data rates the scope always needs to terminate the line to 50 Ω to avoid reflections, and the load then becomes a constant 50 Ω (and *not* capacitive, except for small parasitics that we will talk about later). This may be fine in a device characterization setting where all the driver needs to drive is the scope, but it will wreak havoc if we attach such a "probe" to the

[38] The larger the load, the more the system behavior will be influenced by the presence of the probe, which is normally not a desired property.

[39] For simplicity we will name only oscilloscopes, but all that is said implicitly applies to any other instrument (BERT, spectrum analyzer, production tester) as well.

middle of a data bus – chances are the additional load (and the impedance mismatch created by it) will make the data transmission fail immediately, not even considering the fact that this will not show us what the signal really looks like without the probe attached.

One solution is to increase the probe impedance by adding a resistor at the probe tip so it acts as a voltage divider (made up by the resistor and by the characteristic line impedance). Ratios of 1:10 or 1:20 are typical (i.e. 500 Ω or 1 kΩ load impedance, respectively[40]). In the simplest incarnation the divider is a single higher-impedance resistor in combination with the 50 Ω cable/scope impedance. While this does reduce the load on the system under test, it also greatly reduces the signal amplitude transmitted to the oscilloscope, so the scope's noise has a larger impact (reduced signal-to-noise ratio), which will make it difficult to observe and measure very small signals. What's more, large impedances make the probe extremely sensitive to any parasitic capacitance (remember that the time constant is $R \times C$ and the bandwidth is inversely proportional to that). The resistor needs to be placed right into the probe tip; otherwise the transmission line stub from the probe tip to the first resistor will cause reflections and ringing. The residual stub plus some unavoidable fringe fields at the tip inevitably add some parasitic capacitance to the load presented by the probe – the smaller this capacitance is, the better (more to that later). This capacitance is usually given in the manufacturer's specifications for the probe.

It is easy to calculate that the impedance mismatch created by an additional 500 Ω load placed on a 50 Ω transmission line will cause approximately 5% reflection and will reduce the signal amplitude transmitted to the receiver by almost 5%. It depends on the particular case if such an impact is acceptable or not.

That said, passive voltage-divider probes can perform well up to several GHz, and their simplicity makes them a valuable tool for low- and midrange applications. If the bandwidth requirement is not very high, it is easy to build such a probe out of a small surface mount resistor soldered onto the tip of an SMA connector.

4.3 Active Probes

Putting an amplifier – i.e. an active element – in the probe head is clearly a way to overcome the loading-vs.-signal-amplitude tradeoff. Well designed

[40] For 500 Ω probe impedance, the necessary series resistor is 450 Ω. In series with the characteristic line impedance of 50 Ω this adds up to the desired value, and makes for a 1:10 division ratio. Note that for a 50 Ω signal source this reduces the swing visible at the scope by only a factor close to 5, not 10 (because the load is now negligible, while a 50 Ω load reduces the driver swing by half).

integrated MOSFET amplifiers exhibit input impedances up to several hundred kΩ or more, virtually eliminating any static load on the measured system. Leading-edge probes achieve bandwidths exceeding 10 GHz. The main detractor is the parasitic capacitance (again the amplifier has to be placed right at the probe tip, but some minimal stub is unavoidable), which we will look at more closely shortly, but a good high-performance probe will have only a fraction of a pF.

Not surprisingly this improvement in performance comes at a price. Fast transistor circuits are very sensitive to overvoltage or electrostatic discharges – simply touching them with our finger when we are not well grounded can destroy them, and a good active probe is expensive (much more than a good passive probe). They need an external power supply (but this is often provided built into the oscilloscope). Being active circuits they will always add some noise and some random jitter to the signal, but this is usually more than compensated by the increased signal amplitude delivered to the oscilloscope. And finally, there is a clear tradeoff between electrical and mechanical performance – to minimize parasitics and to allow to probe today's narrow-pitch device packages and connectors, the tips of high-bandwidth probes (and the ground leads as well) must be small and correspondingly fragile.

4.4 Probe Effects on the Signal

As we have already mentioned in the beginning of this section about probes, some influence of the probe on the system under test (and consequently on the signal to be measured) is unavoidable. In this section we will investigate this in a bit more detail, with the goal of understanding typical problems and looking for possible remedies.

4.4.1 Basic Probe Model

Figure 32 displays a simple, very generic model of a probe. While it does not claim to be a very realistic or detailed model of any real probe, it does include all the necessary components needed to get a good general understanding of how a typical probe behaves in the system. We will have a closer look on each of those probe components in the following sections.

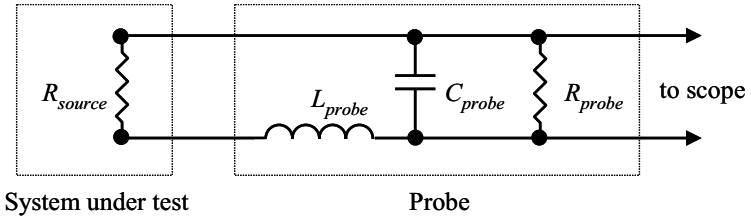


Figure 32: Basic, simplified model of an oscilloscope probe.

4.4.2 Probe Resistance

First, every probe has some static input impedance, represented by the ohmic resistance R_{probe}[41]. We have already seen that this resistance is rather low (usually less than a kΩ) for passive probes, and much higher (at least a few ten kΩ) for active probes. This load will reduce the signal amplitude in the system under test and may even potentially affect the circuit's operation. It will also create additional reflections when the probe point is in the middle of a signal line. Fortunately, at least for active probes, this load is too small to be of much importance in real life, unless the source impedance is very high (which is more of an issue with on-chip probing than with line drivers, since the latter have to be strong – i.e. low impedance – enough to drive a 50 Ω transmission line).

4.4.3 Parasitic Probe Capacitance

Of much more concern is the parasitic probe capacitance C_{probe}. It is the sum of the probe tip's fringe capacitance, the stub leading to the resistor or amplifier, and – for active probes – the input gate capacitance of the amplifier. As we all know the impedance of a capacitance C_{probe} is inversely proportional to the frequency f, namely

$$Z_C = \frac{1}{2 \times \pi \times f \times C_{probe}}. \tag{28}$$

For frequencies above a certain limit this shunt capacitance will have much lower impedance than the ohmic probe resistance, thus completely dominating the probe's total impedance. The effect is illustrated graphically in Figure 33: From that we see that the probe capacitance is a more

[41] The scope's input resistance – or more precisely, the characteristic impedance of the cable between the probe head and the scope – is of course in series with the impedance of a passive probe, but it is small compared to the probe resistance.

important factor for high-speed probe performance than its DC resistance: The smaller the capacitance, the higher the frequency until which the probe impedance stays sufficiently high (i.e. much higher than the signal source impedance) to allow meaningful measurements. Another – maybe surprising – result is that for frequencies in the higher GHz range (corresponding to rise times below a few hundred ps) a straight cable connection is actually superior to all "real" probes – the great-looking 1 MΩ DC resistance specification of one's expensive active probe is of no real use here!

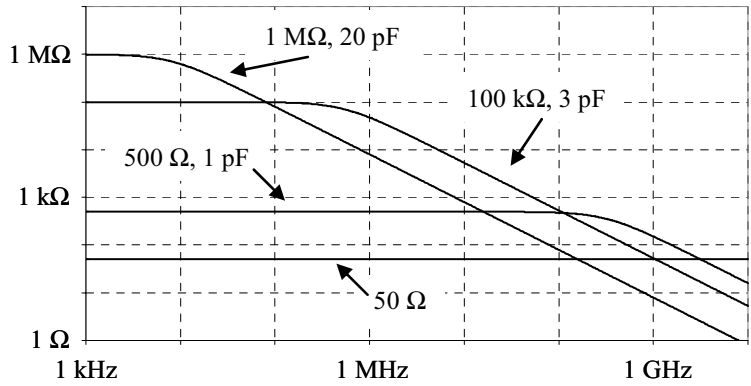


Figure 33: Effective input impedance vs. frequency for different probe types.

What's more, when probing in the middle of a signal line, the probe capacitance will cause a reflected spike going back to the driver, and it will degrade the rise time of the transmitted signal. As we know from our discussion about transmission lines and parasitics the rise time and bandwidth of such a low-pass R-C filter is $T_{10/90} \approx 2.2 \times C \times Z_0/2$ and $BW_{-3dB} \approx 0.35/T_{10/90}$, respectively, so apart from reflections and circuit loading the parasitic capacitance also limits the achievable probe bandwidth[42].

As an example, let's assume we have an active probe with a DC resistance of 100 kΩ and a parasitic capacitance of 1 pF, and we want to probe a signal with 100 ps rise time (the edge looks Gaussian). What can we expect to see? First, the rise time corresponds to a signal bandwidth of 0.33/0.1 = 3.3 GHz. At this frequency the impedance of the capacitance is around 48 Ω, much lower than the DC resistance and closing in onto the

[42] Note that the factor ½ in the rise time formula is only valid if the probe point is either in the middle of the transmission line, or at the end when the line it is terminated with matched termination (in either case the two sides act together to form a 25 Ω Thevenin equivalent source). If probing at the – unterminated! – end of a line, the source impedance is 50 Ω which doubles the effective filter rise time and halves the bandwidth.

25 Ω effective source impedance, so we should expect to see some impact on the signal. The filter rise time is $2.2 \times 1\,\text{pF} \times 50\,\Omega / 2 = 55\,\text{ps}$. The rise time seen by the probe amplifier (not counting the amplifier's own bandwidth limitations) will then be $\sqrt{100^2 + 55^2} = 114\,\text{ps}$. In addition attaching the probe to the transmission line will degrade the transmitted signal rise time to the same value, and also cause a reflected spike of approximately 20% of the amplitude, which is far from negligible[43]!

4.4.4 Parasitic Probe Inductance

The only component we haven't talked about yet is the probe inductance. It is caused by the current loop consisting of the signal path, the probe ground return path, and the return path through the system under test, as illustrated in Figure 34. As a general rule, the larger the enclosed area, the larger is the total inductance. This is one more reason why high-speed probes have to be so tiny.

