CHAPTER 14

Electronic Building Blocks

> And then of course I've got this terrible pain in all the diodes down my left-hand side.. . . I mean, I've asked for them to be replaced but nobody ever listens.
>
> —Marvin the Paranoid Android, in *The Hitchhiker's Guide to the Galaxy* by Douglas Adams

14.1 INTRODUCTION

The subject of electronic components and fundamental circuits is a whole discipline unto itself and is very ably treated in lore-rich electronics books such as Horowitz and Hill's *The Art of Electronics*, Gray and Meyer's *Analysis and Design of Analog Integrated Circuits*, and (for a seat-of-the-pants approach to RF work) the *ARRL Handbook*. If you're doing your own circuit designs, you should own all three.

There's more to it than competence in basic circuit design. Electro-optical instruments present a different set of problems from other signal processing applications, and that means that there's lots of stuff—basic stuff—that hasn't been thought of yet or hasn't been taken seriously enough yet. That's an opportunity for you to do ground-breaking electronic design work, even if you're not primarily a circuit designer; not everybody with a circuit named after him can walk on water, after all, even in the more highly populated fields.[†]

If we can gain enough intuition and confidence about the basic characteristics of circuitry, we can try stepping out of the boat ourselves. Accordingly, we'll concentrate on some important but less well known aspects that we often encounter in electro-optical instrument design, especially in high dynamic range RF circuitry. You'll need to know about Ohm's law, inductive and capacitive reactance, series and parallel *LC* resonances, phase shifts, feedback, elementary linear and digital circuit design with ICs, and what FETs, bipolar transistors and differential pairs do.

The result is a bit of a grab-bag, but knowing this stuff in detail helps a great deal in producing elegant designs.

[†] Jim Williams's books have a lot of lore written by people like that and are worth reading carefully.

Building Electro-Optical Systems, Making It All Work, Second Edition, By Philip C. D. Hobbs

14.2 RESISTORS

Resistors are conceptually the simplest of components. A lot of nontrivial physics is involved, but the result is that a good resistor has very simple behavior. A metal resistor obeys Ohm's law to extremely high accuracy (people have mentioned numbers like 0.1 ppm/V, a second-order intercept point of +153 dBm for 50 Ω—don't worry about it). In addition, its noise is very well described, even at frequencies below 1 Hz, by the Johnson noise of an ideal resistor. Not bad for a one-cent part.

There are only a couple of things to remember about metal film resistors. The first is that the cheap ones have biggish temperature coefficients, as high as 200 ppm/°C. This amounts to 1% drift over 50 °C, which may be a problem, especially if you're using trims. You can get 50 ppm/°C for only a bit more money, and you can do much better (1–5 ppm/°C) with expensive resistors. Clever use of cheap resistor arrays can achieve voltage divider ratios with this sort of stability as well. The other thing is that they are allowed to display some amount of $1/f$ noise when biased with high voltages: the specs allow rms noise from this source of 10^{-7} times the bias voltage applied, when measured in one decade of bandwidth. This is significantly more obnoxious than the nonlinearity; a 10 kΩ resistor with 15 V across it could have 10–100 Hz noise voltage of around 1.5 μV rms, corresponding to a noise current of 150 pA in the same bandwidth, 22 dB above the Johnson noise. In practice, the noise is nowhere near this large. The author measured some 11 kΩ RN60C type resistors in a 2:1 voltage divider with a 9 V battery and found that the coefficient was lower than 5×10^{-9} per decade all the way down to 1 Hz, which was the resolution limit of the measurement.

Other resistor types, such as metal oxide, carbon film, conductive plastic, cermet, and carbon composition, are worse in one or more ways. For example, metal oxide resistors (the cheapest 1% units available) have huge thermoelectric coefficients, so that a temperature gradient will produce mysterious offset voltages and drifts. Like everything else, resistors have some parasitic capacitance; an ordinary $\frac{1}{8}$ W axial-lead resistor has about 0.25 pF in parallel with it.