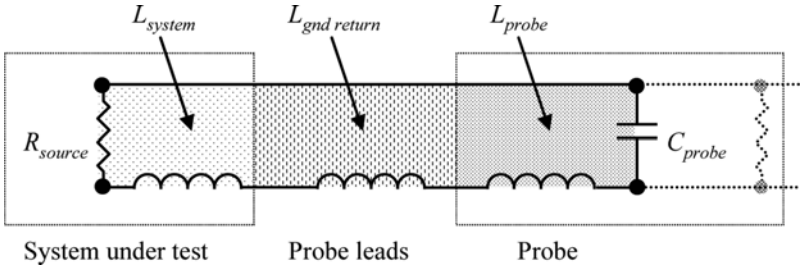


Figure 34: Current loop of system, probe, and ground return path, causing parasitic inductance.

In combination with the probe capacitance and the ohmic source resistance[44] this inductance forms a damped, resonant LCR circuit. Its resonance frequency – assuming small damping – is

$$f_{res} \approx \frac{1}{2 \times \pi \times \sqrt{L_{probe} \times C_{probe}}}. \qquad (29)$$

[43] For the calculation remember that in the reflection formula the rise time is the 0%/100% rise time, which we can approximate as $T_{0/100} \approx 1.25 \times T_{10/90}$.

[44] In most cases the effective source resistance is much lower than the probe's DC resistance, so the former completely dominates. This is what we assume here.

How strongly damped (or how well oscillating) this circuit is depends on the so-called Q-factor

$$Q = \frac{\sqrt{\frac{L_{probe}}{C_{probe}}}}{R_{system}}. \tag{30}$$

A Q-factor much smaller than 1 means strong damping, R is dominant, so there is no ringing and the probe input largely behaves as the simple RC low-pass filter we discussed before. A high Q-factor on the other hand indicates very little damping, so an incident edge will excite strong, long lasting resonant ringing in the probe (at the frequency f_{res} given above). For the case just in between (close to a Q-factor around 1, also called "critically damped"), where the damping is too strong for real resonance, the RCL filter rise time is given by

$$T_{10/90} \approx 3.4 \times \sqrt{L \times C}\,. \tag{31}$$

In any case we see that reducing the parasitic inductance is always in our interest, since there is little we can do about the resistance of the system under test[45].

We have several possibilities to minimize the loop inductance: First, make the signal path as well as the ground connection as short as possible. A long cable that conveniently reaches a far-away ground point is an absolute no-no if we want anything close to decent performance. If we do have to use a short cable for the ground return, press it as close to the probe enclosure as possible, to avoid any unnecessary loop area. We may not always have much choice in the ground path inside the system under test, but choosing the closest possible ground attachment is usually a good idea. More than one ground connection is even better since it means we have several inductances in parallel, which further reduces the total inductance. Finally, as should be abundantly clear by now, having no ground connection at all is a capital crime in the world of probing, even though one may have heard people argue that "It worked just fine when I tried": What happens is that the return current will find *some* way back, probably through a huge detour comprising the fixture power supply and the scope power supply among other nasty

[45] At first glance increasing the probe capacitance to reduce the Q-factor and avoid ringing may look tempting, but all it will really do is trade in this resonance for severe rise time degradation because of the increased low-pass RC filter time constant. Reducing the inductance is our only valid option.

elements. The huge ground loop created by this will prevent any meaningful measurement above maybe a MHz or so (which probably was the speed those people were running their system at).

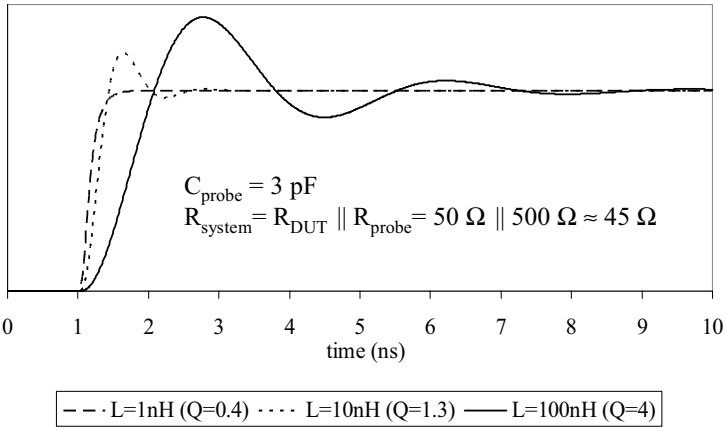


Figure 35: Acquired waveforms for different probe grounding schemes (resulting in different probe inductances), for a passive 500 Ω probe attached to a 50 Ω driver.

To get a *rough* estimate for the parasitic inductance of a given ground return loop we can use the following approximation formula:

$$L \approx 6 \cdot 10^{-7} \times D \times \left(\ln \frac{8 \times D}{d} - 2 \right), \tag{32}$$

where L (in H) is the inductance, D (in m) is the diameter of the loop (assumed to be a circle), and d (in m) is the diameter of the wire. Since the logarithm is a very slowly changing function, the inductance is mostly proportional to the loop diameter, the wire diameter being only of secondary importance. For example, if we have a loop of 1 cm diameter of a wire of 0.5 mm thickness, then the resulting inductance is

$$L \approx 6 \cdot 10^{-7} \times 10^{-2} \times \left(\ln \frac{8 \times 10^{-2}}{5 \cdot 10^{-4}} - 2 \right) \approx 18\,\text{nH}\,.$$

This is a huge inductance indeed for high speed purposes and will severely limit our measurement bandwidth. For example if we have a probe capacitance of only 1 pF (which is a pretty good and probably expensive probe), then the ringing frequency would be a measly

$$f_{res} \approx \frac{1}{2 \times \pi \times \sqrt{18 \cdot 10^{-8} \times 1 \cdot 10^{-12}}} \approx 380\,\text{MHz},$$

in other words, we would not be able to accurately measure anything even close to 400 MHz (or corresponding rise times of 1 ns or less).

To illustrate the preceding paragraph graphically, Figure 35 shows the acquired waveforms with different grounding schemes (long cable, short cable, minimum length ground pin).

4.4.5 Noise Pickup

In addition to degrading our probe rise time and causing ringing, the ground return loop has one more unpleasant effect: It picks up noise from electromagnetic fields, as shown in Figure 36. The process is simple magnetic induction – the signal loop acting as a coil that is traversed by a magnetic field changing in time. The induced voltage is proportional to the enclosed loop area, so here is another reason why reducing this area makes a lot of sense!

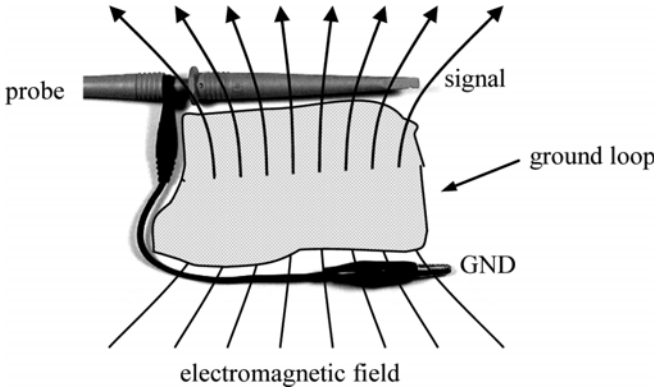


Figure 36: Noise pickup through the probe ground loop.

How can we find out if a strange-looking feature or some excessive noise on our waveform is caused by noise pickup? As a first step, disconnect the probe from the system under test and connect the probe tip directly to the ground lead. Without noise pickup we should not see any signal on the scope. Any signal we do see is external noise coupling into our loop. To further prove that this is the source, change the size of the loop (e.g. by squeezing the ground return closer to the probe tip) – the noise amplitude should change accordingly.

Since only magnetic fields *traversing* the loop can cause an induced signal, rotating the loop so the field lines are simply passing by is another way to reduce the size of the induced noise. This directional sensitivity can also help in tracking down the source of the interference.

4.4.6 Avoiding Pickup from Probe Shield Currents

Figure 37 shows another practical case of external noise coupling into the oscilloscope (or other measurement device). Chances are every test and characterization engineer has been bitten by this situation at least once. The troublemaker here is the huge loop that is closed by the separate ground connections of the system under test and the oscilloscope. If there is even just a weak field present (e.g. 60 Hz noise from other power lines or from fluorescent lights) it will induce a sizeable loop current I_{noise}. This in turn will produce voltage noise because even the best cable connection has some ohmic resistance. The important parameter here is the resistance R_{shield} of the signal return path (the outer shield for a coaxial cable); it will cause a voltage offset V_{noise} between the system ground and the instrument ground:

$$V_{noise} = I_{noise} \times R_{shield} \tag{33}$$

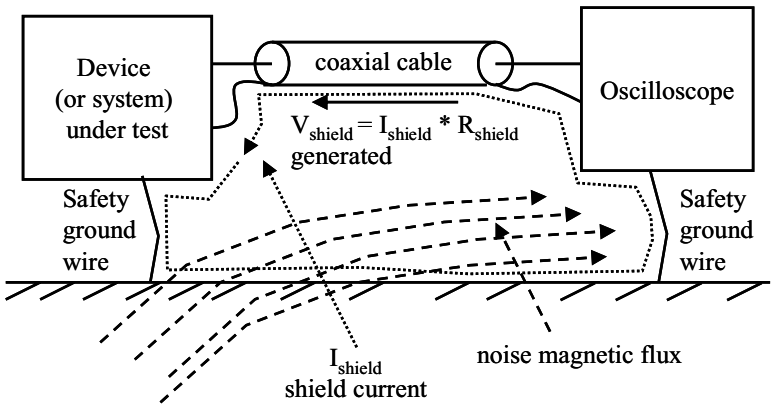


Figure 37: Noise caused by probe shield currents.

One can try several things to mitigate this noise:

- Reduce the current loop by attaching scope and system to same power outlet and keep the power cables close to each other. This is good practice anyway.

- Add a shunt capacitance between the scope and logic ground on the test fixture. This causes more of the noise current to flow through the shunt and less through the shield.
- Put a big inductance in series with the shield. This raises the inductance of the probe shield, lowering the current. An easy way to do this is to wind the probe cable a few times through a ferrite choke. This method is useful between maybe 100 kHz and 10 MHz – below that range the achievable inductance is too small, above the ferrite material ceases to be effective.
- If using a passive 10:1 probe, replace it with a 1:1 probe. A 10:1 probe attenuates actual logic signals and makes shield voltages appear relatively ten times larger. But 1:1 probe often has much larger parasitics, and the circuit may not like the increased load.
- At the very last, the best method is to use true differential probing. A differential probe has both of its connections floating (i.e. not connected to ground or the shield), thus it picks up the correct ground level at the fixture. This is one of the major advantages of differential probes compared to single-ended probes.

4.4.7 Rise Time and Bandwidth

As we have just seen, parasitics in the probe front-end section limit the maximum bandwidth of the probe and thus increase the rise time of the signal. If our probe is an active probe then the amplifier will compound this effect, and in any case the oscilloscope will degrade the signal further. As usual, those rise times add up approximately as root-mean-square, that is

$$T_{r,display} \approx \sqrt{T_{r,signal}^2 + T_{r,probe}^2 + T_{r,scope}^2}\ , \tag{34}$$

where $T_{r,display}$, $T_{r,signal}$, $T_{r,probe}$, $T_{r,scope}$ are the displayed rise time, the "real" signal rise time, the probe rise time, and the scope rise time, respectively.

The equivalent bandwidth of the scope-and-probe combination[46] is of course[47]

$$BW_{scope+probe} = \frac{1}{\sqrt{\frac{1}{BW_{probe}^2} + \frac{1}{BW_{scope}^2}}} \tag{35}$$

Note that for high-performance probes the specified probe bandwidth only holds true with optimum grounding (minimum length ground lead). It should be clear that between scope and probe the weaker link determines overall performance – a low-bandwidth probe (or even a good probe, but with long ground connection that causes ringing and bandwidth degradation) can obliterate the performance of even the best – and most expensive! – oscilloscope. Since probes are typically much less expensive than a scope of similar bandwidth it always makes sense to get a probe with somewhat higher bandwidth rating than the scope.

4.5 Differential Signals

So far we have always regarded our signals as being single-ended (i.e. a single signal line referenced to ground). However, with faster data rates differential signaling is becoming more and more prevalent and is most likely to replace single-ended signaling in most applications. On the other hand, many traditional oscilloscopes and probes are only single-ended. Even though we haven't gotten into jitter – we will do so in the next chapter – it probably seems very plausible even now that in order to properly characterize all the jitter characteristics of a differential signal, the acquisition must be done differentially as well.

4.5.1 Probing Differential Signals

The most straightforward method for this is of course to use a true differential probe, which virtually always is an active probe. There is no

[46] We should note that some scope manufacturers directly specify on their probes the total bandwidth of scope plus probe, assuming of course that one uses one of their probes with the intended scope of theirs. For high-bandwidth oscilloscopes this actually makes a lot of sense because in this performance range scope and probe are usually optimized for each other (e.g. for minimum ringing), so using a third-party probe can often result in sub-optimal performance.

[47] One gets this result by simply replacing the rise times with the bandwidth in the formula for adding bandwidth, using $T_r \approx k / BW$, and assuming that the k factors are the same for probe and for scope (which may not be exactly the case).

fundamental design difference between a single-ended active probe and a differential one apart from the fact that the input of the latter is freely floating and referenced to the second input, while the former has ground as its reference. Both probe types transmit the same type of signal to the oscilloscope, i.e. the scope does not care if a differential or a single-ended probe is attached to it. So in other words, any active differential probe, while giving the flexibility to probe differential signals, can also be employed to probe single-ended signals without restriction.

It also means the same limitations apply, namely bandwidth restrictions – even the most advanced commercial differential probes max out around 12 GHz and are thus ill suited for data rates beyond 6 Gb/s. And since they employ active elements (the amplifier), they inevitable add additional noise and jitter onto the signal. Still, as long as these limitations are beyond the required performance, these probes are the tool of choice for differential probing. They also offer a high common mode rejection ratio (CMRR) at lower frequencies (a few MHz), but the CMRR falls of rapidly with increasing common signal frequency.

The great advantage of direct measurement of the differential signal (using a true differential active probe) is that all timing and jitter measurements are then absolutely identical to single-ended measurements – after all, what the probe delivers is a single-ended signal proportional to the difference between the two input signals. Second, with a high impedance probe we can directly probe signals in a system, e.g. signals running on a bus, without adding excessive load (which otherwise may make the system fail since additional loading causes additional attenuation).

Unfortunately where we encounter differential signals is predominantly at highest speeds (where the advantages of differential signaling are becoming absolutely necessary to retain enough noise margin), while at the same time those probes – being active devices – are restricted in bandwidth. In other words, just where we would need them most, active differential probes hit their performance limit.

Most high-end bit error rate testers also offer differential capabilities (both for drive and for receive). Compared to oscilloscope probes their task is made somewhat easier by the fact that they do not intend to reconstruct the full waveform, but only need to trigger at a certain threshold.

Another option is to feed the two single-ended components (true and complement signal) of the differential signal into two separate channels on the oscilloscope (let's call them channels *A* and *B*), and use the built-in math functions to display the difference signal $D = A - B$. That way one is not limited by the maximum active probe performance (12 GHz at best), but gets the full bandwidth of the oscilloscope – especially in the case of equivalent-time sampling scopes a large benefit. Of course this will put a 50 Ω load on

the circuit, which is usually okay in a device characterization situation (where the only load is the scope, i.e. no in-circuit probing is necessary), but would otherwise be unacceptable. It also only works well if the signals don't carry excessive common mode components (depending whether they are DC balanced or not, AC-coupling could potentially remove this problem). Otherwise this is a perfectly viable option as long as we consider two things:

In order for this scheme to work the two channels (probes, cables, and scope input amplifiers and samplers) have to be extremely well matched, both in delay as well as in gain. Otherwise a sizeable portion of spurious common mode signal will show up (that is not present on the real differential signal), in other words, the common mode rejection ratio will be very poor. Out of this reason until a few years ago single-ended probing of differential signals was usually strongly discouraged, since especially analog scopes had no way of tightly controlling their amplifiers' gain and gain linearity. But today, where digital high-end instruments can employ sophisticated gain and linearity compensation and calibration, gain mismatch has largely become a non-issue. As for timing, given a good cable vendor it is possible to obtain cable delay matching down to 1 ps, which is enough even for the fastest speed we encounter today (delay should be matched to within a small fraction of a rise time, and even 12 Gb/s data signals have 20/80 rise times around 30 to 40 ps).

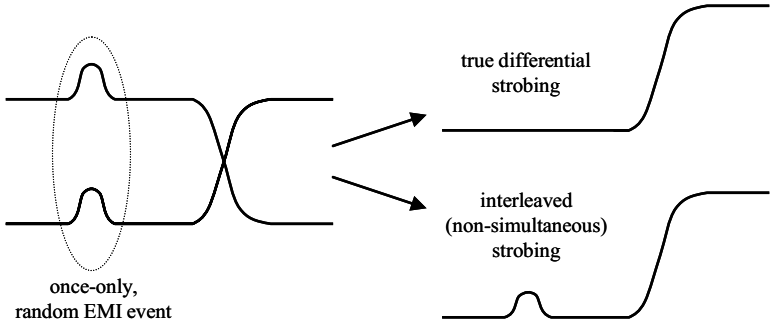


Figure 38: For differential signals, true differential (simultaneous) strobing of the signal pair can yield very different results to interleaved strobing.

Second, in order to look at random jitter and noise, the scope must be able to strobe both channels simultaneously. Otherwise the math waveform will not display the true differential signal because each point consists of two data points (from *A* and from *B*, respectively) taken at different times – so the instantaneous random noise would be completely unrelated between those two. As an example, look at Figure 38. The signal is experiencing a common mode spike that moves the levels of both single-ended signals simultaneously and in the same directions. If both *A* and *B* are sampled at

this same instant, the scope will correctly display the differential signal, unaffected by the spike (because it cancels in the difference). But is A is sampled at this instance, and B is sampled at the next pattern repeat (assuming this time no spike occurs), then the spike is only present at A and thus on the display – in contrast to the real situation on the signal.

As a matter of fact, many oscilloscopes employ interleaved (non-simultaneous) strobing, either to save cost, or because of hardware limitations. Unless we know for sure what our situation is, how can we check if scope does simultaneous strobing or not? The easiest possibility is to take a differential pattern generator and let it drive a pseudo-random bit stream (many generators have this capability built in; otherwise they may allow to program an arbitrary pattern). Feed one channel (true) into scope channel A, the other (complement) into scope channel B. Trigger the scope on a clock running at the data rate speed (i.e. don't use a trigger that only occurs once per pattern!). Set the scope up to display $D = A - B$. If sampling is simultaneous, D will display as a two-level eye diagram[48] as in Figure 39(a) because the signals at the same instant are always the opposite of each other: Either $A = low$ and $B = high$, thus $D = high$; or $A = high$ and $B = low$, thus $D = low$). If not, then *three* levels will appear because the level on A (taken at some time) and the one on B (taken at some other time) are completely unrelated to each other (third level: $A = low$ and $B = low$, or $A = high$ and $B = high$, both resulting in $D = midlevel$ (zero) between high and low), as shown in Figure 39(b).

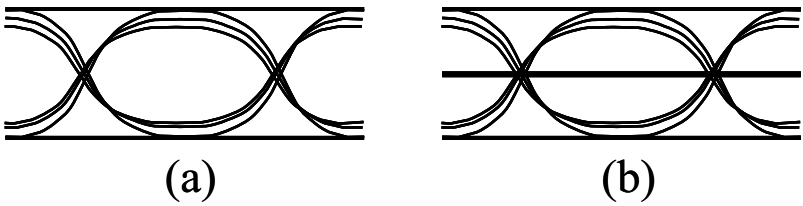


Figure 39: An eye diagram of the differential signal, acquired through two single-ended channels, reveals if the oscilloscope acquires the two components simultaneously (a) or interleaved (b).

A final word of caution: Some scopes on the market do simultaneous strobing when the deskew adjustment between their channels is set to zero, but when this adjustment is non-zero (even a single ps is enough!) they switch over to interleaved strobing. In this case if we need to remove skew between those two channels our only option is to add an adjustable hardware delay line (see section 5.2.7) in front of the scope inputs and leave the internal scope deskew setting at zero.

[48] We will talk about pseudo-random patterns and about eye diagrams in the next chapter.

Second, even when an oscilloscope acquires both channels simultaneously this does not necessarily mean the strobe jitter of those channels is correlated, because some part of the two strobe delivery paths may be separate. In other words, the observed random jitter of the displayed differential signal may or may not be the true differential jitter on the actual signal.