14.2.1 Resistor Arrays

You can get several resistors deposited on a common substrate in an IC package. These are wired either completely separately or all having one common pin. Arrays typically come in a 2% tolerance, but the matching between devices is often much better than that, and the relative drift is very small indeed, because the temperature coefficients are very similar and the temperature is fairly uniform across the array. The matching becomes even better if you use a common centroid layout, as in Figure 14.1. From an 8 resistor array, we can make two matched values by wiring #1, 4, 5, and 8 in series to make one, and 2, 3, 6, and 7 to make the other. This causes linear gradients in both temperature and resistor value to cancel out. (Note: The absolute resistance isn't stabilized this way, only the ratios.) You can make four matched values using #1 + 8, 2 + 7, 3 + 6, and 4 + 5, or even a reasonable 3:1 ratio using 3 + 6 versus the rest, though these fancier approaches may not be quite as good, because there's liable to be a second-order (center vs. edge) contribution beside the gradients.

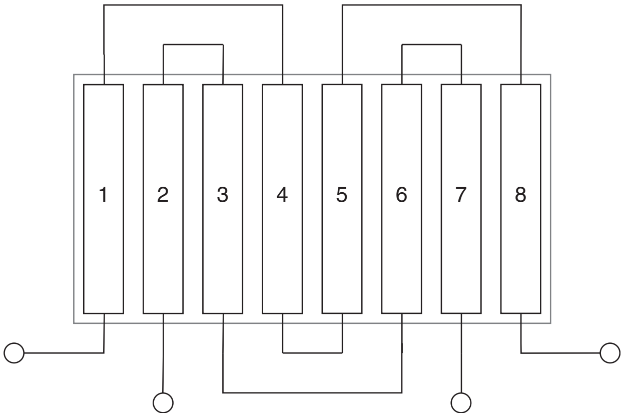

Figure 14.1. Common centroid layout for resistors in series: the two series strings will match closely because the layout ensures that linear gradients in surface resistivity cancel.

14.2.2 Potentiometers

A potentiometer is a variable resistor with three terminals: the two fixed ends of the resistive element and a wiper arm that slides up and down its surface. Because the resistance element is a single piece, the voltage division ratio of a pot is normally very stable, even though the total resistance may drift significantly.

The wiper is much less reliable than the ends. Try not to draw significant current from it, and whatever you do, don't build a circuit that can be damaged (or damage other things) if the wiper of the pot bounces open momentarily, because it's going to (see Example 15.2).

Panel mount pots are getting rarer nowadays, since microcontroller-based instruments use DACs instead, but they're still very useful in prototypes. Conductive plastic pots adjust smoothly—if you use them as volume controls they don't produce scratchy sounds on the speaker when you turn the knob—and have long life (>50k turns), but are very drifty—TCR ≈ 500–1000 ppm/°C—and have significant $1/f$ noise. Cermet ones are noisier when you twist them, and last half as many turns, but are much more stable, in the 100–200 ppm/°C range. Wire-wound pots also exist; their temperature coefficient is down around 50 ppm/°C, but since the wiper moves from one turn to the next as you adjust it, the adjustment goes in steps of around 0.1–0.2% of full scale, like a mechanical DAC. Their lower drift makes wire-wound pots much better overall.

14.2.3 Trim Pots

Trim pots, which are adjusted with a screwdriver, are much more common. They come in single-turn styles, where the shaft moves the wiper directly, and multiturn, which move the wiper with a screw. Stay away from the multiturn styles when possible. Their settability is no better in reality, and their stability is much poorer. If you need accurate fine adjustment, put low and high value pots in series or parallel, or use the loaded pot trick. Trimmers have much larger minimum resistances, as large as 5% of the end-to-end value, and won't stand much twiddling: typically their specs allow the resistance to change 15% in only 10 complete cycles.

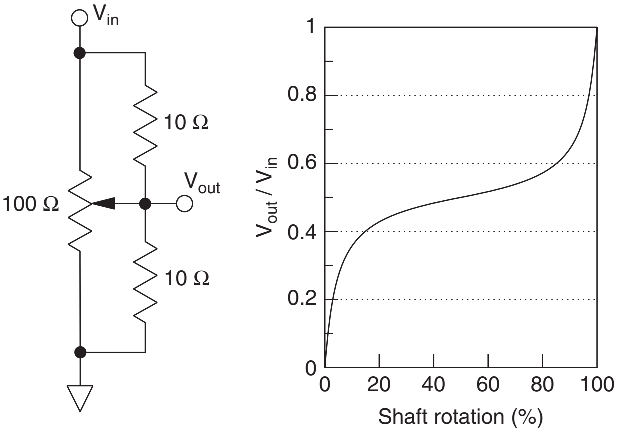

Figure 14.2. Loaded pot.