4.5.2 Single-Ended Measurements on Differential Signals

The remaining problem is how to relate the jitter measured on the difference trace on the scope to the actual jitter on the real differential signal. For deterministic, static jitter types like data dependent jitter or duty cycle distortion this is trivial – they are the same.[49] Unfortunately this is not necessarily true in the case of random jitter or any type of jitter (uncorrelated periodic jitter comes to mind) that requires single-shot acquisition:

Let's assume our single-ended signals are S_1 and S_2, and we let the scope (or tester) calculate the differential signal M as

$$M = S_2 - S_1 . \tag{36}$$

If S_1 and S_2 are sampled simultaneously then M corresponds to the real differential signal and we are done – just perform any measurement and analysis on M instead of the single-ended signals. But many scopes – and equivalent-time sampling scopes in particular – interleave the sampling between the two channels; e.g. the scope may first acquire a full trace of S_1, and then a full trace of S_2, or maybe acquire one sample of S_1 and then one sample of S_2, and so on. Nobody can guarantee that the instantaneous jitter at the first sample instant is the same as the jitter at the second instant – especially for random jitter. In fact, we can almost guarantee it is not!

This is less of a concern for data dependent jitter or duty cycle distortion because it will not change from one run (or sample) to the next – per definition it only depends on the pattern driven. Thus we can average the two curves over several acquisitions (to get rid of uncorrelated and random jitter), and the differences of the averages will be identical to the average of the differences. Done.

There is no such simple way out for random jitter and other uncorrelated jitter. Let's have a look on Figure 40 which shows the two single-ended components of the differential signal:

In Figure 40(a) S_1 and S_2 jitter together (their timing jitter is correlated). This is a common case when the signals are generated by a truly differential,

[49] We are getting a little ahead of ourselves here. For more details about the different types of jitter refer to the next chapter.

low-noise driver that does not add much jitter (through noise) on its own, so the timing jitter mostly comes from earlier stages of the signal generation and will affect both signals' timings the same way. In this case the timing jitter of the true differential signal is equal to the timing jitter of each of the single-ended signals, which is easily measured. This case also applies to all types of data dependent effects as long as the two signal paths are comparable (same bandwidth etc.).

Figure 40(b) displays the opposite case – here the jitter of S_1 is the opposite of the jitter on S_2 (i.e. when one signal is early, the other one is late). The most frequent cause of this situation is common noise on both signals (e.g. from crosstalk or EMI) that makes both signals move higher (or lower) in lockstep, as indicated in Figure 40(b). In this extreme the differential signal has no timing jitter whatsoever (the crossing point moves up and down in voltage, but stays still timing-wise), while each single-ended signal on its own certainly has. Measurement of the single-ended jitter here would be completely misleading since it does show significant jitter.

Finally, there is Figure 40(c) where both signals' jitter is completely independent from each other. This can come e.g. from random noise on the final driver stage of the signal source. Let's assume at some given transition signal S_1 has the ideal timing, while S_2 is late by some amount – as shown in Figure 40(c). The differential crossing point is moved in both voltage and timing, but the timing movement is only *half* of the displacement of S_2. Thus the differential jitter of M contributed by S_2 is only half the single-ended jitter on S_2. At the same time, S_1 also contributes some jitter. If we assume the jitter on S_1 and S_2 to be the same amount (very likely if the two paths are designed the same), and further assume the jitter to be purely random (Gaussian distribution), then their values will add up as RMS, i.e.

$$jitter_M = \sqrt{\left(\frac{jitter_{S1}}{2}\right)^2 + \left(\frac{jitter_{S2}}{2}\right)^2} = \frac{jitter_{S1}}{\sqrt{2}}. \qquad (37)$$

In other words, the jitter on the true differential signal is smaller than the measured single-ended jitter by $\sqrt{2}$.

So with the three cases above for given (measured) single-ended jitter, the true differential jitter on the differential signal can either be zero, or equal to the single-ended jitter, or smaller (by $\sqrt{2}$) than the single-ended jitter. In any practical situation there will always be some mixture of those three cases (although one can be dominant), but in all of the cases the true differential jitter is always *smaller* or at worst equal to the measured

single-ended jitter. This means that if we only measure single-ended, the result is a worst-case, pessimistic estimate of the true differential jitter.[50]

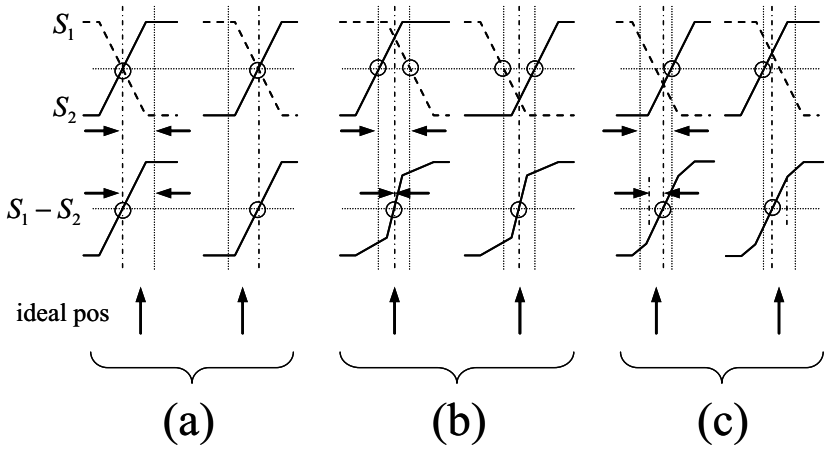


Figure 40: Possible relationship between the random jitter of a differential signal vs. the random jitter of its single-ended components: (a) correlated, i.e. differential jitter equal to single ended jitter, (b) anticorrelated, thus differential jitter zero, (c) uncorellated, thus differential jitter smaller than single ended jitter.

4.5.3 Passive Differential Probing

If we absolutely need to determine the uncorrelated differential jitter (random, periodic, uncorrelated deterministic) and only have single-ended measurement capability, one way is to use a *passive* Balun (see section 5.2.2) to convert the differential signal into a single-ended ones. Most Baluns will restrict our bandwidth and will distort the signal to some extent, but we are not out to measure data dependent effects with that setup anyway. Rather we can take advantage that – being a passive element – other than limited in bandwidth, random jitter is not influenced by such a device, and not worry too much about other signal distortions. Data dependent effects can later be measured single-ended without the Balun in the path.

[50] This favorable relationship between singled ended and differential jitter is also one of the reasons (but by far not the only one!) why differential signaling is so popular for higher data rates or noisy transmission situations because it provides a reduction in jitter compared to the single-ended case, of course at the cost of doubling the channel count.

5. ACCESSORIES

5.1 Cables and Connectors

There may be no other part of the signal chain that is so often overlooked as the humble coaxial cable. As long as speeds are moderate, it is merely seen as a passive interconnect that transports the electrical signal from the source (in our case the device under test) to the receiver (the oscilloscope or similar instrument). But as a matter of fact, the electrical quality of the cable used determines to a significant amount the high-frequency performance of the overall measurement system.

Coaxial cables – having a homogeneous geometry and therefore homogeneous capacitance and inductance per unit length – act as transmission lines, in our case most likely with a characteristic impedance of 50 Ω. As any real transmission line they are subject to losses – DC (ohmic) resistance, skin effect, and dielectric loss. Those losses give rise to an attenuation of the signal, which is usually specified in dB per length (e.g. dB per meter), i.e. on a logarithmic scale. One of the advantages of this scale is that the loss numbers for series of cables simply add up, i.e. when cable 1 has a loss of 2 dB and cable 2 has a loss of 3 dB at a given frequency, then the series combination of those two cables will have a loss of $2+3 = 5$ dB at this frequency. As we recall, a loss of 3 dB corresponds to a 30% reduction in amplitude, and 6 dB mean we lost already half of our signal. In the previous chapter we saw that skin effect losses increase with the square root of frequency and dielectric loss linearly proportional to the frequency, so at higher speeds those effects will become more and more important. High-quality cables have usually specified their total losses at several different frequencies.

5.1.1 Cable Rise Time/Bandwidth

For digital signals in particular there is more to losses than a simple reduction in amplitude. Keeping in mind that frequency is the inverse of amplitude, it is immediately plausible that losses affect the strongest the part of a waveform at or immediately after a transition because short time scales correspond to high frequencies. The outcome is that those frequency dependent losses degrade (increase) the rise time of edges, which in turn not only affects the accuracy of rise time measurements, but it worsens data dependent timing errors (see later) and closes the data eye because the signal does no longer reach its full level before the subsequent transition: Since at least in theory both skin effect and dielectric loss only get smaller, but never

really disappear with decreasing frequency (increasing time scale), they are effective even long time after some transition; in reality the exponential behavior of dielectric effects makes them negligible rather soon, while the square root behavior of the skin effect causes it to stick around for much longer.

It should be obvious that a longer cable of the same type as a shorter one means more losses, but what is the exact dependency?[51] Let's say we have a cable of some length that – due to skin effect and dielectric loss – has some rise time T_r. What happens if we put two of those cables in series, e.g. to be able to connect our system under test to a bulky oscilloscope that we can't (or don't want to) approach too close to the system? We may remember the RMS rule (adding up the squares) to add up rise times from the previous chapter. It is tempting to apply this rule blindly to the new situation, but this neglects the fact that the rule is only valid for Gaussian (and closely valid for simple one-pole exponential) edges – unfortunately neither skin effect nor dielectric loss produce Gaussian edge shapes.

For dielectric loss it turns out that the rise time instead increases *linearly proportional* to the length of the cable. In other words, if we double the length of the cable, the equivalent rise time of the (longer) cable will be twice as large as before. This is a far stronger increase than the $\sqrt{2}$ increase that we would have expected from the RMS rule!

But it gets even worse for skin effect – the rise time in this case increases *proportional to the square* of the cable length. This odd behavior becomes more plausible if we remember that in frequency domain the skin effect increases with the square root of the frequency, thus in time domain it disappears proportionally to $1/\sqrt{time}$. Skin effect is based on the ohmic resistance of the cable material (aggravated by the inhomogeneous current distribution), and ohmic resistance doubles when the cable length doubles. Thus in order for the skin effect of a cable of a certain length to reduce to the same value as a cable of only half that length, we have to wait *4 times* as long, or in other words, the skin effect loss increases with the square of the length.

The strong dependency on the cable length for both skin effect and dielectric loss illustrates well the paramount need for (and the great benefits associated with) keeping cables as short as possible. For high-end oscilloscopes there are often extenders available that allow us to bring the (small) sampling head as close as possible to the system under test – the

[51] As stated before, in the frequency domain (i.e. corresponding to RF type signals) things are easy: If the loss for some cable for some frequency is x dB, then a cable twice as long will have 2x dB loss. But digital signals, which contain a very wide band of frequencies (each of them experiencing a different amount of loss), the situation becomes more complicated to describe, especially in the time domain.

extender cables carry only the strobe signals whose exact shape is only of secondary importance, but it enables us to minimize the analog signal path from the system through the cable to the sampling head.

Given the rise time of the cable used (or maybe even the relative contributions of skin effect and dielectric loss), the output rise time for a given input signal rise time can still, at least approximately, be calculated using the RMS rule, as long as we calculate the cable rise time based on the behavior above (linear and quadratic dependency, respectively):

$$T_{r,out} \approx \sqrt{T_{r,in}^2 + T_{r,cable}^2} \approx \sqrt{T_{r,in}^2 + T_{r,skin}^2 + T_{r,dielectric}^2} \,. \tag{38}$$

The approximate bandwidth is then given as usual by

$$BW_{cable} \approx \frac{0.33}{T_{r,cable}} \,. \tag{39}$$

5.1.2 Skin Effect Compensation

By now it should be obvious that the main enemy of digital signal transmission through a cable is less the absolute attenuation than rather the dependency on the frequency, because this is what distorts the signal edges, and what causes data dependent jitter.[52] A possible workaround is to make a tradeoff between static losses (ohmic, DC) and variations due to skin effect. Figure 41 shows a cross-section of such a skin-effect compensated cable. Its center conductor (which has a much smaller surface than the shield and thus contributes most of the skin effect losses) consists of a rather resistive core, plated with a thin layer of smooth and highly conductive material (e.g. silver). The high core resistance means that even at low frequencies (where the current takes the path of least ohmic resistance) most of the current will already flow in the thin outer layer, so when skin effect takes hold at higher frequencies, there is only little change in the current distribution, and therefore the effective resistance does not change much. Of course this is only true as long as the skin depth is larger than the thickness of the outer layer – once it becomes smaller, the same old $\sqrt{frequency}$ behavior takes hold again. Overall, this provides good compensation of skin effect (i.e. a very flat loss profile vs. frequency) up to a certain maximum frequency.

[52] Constant (frequency independent) losses merely reduce the signal levels at each instant by a constant factor, but they do not cause any waveform distortion. Such losses can easily be taken into account by adjusting the driver levels and/or the receiver threshold levels.

One can improve this scheme further by using some ferromagnetic material for the core. That's one reason why cables with silver-plated steel cores are so popular (apart from their mechanical strength). Looking at the skin effect formula for a cylindrical cable in section 1.6.5.2 we see that the skin depth is inversely proportional to the square root of the magnetic permeability μ, which for steel can be thousands of times higher than for non-ferromagnetic metals. As a result the current gets pushed out of the steel core already at very low frequencies, flowing almost entirely in the copper (or silver) plating, and the frequency response is flat from low frequencies on up to where the skin depth becomes smaller than the thickness of the plating.

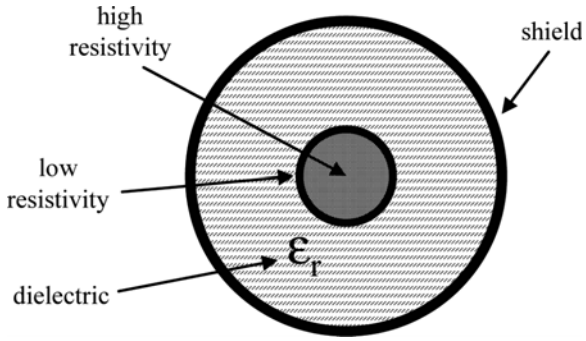


Figure 41: Cross-section of a coaxial cable with skin-effect compensation.

5.1.3 Dielectric Loss Minimization

For dielectric loss the most straightforward approach is to use a lower-loss dielectric. Vacuum (or air) would be ideal, but except for absolutely rigid transmission line structures it would pose insurmountable technical challenges. So the next best thing, which is employed in high performance cables, is to use foamed Teflon. Teflon has very small losses up to high frequencies to begin with, and they are further reduced by the amount of air contained in it because the signal now sees a mixture of Teflon (low losses) and air (close to zero losses). Of course the size of the air bubbles in the Teflon must be orders of magnitude below the wavelength of the highest frequency component to be transmitted so the dielectric really acts as a homogeneous body.

The only downside of this foamed dielectric (as opposed to a solid one) is reduced mechanical ruggedness. High-performance cables don't like to be flexed too often and will wear out rather fast. This wear-out can manifest itself in changes in delay, loss, and signal shape when the cable is flexed.

For the coaxial connectors, dielectric loss minimization means replacing the solid dielectric with minimum-size standoffs that are just sufficient to keep the center conductor in place. The price is again reduced mechanical stability.

5.1.4 Cable Delay

As long as we do not have to deal with propagation delay matching between the two lines of a differential pair, the absolute delay of our cables is normally of secondary importance. What we do care however are changes in the delay during the measurement.

All dielectrics exhibit more or less pronounced changes of their dielectric constants with temperature – and thus changes of the propagation delay since the propagation speed varies with $\sqrt{\varepsilon}$. So keeping the ambient temperature constant during a long measurement becomes very important for highest accuracy requirements.

Second, all cables show some change in delay when they are bent and flexed since the dielectric deforms and thus the cable geometry changes. For quality cables the maximum amount of change (for flexure down to the minimum allowed bend radius) is usually given in the cable specifications. For good cables the change is in the order of 1 ps, but cheaper, lower-performance cables can produce many times that amount. In any case for highest accuracy it is advisable to avoid any movement of the cables during measurement; another option is the use of a rigid or semirigid cable assembly (of course at the cost of setup flexibility and ease of use).

5.1.5 Connectors

Just as important as it is to get the signal *through* the cable (or some other component of the path) it is to get the signal *into* and *out of* the cable. This is where connectors come in. The best, highest bandwidth cable will not help us if we cannot connect it to the rest of the setup with equally high bandwidth.

For high-speed, high-accuracy laboratory environments – where performance is paramount and the number of connections is usually small – there exists a relatively small number of standard connector types, each geared towards a specific frequency range and level of ease of use.[53] Table 1 shows an overview for the common types (ordered from lowest to highest

[53] The situation would be different e.g. for consumer applications or large-scale production setups where cost per pin/channel has to be low and often a large number of connections has to be made.

performance): BNC, Type-N, SMA, 3.5 mm, 2.92 mm, 2.4 mm, 1.85 mm, and 1.0 mm.

Some common trends are visible when going to higher and higher performance connectors (and the general trends hold true for connectors other than the ones discussed as well):

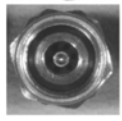
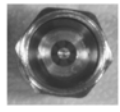
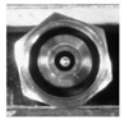
First, all the connectors are coaxial, impedance-controlled designs. Good impedance control is easiest achieved for straight barrels, so any bends (i.e. connectors that have a 90 degree launch into the cable) risk to introduce discontinuities unless the manufacturer went to great lengths in the geometric design to keep the impedance constant throughout the path within the connector. Thus, when in doubt it is always better to use a connector with a straight launch into the cable.

Second, like any transmission line, connectors exhibit losses that limit high-frequency performance. To reduce the dielectric loss contribution, good connectors must either use a high-performance dielectric like Teflon, or try to minimize the amount of dielectric altogether. Thus in the highest bandwidth types one will only find small standoffs that keep the center pin in place, rather than a full solid barrel of dielectric. This is also a giveaway when looking at an unknown connector – e.g. SMA and 2.92 mm are mechanically compatible (same thread, can mate to each other) but the (higher-bandwidth) 2.92 mm connector will contain largely air.

Third, in order to avoid any dispersion (and thus signal edge distortion) in the connector, the diameter of the connectors must be much smaller than the wavelength of the highest frequency component to be transmitted (it is the inner diameter of the outer (shield) conductor that counts). Otherwise there will be "strange" modes of propagation – cavity resonances – that introduce loss and distortion. So it is not to surprising that BNC connectors have a much larger diameter than SMAs, and those have larger diameters than e.g. 2.4 mm connectors (in fact the mm number gives this diameter; it does not tell anything about the outer thread diameter, so e.g. 3.5 mm and 2.92 mm connector have compatible threads.

Finally, any slots or holes in a connector cause resonances and radiation losses – thus only the low-bandwidth BNC type has such slots in the outer conductor, and the highest-bandwidth connectors avoid even slots in the center barrel of the female type that mates with the male plug. This of course requires extremely high geometrical precision during manufacturing to assure reliable contact without the compliance provided by a slotted and thus elastic receptacle.

Table 1: Commonly used coaxial connector types.

Name	Bandwidth GHz	Remarks	Picture
BNC	2...4	Cheap, easy to connect (latch-on). Slots leak radiation. 50 Ω and 75 Ω versions.	
Type-N	up to 18	Often used in RF equipment (e.g. sine generators); 50 Ω and 75 Ω versions.	
SMA	12...20	Widely used; poorly standardized tolerances.	
Super-SMA	26	Improved, high-precision SMA version.	
3.5 mm	34	Mechanically compatible to SMA.	
2.92 mm (K^{54})	40	Mechanically compatible to SMA.	
2.4 mm	50		
1.85 mm (V^{54})	65	Mechanically compatible to 2.4 mm	
1 mm	110		

[54] "K" and "V" are Anritsu's copyrighted designations for their 2.92 mm and 1.85 mm connectors, respectively.

A final word about good care for connectors (and any other accessories of our setup): High-bandwidth connectors are very sensitive to abuse and their performance will degrade quickly if not handled properly – and those changes may go unnoticed for a long time while degrading our signal integrity and our measurement results. Please keep the rubber dust caps they came with on whenever the connector is not in use. Use a torque wrench of the proper torque to tighten them – this assures a solid connection (and good signal integrity) while at the same time preventing damage from over-torqueing. When tightening the screw, always turn the outer barrel and *not* the connector (or cable) itself, because that would grind down the internal mating surfaces. To clean a dirty connector use some alcohol on a cotton swab, but don't touch the center pin of male 3.5 mm, 2.4 mm or faster connectors – there is not much dielectric to hold them in place and we would risk breaking or dislocating them.