14.2.4 Loaded Pots

You can do some interesting things by putting small resistors between the ends of the pot and the wiper, the so-called *loaded pot*, as shown in Figure 14.2. A 1 kΩ pot with a 10 Ω resistor across each side has the same adjustment sensitivity as a 51 turn, 20 Ω pot, but is dramatically more stable. Big twists of the shaft cause small changes in balance, until you get right near one edge, when the pot starts to take over. Loaded pots are great for offset adjustments, another example of the virtue of a vernier control. A popular parlor game is figuring out new nonlinear curves you can get from a pot and a few resistors; it's a fun challenge, like wringing new functions out of a 555 timer, and is occasionally useful.

14.3 CAPACITORS

Capacitors are also close to being perfect components. Although not as stable as resistors, good capacitors are nearly noiseless, owing to their lack of dissipative mechanisms. This makes it possible to use capacitors to subtract DC to very high accuracy, leaving behind the residual AC components. The impedance of an ideal capacitor is a pure reactance,

$$Z_C = \frac{1}{j\omega C}, \tag{14.1}$$

where the capacitance C is constant.

Aside: Impedance, Admittance, and All That. So far, we've used the highly intuitive idea of Ohm's law,[†] $V = IR$, for describing how voltages and currents interact in circuits. When we come to inductors, capacitors, and transmission lines, we need a better

[†] Purists often call the voltage E (for electromotive force) rather than V, but in an optics book we're better off consistently using E for electric field and V for circuit voltage.

developed notion. A given two-terminal network[†] will have a complicated relationship between I and V, which will follow a constant-coefficient ordinary differential equation in time. We saw in Chapter 13 that a linear, time-invariant network excited with an exponential input waveform (e.g., $V(t) = \exp(j\omega t)$) has an output waveform that is the same exponential, multiplied by some complex factor, which was equivalent to a gain and phase shift. We can thus describe the frequency-domain behavior of linear two-terminal networks in terms of their impedance Z, a generalized resistance,

$$V = IZ. \tag{14.2}$$

The real and imaginary parts of Z are $Z = R + jX$, where X is the reactance. A series combination of R_1 and R_2 has a resistance $R_1 + R_2$, and series impedances add the same way. Parallel resistances add in conductance, $G = G_1 + G_2$, where $G = 1/R$, and parallel impedances add in admittance, $Y = 1/Z$. The real and imaginary parts of the admittance are the conductance G and the susceptance B: $Y = G + jB$. (See Figure 14.3.)

Whether to use admittances or impedances depends on whether you're describing series or parallel combinations. The two descriptions are equivalent at any given frequency. Just don't expect X to be $-1/B$ or R to be $1/G$ for complex impedances, or that series RC circuits will have the same behavior as parallel ones over frequency.

The linear relationship between AC voltage and current in a reactive network is sometimes described as "Ohm's law for AC," but the physics is in fact quite different; Ohm's law arises from a large carrier density and many dephasing collisions in materials, whereas the linearity of Maxwell's equations means that reactance is still linear even in situations where resistance is not (e.g., in an electron beam or a superconductor).

14.3.1 Ceramic and Plastic Film Capacitors

The two most popular types of capacitor are ceramic and plastic film. Ceramic dielectrics come with cryptic names such as NPO, X7R, and Z5U. Table 14.1 should help sort out the confusion. Plastic film and low-k[‡] monolithic ceramic types are very linear and will

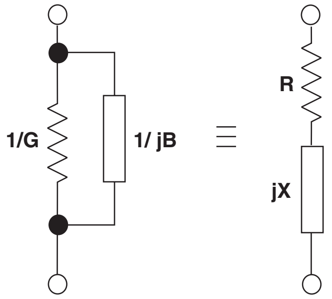

Figure 14.3. Series and parallel equivalent circuits for an impedance $R + jX$. Note that these only work at a single frequency due to the frequency dependence of X.

[†] *Two-terminal* is the same as *one-port*, that is, one set of two wires.

[‡] For historical reasons, we refer to the relative dielectric constant ϵ as the relative permittivity k when discussing capacitors. They're the same thing, don't worry.