5.2 Signal Conditioning

5.2.1 Splitting and Combining Signals

Sometimes we need to send the same signal into more than one input. A typical example is an equivalent-time sampling scope (that needs an external trigger) when we want to trigger on the signal itself. Or we may want to monitor a signal on a bus by picking off a small portion of it.

The worst thing we can do in this case is to use some of the (readily available) T-connector pieces (shown in Figure 42(a)). As a 1:2 splitter they are horrible because they make no effort at all to match input and output impedances: The incoming signal hits a combination of two transmission lines in parallel, so strong reflections will occur. Two better solutions are the power splitter (Figure 42(b)) and the power divider (Figure 42(c))[55]. The resistors in series combination with the transmission line impedances provide the necessary impedance matching. While the power splitter only looks like 50 Ω for the input port (left side in Figure 42(b)), the power divider is matched for any of the three ports.

[55] We should mention, however, that the impedance matching comes at a price, namely ohmic power dissipation in the resistors. The output amplitude on each output is only half the input, so overall half the power is lost. The T-connector on the other hand does not dissipate any energy at all (it only reflects and transmits it), and the transmitted power is thus higher than for the other two solutions (transmission coefficient of 66% vs. 50%). If the signal sources provide matched source termination, thus swallowing the reflections, it can sometimes be the solution of choice to maximize the transmitted signal amplitude.

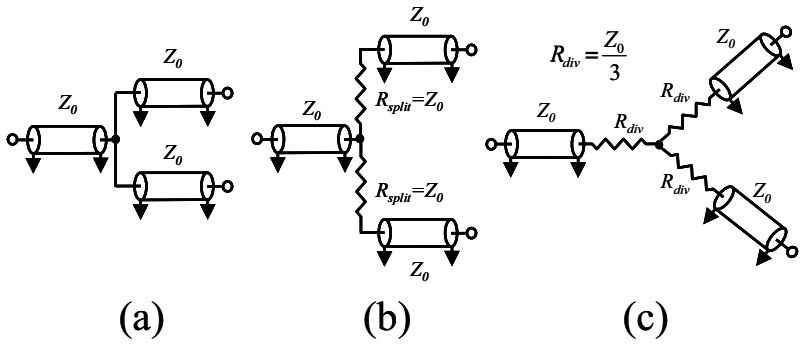


Figure 42: Different designs for broadband signal splitters: (a) simple T-element, (b) power splitter, (c) power divider.

Of course, when driven in the opposite direction (two signals driven in, one coming out), a divider will act as a power combiner, i.e. it will perform a summing operation on its two input signals. This can be useful e.g. to look at the common mode voltage of a differential signal if the oscilloscope can't add up two channels, or for adding some controlled amount of noise to a signal (see also section 9.2.13).

Beware of any splitters (or dividers) that are not explicitly sold for time domain applications! Splitters for RF type applications are normally only designed to work over a narrow frequency band (the main goal here is usually minimum losses, so they do not employ lossy resistors but rather matched-length transmission line couplers and other "esoteric" stuff) and they won't perform at all when presented with a – wide-band! – digital data signal.

5.2.2 Conversion between Differential and Single-Ended

When dealing with differential signals there is often a need to either convert a single-ended signal into a differential one (e.g. we want to generate a differential signal but almost all RF generators are single-ended) or vice versa (e.g. we only have single-ended inputs on our oscilloscope but need to measure a differential signal). A Balun (from *bal*anced-*un*balanced[56]) is a useful accessory that achieves this conversion. The choice here is between active and passive elements.

The block diagram of a passive Balun is shown in Figure 43(a). While slower-speed passive Baluns (up to a GHz or so) actually use real wide-bandwidth transformers, faster designs make use of transmission line effects –

[56] Depending on the application (especially RF vs. digital), "differential" is also called "balanced", and "single-ended" is called "unbalanced".

an example[57] which works from a few kHz all way up into the multi-GHz region is shown in Figure 43(b). Because of their design, passive Baluns always have some lower frequency cutoff in addition to their high-frequency limit. Their main attraction comes from the fact that – being passive devices – they don't add any random jitter (but they may exhibit nonlinearities and data dependent effects).

An active Balun can be a simple differential high-speed op-amp circuit (including feedback of course). The advantage is that it can potentially operate all the way down to DC, but on the other hand any active element inevitably adds some random noise and nonlinear distortion to the signal (although those effects may be small enough to neglect them if using top-of-the-line components and good circuit design). In addition, while passive Baluns work in both directions (differential-to-single-ended conversion as well as single-ended-to-differential conversion), active Baluns typically can only perform either in one or in the other. This is not necessarily a disadvantage since it provides isolation against reflected signals.

Again, beware of components geared towards RF type applications (which have been around for a long time and are widely available) – most likely they will not work over the wide band of frequencies that a digital application needs. Always check the operating range and the gain (loss) flatness.

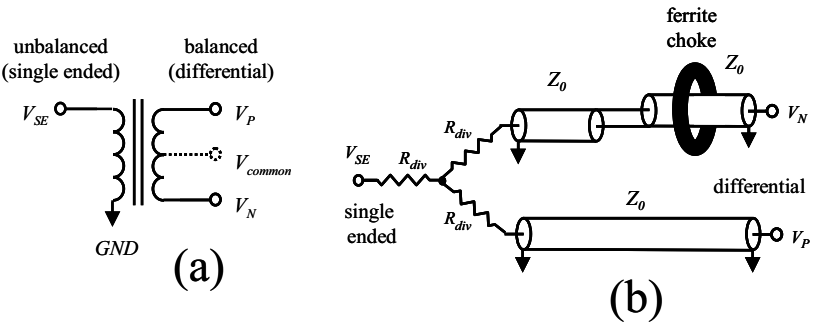


Figure 43: Different passive Balun designs: (a) transformer, (b) crossed transmission lines (semirigid coax) with RF choke (the second transmission line assures delay matching between true and complement path).

[57] Why the setup works may not be immediately obvious, so it deserves some explanation: The RF choke assures that the current coming out of the ground conductor on the balanced side *must* come back through the center conductor (and not through some other ground path) since every other path would loop around the choke and thus have huge inductance – which effectively prevents current through those paths down to rather low frequencies (a few kHz). For higher frequencies, where the ferrite loses its efficiency, the small parasitic inductance of the other ground paths takes over that role, since the shield in any case provides the lowest possible total path inductance..

5.2.3 Rise Time Filters

In digital (time-domain) applications, the only filter type used regularly is the low-pass filter[58]. Such a filter increases the rise time of the transmitted signal (in the frequency domain this corresponds to an attenuation of all frequencies above a certain threshold). Typical use cases are adjustment of the rise time of a signal source so the signal matches more closely the intended application, or simulation of the limited bandwidth of a slower receiver when measuring with a higher-bandwidth oscilloscope.

Some caveats apply here as well. First, depending on the actual design, such filters may be reflective or absorptive. The first type simply reflects the unwanted signal components – which can spell trouble of the driver cannot "digest" (terminate) strong reflections or if there are other impedance mismatches or parasitics in the signal path (because they will re-reflect those components to the receiver). The latter type absorbs them so they cannot interfere with our measurement. This is of course the preferred type, because in order to provide pure absorption (no reflections at the filter input) those filters have to provide matched 50 Ω impedance.

The second trap is the filter type. When one looks for filters (e.g. browsing the internet), chances are most of the search results will be filters for RF (frequency domain) applications, and again they are usually not well suited for the requirements of time domain applications. To avoid signal distortion other than pure rise time increase (e.g. overshoot, ringing), the group delay has to be largely constant, so the filter types of choice for time domain applications are ones with Bessel and Gaussian type filter responses.

Given the typical high bandwidths needed for present-day digital applications, all filters used here are passive filters (combinations of resistors, capacitors, and inductors).

5.2.4 AC Coupling

In quite a few interfaces, and also as an option in most oscilloscopes, one finds so-called *AC coupling*. What this means is that a capacitor is placed in series with the data line(s), as shown in Figure 44(a). This forms a high-pass filter, which effectively avoids any DC current flow, only AC components above a certain frequency (the filter bandwidth) can get through[59]. There can be several reasons for this. In an oscilloscope this allows to look at a small signal component that rides on a large DC offset (otherwise the large offset

[58] The only common exception to this rule is AC coupling (see section 5.2.4), which constitutes a high-pass filter.

[59] Which is the reason AC coupling is also called "DC blocking" .

would require a rather insensitive range setting which would not resolve the small signal well). Or it may be that the device driver must not be terminated to ground like most benchtop instruments (oscilloscopes, spectrum analyzers) do – CML (current-mode logic) pull-down drivers being a prime example which work well with AC coupling because they provide their own biasing. Or it can be a way to prevent damage to the sensitive I/O circuitry for hot-pluggable daughter cards.

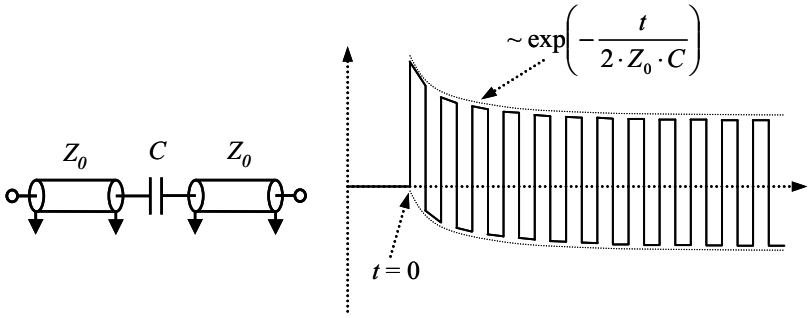


Figure 44: (a) AC coupling of a transmission path. (b) Initial settling behavior ("warm-up") of an AC coupled clock signal.

In order for AC coupling to work, the data stream (and the coupling capacitor) must fulfill certain criteria: First of all the data stream must be DC balanced (i.e. it must spend equal times in the high and the low state) within every not-too-long section of the stream. The figure of merit here is the *maximum running disparity* (RD_{max}) which tells us how many unbalanced "excess bits" (either ones or zeros) can accumulate at most within the data stream. The filter time constant $T_C = 2 \times Z_0 \times C$ must then be large compared to RD_{max} times the bit period to avoid excessive DC wander (change in average level on the output). The maximum wander A_{wander} is given by

$$A_{wander} < A_{signal} \times \frac{RD_{\max} \times T_{bit}}{Z_0 \times C}, \tag{40}$$

where A_{signal} is the signal amplitude, T_{bit} is the bit period, C is the capacitance, and Z_0 the line impedance. In addition, this scheme needs sufficient warm-up cycles to reach a steady state (in the end the output will always be centered around zero if the duty cycle is 50%), as illustrated in Figure 44(b) which shows a clock signal going through an AC coupled path. The settling time constant is identical to the filter time constant T_C. This settling behavior is a frequent cause of "My test run on the production tester fails, but signal looks fine when I check it with a scope" – the test strobes

come after an insufficient number of warm-up cycles, but when looking at the scope, the pattern has been looping for a long time. To minimize this warm-up, we should make *C* large enough to avoid droop, but not any larger.

Examples for data streams that will work with AC coupling:

- A clock with 50% duty cycle ($RD_{max} = 1$).
- Manchester-coded data in NR[60] format where each data bit is coded as two bits (either 10 or 01; $RD_{max} = 1$).
- 8b/10b encoded data ($RD_{max} = 3$) which is frequently used in modern serial transmission schemes like SerDes.
- PRBS[61] data of sufficient length (a 2^N-1 PRBS pattern has $RD_{max} = N$, but is not exactly DC balanced because the pattern contains one more "one" than it contains zeros, but for long patterns, around N > 7, this becomes negligible).

Cases where AC coupling will not work are data streams in RZ or RO format[62], patterns where there is no guaranteed DC balance (e.g. data from an A/D converter) or where RD_{max} can be excessively long (e.g. memory address lines).

As an example for how to size the coupling capacitor, assume we have a SerDes data stream at 1 Gb/s, 8b/10b encoding ($RD_{max} = 3$), $Z_0 = 50\ \Omega$, signal swing 400 mV, maximum allowed eye closing 1% = 4 mV:

$$C \geq \frac{A_{signal}}{A_{wander}} \times \frac{RD_{\max} \times T_{bit}}{Z_0} = \frac{400\,\text{mV}}{4\,\text{mV}} \times \frac{3 \times 1\,\text{ns}}{50\,\Omega} = 6\,\text{nF}\,.$$

To be on the safe side (among other issues capacitors can have large variations), let's choose C = 10 nF. We will then need to run warm-up cycles for at least several time constants, let's say 5 (this gives less than 1% residual level error), i.e.

$$T_{warmup} = 5 \times 2 \times Z_0 \times C = 10 \times 50\,\Omega \times 10\,\text{ns} = 5000\,\text{ns}\,,$$

which corresponds to 5000 cycles (bits)!

[60] NR = non-return, the signal stays at the same level until a bit of opposite polarity is sent.
[61] Pseudo-random bit stream - we will talk more about PRBS in section 9.2.4.
[62] Return-to-zero (one), the signal always returns to low (high) level before the next bit period.

5.2.5 Providing Termination Bias

The inputs of virtually any bench-top test equipment we encounter are either high impedance (slower oscilloscopes for example, or active FET probes) or provide matched 50 Ω termination. In the latter case more likely than not the termination goes straight to ground. This is fine in many cases, but certain device driver architectures (ECL and PECL for example) require that the end of the line be terminated to some other voltage, e.g. the positive or negative supply voltage. They may not even work at all if the termination pulls them towards ground! And they may not be so user-friendly and provide their own bias if AC coupled (like CML drivers would do). So what can we do?[63]

Using a high-impedance passive or active probe may work in some cases (this again depends on the particular architecture of the driver), but this means we leave the receiver end of the line completely unterminated – strong reflections are guaranteed. As long as the driver itself provides matched source termination this is of less concern, but first we cannot always be sure about the quality of its termination (keep in mind that what we want to *test* is the driver, so it isn't good practice to make too many assumptions about its good behavior), second this may not correspond to the actual use in the final application (in case the receiver there does provide termination), so the validity of our test results becomes questionable. Thus something more elaborate is required.

First, we could think of still using the high-impedance probe, but adding some termination circuitry of our own. This is indicated in Figure 45(a). As a side effect, the termination cuts the effective (Thevenin) source impedance in half, so the effect of any probe capacitance is reduced by half as well. (Of course the voltage swing is only half, too). Unfortunately a simple resistor connected to a power supply will not do the trick: The current through the resistor has to change as fast as the signal rises, which can be as fast as a few picoseconds. Clearly no power supply is up to that task, since already the cabling would cause many orders of magnitude more delay than that. So we need to provide some decoupling as well, indicated as a capacitor in Figure 45(a).

Things get even more complicated when the scope input is not high impedance but rather 50 Ω. In this case we cannot simply add some resistor since the termination would no longer be matched. Instead we have to create a "Thevenin equivalent" network that looks like 50 Ω at its input. At the very least this needs two resistors (plus the power decoupling mentioned

[63] We should note that most digital semiconductor testers have adjustable termination voltage on their digital pins, but not necessarily on their analog instruments. So in this chapter we will focus on benchtop equipment, especially oscilloscopes.

before of course); Figure 45(b) shows one possibility. The signal swing at the oscilloscope is attenuated, the attenuation being a tradeoff against the voltage necessary to drive the DC offset, and with the immunity against impedance tolerances of the oscilloscope.

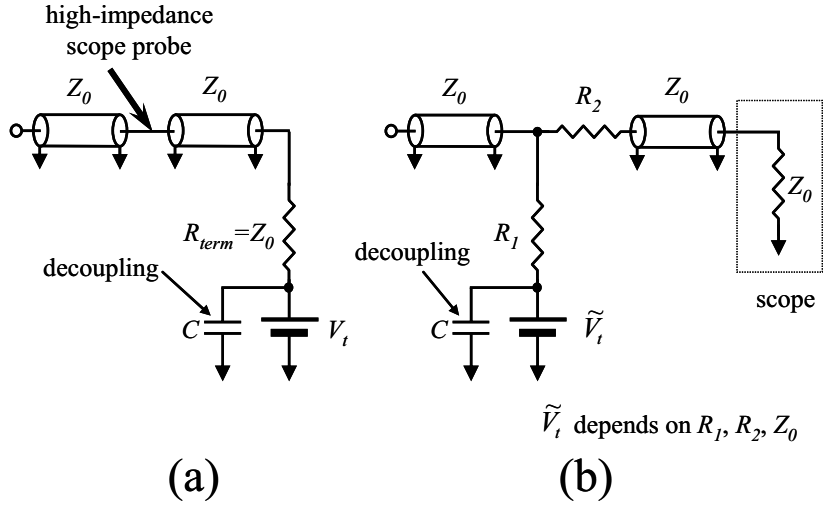


Figure 45: Providing matched termination to a bias level: (a) For a high-impedance probe. (b) Thevenin-equivalent termination for a matched impedance probe that by itself terminated to ground.

Second, we can go out and look for a probe with adjustable termination voltage. As a matter of fact, the selection is not very large, but there are models with SMA inputs where we can either program the termination voltage, or supply it with our own power supply (the probe takes care of the decoupling). This is an excellent and flexible solution with a single downside: Being an active probe, even though one of the best available, its bandwidth is limited to around 12 GHz or less, so we can't use it for faster data rates than maybe 6 Gb/s.

A second very common solution is the use of so-called *bias tees*. These devices are available off-the-shelf from several vendors. Such a bias tee, shown in Figure 46, is a combination of a DC block (a capacitor in series with the transmission path, i.e., AC coupling) and an inductor hanging off the path. The capacitor acts as a high pass filter and blocks any DC current into the oscilloscope. A DC power supply attached to the inductor provides the bias voltage, while the inductor prevents any AC signal to leak out into the power supply. While often used, this solution has several limits and shortcomings:

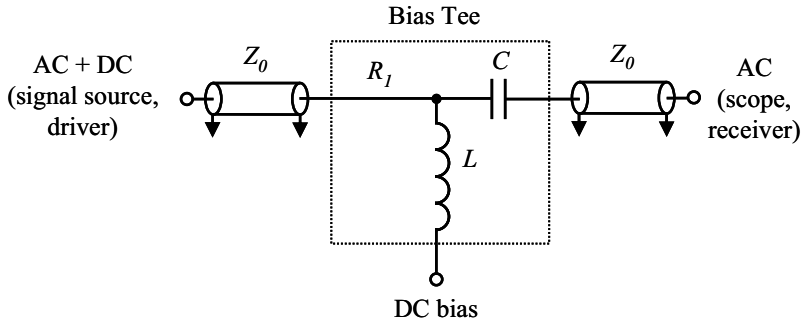


Figure 46: A bias tee is a good way to provide a DC bias to the driver, as long as the signal fulfills certain criteria.

First, the capacitor in the line constitutes AC coupling, i.e. a high-pass filter, so only frequency components *above* a certain limit are transmitted to the scope. Good bias tees have a limit as low as a few kHz. One cannot make the capacitor too large either (to lower the limit) because then its parasitic imperfections will degrade its high-speed performance, and settling will be too long (see section 5.2.4). The inductance shall *block* high-frequency components, so we want to make it as large as possible. In addition the inductance in combination with the capacitance form a resonant circuit with an approximate resonance frequency of

$$f_{res} = \frac{1}{2\pi \times \sqrt{LC}}, \tag{41}$$

which again gives motivation to make L and C as large as possible to lower the operating frequency limit. If we remember that long sequences of ones (or zeroes) correspond to low frequencies, it becomes clear that we cannot transmit really arbitrary patterns; the capacitor will charge up and block, and the inductor will open up and conduct. What's more, real coils are rather bad approximations to an ideal inductor. They have a lot of parasitic capacitance between their turns (which short-circuits the high-frequency components the inductor is supposed to block) as well as non-negligible ohmic resistance that will cause a voltage drop whenever current flows into the device under test – we will need to compensate for that. Practical designs aim to minimize the capacitive parasitics through the use of conical coils (coil diameter increasing from the signal path to the other end) where the first, small, low inductance section blocks the highest frequency components with very little parasitics, and subsequent, increasingly wider sections (thus also with increasing parasitics) block lower frequency components.