TABLE 14.1. Common Dielectrics Used in Capacitors

Dielectric	Tol	TC	Range	(ppm/° C)
Ceramics				
NPO/C0G	5%	0 ± 30	1 pF–10 nF	Good for RF and pulse circuits; best TC, stable, low dielectric absorption.
X7R	10%	-1600	1 nF–1 μF	Like a cheap film capacitor; TC very nonlinear.
Z5U	20%	$+10^4$	10 nF–2.2 μF	Good for bypassing and noncritical coupling, but not much else.
Silver mica	5%	0 ∓ 100	1 pF–2.2 nF	Excellent at RF, bad low frequency dielectric absorption.
Plastic film				
Polyester (mylar)	10%	700	1 nF–10 μF	General purpose. Very nonlinear TC—it's cubic-looking.
Polycarbonate	10%	$+100$		Similar to polyester
Polypropylene	1–5%	-125	47 pF–10 nF	Best for pulses; very low leakage, loss, dispersion, and soakage
Teflon	5%			Unsurpassed low leakage and dielectric absorption, good for pulses
Electrolytics				
Aluminum	$+80/-20$		1 μF–1.0 F	big C, but slow, noisy, lossy, short-lived and vulnerable to frost.
Solid tantalum	20%		0.1 μF–220 μF	Better than wet Al; low ESR and ESL; big ones are very expensive.

cause you no problems apart from temperature drift and possibly a bit of series inductance. Polypropylene especially is an inexpensive material with really good dielectric properties and decent high temperature performance.

In order to pack a lot of capacitance into a small volume, you need many very thin layers of high-k material. Unfortunately, to get high k values, you have to trade off everything else you care about: stability, low loss, linearity, and so on. An X7R ceramic is allowed to change its value by 1%/V. This will cause serious distortion when X_C is a significant contributor to the circuit impedance. A Z5U ceramic's capacitance can go up by half between 25 °C and 80 °C. High-k ceramic capacitors can even become piezoelectric if mistreated, which is really fun to debug (look for weird reactance changes below about 1 MHz).

Capacitor tempco can also be a source of drift. An RC with bias voltage V_{bias} will produce a thermal offset of

$$\Delta V = V_{\text{bias}} R \frac{dC}{dT}\frac{dT}{dt}, \tag{14.3}$$

which can be serious, for example, if $dT/dt = 10$ mK/s, a 1 second TC using an X7R capacitor with 2.5 V across it can produce a 40 μV drifty offset.

The wisdom here is to let high-k devices do what they're good at, so in high level circuits, use values large enough that you don't have to care how much they vary.

Since the big shift to surface mount, plastic film capacitors are not as common as they were. The reason is that the capacitor body gets very hot in surface mount reflow soldering. The good film caps come in through-hole only. Newer plastics such as polyphenylene sulfide (PPS) caps come in SMT, but they're not the most reliable things.

Metallized plastic film capacitors are especially useful in situations where transient overvoltages may occur, since if the dielectric arcs over, the arc erodes the metal film in the vicinity of the failure until the arc stops, which amounts to self-repair. On the other hand, film-and-foil capacitors have more metal in them, so their ohmic (I^2R) losses are much lower.

14.3.2 Parasitic Inductance and Resistance

A capacitor doesn't really discharge instantaneously if you short it out; there is always some internal inductance and resistance to slow it down. Inductive effects are mostly due to currents flowing along wires and across sheets, and are lumped together as *effective series inductance* (ESL). Resistive effects come from dielectric absorption and from ohmic losses in the metal and electrolyte (if any). Each charge storage mechanism in a given capacitor will have its own time constant. Modeling this would require a series RC for each, all of them wired in parallel; since usually one is dominant we use a single RLC model with C in series with a single *effective series resistance* (ESR) and ESL.

If you need really low ESR and ESL, e.g. for pulsed diode laser use, consider paralleling a whole lot of good quality film-and-foil capacitors.

14.3.3 Dielectric Absorption

Those other contributions we just mentioned are actually the main problem with capacitors in time-domain applications such as pulse amplifiers, charge pumps, and track/holds. These need near-ideal capacitor performance, and those parallel contributions (collectively known as dielectric absorption, or *soakage*) really screw them up. The imaginary (dissipative) part of the dielectric constant contributes to the ESR and thus puts an exponential tail on part of the discharge current. The tail can be very long; an aluminum electrolytic may still be discharging after 1 s, which would make a pretty poor track/hold. Choose Teflon, polypropylene, or NPO/C0G ceramic dielectrics for that sort of stuff, and whatever you do, don't use high-k ceramics, micas, or electrolytics there. Polypropylene and silicon oxynitride (AVX) capacitors are available in 1% tolerances, so they're also very useful in passive filters and switched-capacitor circuits.