Second, since the DC component of the signal is blocked, the waveform on the scope will always be DC balanced around zero. This spells trouble when the waveform's duty cycle (the time it is high vs. the time it is low) is not constant – the whole waveform will move up and down, with a time constant equal to

$$T_c = 2 \times Z_0 \times C \,. \tag{42}$$

Since such a bias tee contains a DC block, the implications are the same as for AC coupling (section 5.2.4) – namely, the data stream must be DC balanced or at least have constant average duty cycle throughout the pattern (and the duty cycle must not change, lest the output signal levels change as well).

Third, some vendors offer termination blocks with either fixed (PECL, ECL) or adjustable termination voltage. Internally they implement a high-performance version of the power-decoupled Thevenin equivalent network mentioned before. For the driver those terminations look like 50 Ohm to the termination voltage. The limitations are their bandwidth (around 10 GHz is the highest available), and the fact that the output signal is usually attenuated (between 12 and 20 dB, i.e. by a factor between 4 and 10), which means reduced signal-to-noise ratio on the scope.

If our signal is differential (as are more and more high-speed transmission schemes anyway), there is a fourth, even better solution, as shown in Figure 47: First, we assume our lines are uncoupled lines, so even and odd impedance are the same (let's assume 50 Ω cables). We can then implement differential termination in its T-variant (see Figure 19(b)), where the third resistor disappears (because even and odd mode impedance are equal). The midpoint (center tap) voltage never moves, so there is no absolute need for any decoupling (we should still provide one to make the setup more immune against common mode noise). The beauty of this solution is that it retains DC coupling (no warm-up cycles or constant duty cycle necessary) and all parts are available off-the shelf in high-performance versions (20 GHz bandwidth or more), so this is the highest-performance solution available. For the two termination resistors we can (ab-)use an off-the-shelf high-bandwidth power *splitter*. Then all we need in addition are two power *dividers* and a DC power supply. Since the scope inputs draw some current (or, seen differently, the path from the power supply to the scope ground constitutes a voltage divider hooked up to the device driver), we need to set the power supply to some value higher or lower than the required termination voltage.[64] We do need to assure that the true and

[64] The exact voltage to force depends on the input impedance and the termination voltage at the system driver, i.e. on the actual current drawn by the signal source.

complement signal do not have any skew to each other since this would create sudden common mode spikes that we will have a hard time terminating completely[65] even when the center tap capacitor is present, so we need to match all cable delays etc. to a small fraction of the signal rise time.

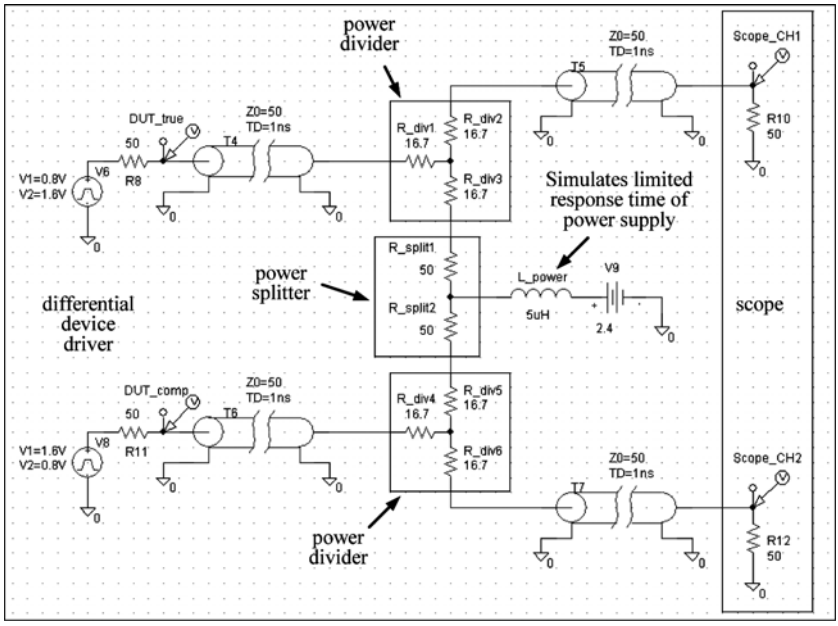


Figure 47: Providing DC biasing for a differential source while retaining maximum system bandwidth, using nothing but off-the-shelf accessories.

5.2.6 Attenuators

Sometimes it is desirable to reduce the signal amplitude to be measured. A frequent case is when using an equivalent-time sampling scope (which in order to maximize bandwidth does not have any amplifiers/attenuators built in before the sampler). If the instrument input is 50 Ω, then a simple resistor in front of the input would do the trick, for the price of an impedance mismatch into the instrument. E.g. to reduce the amplitude into the scope by half, we would place a 50 Ω resistor in front. This constitutes an ohmic 1:2 voltage divider, but now the total input impedance is the sum of the scope impedance and the resistor, i.e.

[65] Even though this termination scheme does provides decent *partial* termination/attenuation of common mode.

$$R_{tot} = R_{scope} + R_{atten} = 50\,\Omega + 50\,\Omega = 100\,\Omega\,,$$

and this would cause strong reflections back into the system under test. This approach would not work at all if the input were high impedance (it would only aggravate the low-pass effect of any parasitic probe capacitance present).

The solution is a three-resistor network as shown in Figure 48. The resistors can always be chosen in such a way that the effective input impedance into each side is 50 Ω *as long as the load on the other side has 50 Ω impedance*. As an example, a -6 dB attenuator (reduces the signal by half) would have $R_1 = R_2 = 16.67\ \Omega$ and $R_3 = 66.67\ \Omega$. This is nothing but the symmetrical power divider we looked at before, but with one transmission line impedance merged directly into R_3. Of course such a network built from discrete resistors would not achieve very high bandwidth because of the significant parasitic capacitances and inductances of real resistors. High-bandwidth commercial attenuators instead employ distributed thin-film resistors and the best achieve bandwidths in excess of 40 GHz.

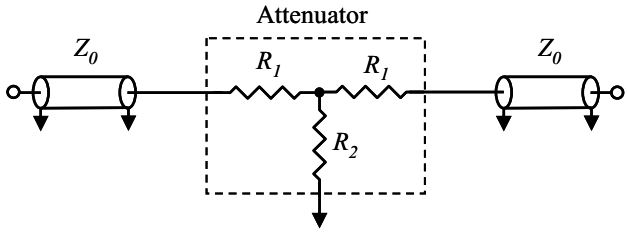


Figure 48: Resistive broadband attenuator.

The design formulas for this attenuator for an attenuation factor of *N* are

$$R_1 = Z_0 \times \frac{N-1}{N+1}, \quad R_1 = Z_0 \times \frac{2N}{N^2-1}, \tag{43}$$

and the attenuation in dB is of course

$$attenuation(\text{dB}) = -20 \times \log_{10} N\,. \tag{44}$$

Apart from simple amplitude reduction, attenuators have another important application: They can be used to reduce the impact of mismatched loads: If an attenuator (attenuation factor -x dB) is inserted into the signal path, the transmitted signal is reduced by -x dB, but reflections from e.g. the receiver going back to the source and there getting re-reflected traverse the

attenuator two additional times and thus experience -2x dB additional loss – as a result, the signal-to-noise ratio improves by 2x dB. The higher the attenuation value, the better the termination (for the price of reduced transmitted amplitude). In theory on could of course always employ a matching network that would provide *exact* 50 Ω termination[66], but such a network would have to be tailored to each specific load and thus is unlikely to be available off the shelf, while high-bandwidth attenuators are easy to come by.

5.2.7 Delay Lines

Sometimes we need to delay signals by a certain amount of time, but as far as possible without affecting their waveform or their jitter. One application is to look at the signal before a trigger with an analog or an equivalent-time sampling scope. Another frequent situation is the need for a stable, linear, well-defined delay as a timing reference (e.g. we may need to step a strobe signal relative to our measured signal, or to measure and compensate for the time base nonlinearity of an oscilloscope). Finally, a finely variable delay will allow to tightly match the timing of the two signals of a differential pair, which may have picked up some skew e.g. because of slightly mismatched cable lengths.

Only passive delay lines can assure that they don't add random jitter (any active circuit will). In principle a passive delay line is just a fancy name for a piece of low-loss transmission line (low-loss to minimize frequency-dependent attenuation and waveform distortion that would introduce additional data dependent jitter). So if all we need is a fixed delay, we can get that by inserting a piece of cable of appropriate length (use semi-rigid or rigid coax to avoid delay changes that inevitably occur when a cable is flexed and deformed). Achievable delays are in the range of a few 10 ps up to a few 10 ns (corresponding to physical lengths between a few mm and several m); above that length losses are likely to exceed our tolerances.

If we need a series of discrete delay values, we could cut a set of cables with those values and put it the appropriate one at the given time, but this is rather cumbersome (not counting the wear and tear on the connectors) and difficult to automate (if we have to switch the delay size under computer control). A better approach is to use high-bandwidth microwave relays to switch those cable sections in and out of the path. If we size their delays in steps powers of 2 (e.g. 1 ns, 2 ns, 4 ns, 8 ns), we can cover a wide range of delays, equally spaced, with a very limited number of sections. The minimum step size is limited by practical restrictions on the minimum feature lengths and manufacturing tolerances for the delay of the cables,

[66] Funny enough, such impedance *matching* networks are called "*mismatch* pads"!

relays, and solder connections. Such cascaded delay lines are commercially available.

Continuously variable and highly linear delay lines can be built as so-called trombones. They basically consist of two parts of transmission line that slide into each other while keeping good electrical contact. Such trombones can be designed for manual operation (both parts are threaded and so move in or out when turned against each other, or be driven by a stepper motor under computer control. If well designed (good impedance control) such delay lines can be extremely linear because they rely only on accurate geometrical parameters for their delay; building them as airlines minimizes losses so they achieve bandwidth reaching many GHz. And the resolution can be almost arbitrarily small (well below 1 ps). That said, they are still mechanical parts that always have some tolerance and some wiggle, so for maximum repeatability and accuracy one must always approach a given delay setting from the same direction (e.g. always start with a smaller delay and than increase it to the desired value).

http://www.springer.com/978-0-387-31418-1

Digital Timing Measurements
From Scopes and Probes to Timing and Jitter
Maichen, W.
2006, XIV, 240 p., Hardcover
ISBN: 978-0-387-31418-1