14.3.4 Electrolytic Capacitors

Electrolytic capacitors have a lot of capacitance in a small volume, which makes them useful as energy storage devices (e.g., power supply filters). Capacitors of a given physical size have a roughly constant CV product; thinning the dielectric increases the capacitance but decreases the working voltage, so that CV remains constant. The energy in the capacitor is $\frac{1}{2}CV^2$, so for energy storage you win by going to high voltage.

They work by forming a very thin oxide layer electrochemically on the surface of a metal; because the dielectric forms itself, it can be extremely thin. Electrolytics can be wet (aluminum) or dry (solid tantalum). Dry electrolytics are better but more expensive. Electrolytics exhibit leakage, popcorn noise, huge TCs, and severe dielectric absorption, so keep them out of your signal paths. The ESR of an aluminum electrolytic climbs spectacularly when the electrolyte freezes.

Ordinary electrolytics are polarized, so that they must not be exposed to reverse voltages. Nonpolarized electrolytics exist but cost a lot and aren't that great. If you reverse an electrolytic capacitor in a low impedance circuit at appreciable voltage, it may explode. Old-style metal can electrolytics were known as *confetti generators* because of this behavior. Solid tantalum capacitors can be very touchy if mistreated, especially by reversal or excessively large current transients (e.g., on the input side of a voltage regulator) that cause hot spots. They go off like solid-fuel rockets, with the sintered metal being the fuel and the manganese dioxide electrolyte the oxidizer—they burn hot and generate shrapnel. Solid aluminum and polymer electrolytics don't share this problem. You needn't avoid tantalums altogether—they have very low impedances at high frequency and don't misbehave at low temperatures the way aluminum electros can. Just treat them gently.

Electrolytics become electrically very noisy when they are reversed, which can be a good diagnostic test. They also can generate tens to hundreds of millivolts of offset on their own—far worse than the soakage in other types. (Of course, being electrochemical cells, they have more than a passing resemblance to batteries.) Don't use electrolytics in the signal or bias paths of any sort of low level, low frequency measurement.

Aside: Eyeball Capacitor Selection. Two capacitors that look identical may be very different in their properties, but there are a couple of rules of thumb: first, caps that are small for their value are probably high-*k* ceramics, and second, tight accuracy specs (J or K suffix) go with better dielectrics such as C0G, polypropylene, or Teflon. All polarized electrolytics have a polarity marking, of course.

14.3.5 Variable Capacitors

You almost never see panel-mount variable capacitors any more, except in very high power applications (e.g., transmitters or RF plasma systems). This is something of a shame, since many of them were works of art, with silver-plated vanes beautifully sculptured to give linear frequency change with shaft angle. Tuning is done with varactor diodes, whose capacitance changes with bias voltage. Adjusting different stages for tracking is now done with DACs or trim pots, setting the slope and offset of the varactor bias voltage.

What we do still use is trimmer capacitors, based on single vanes made of metallization on ceramic plates, whose overlap depends on shaft angle. These have most of the same virtues and vices of NPO and N750 fixed capacitors. The figure of merit for tuning applications is the capacitance ratio $C_{\max}/C_{\min}$, since $\omega_{0\max}/\omega_{0\min} = (C_{\max}/C_{\min})^{1/2}$ for a fixed inductance.

14.3.6 Varactor Diodes

As we saw in Section 3.5.1, a PN junction has a built-in *E* field, which sweeps charge carriers out of the neighborhood of the junction, leaving a depletion region. It is the free carriers that conduct electricity, so the depletion region is effectively a dielectric. How far the carriers move depends on the strength of the field, so that applying a reverse bias causes the depletion region to widen; since the conducting regions effectively move apart, the junction capacitance drops. Devices that are especially good at this are called varactors, but it occurs elsewhere; a PIN photodiode's C_d can drop by 7× or more when the die is fully depleted.[†]

By steepening the doping profile in the junction region, $C_{\max}$ can be increased, increasing the CR to as much as 15 in hyperabrupt devices, with Q values in the hundreds up to at least 100 MHz for lower capacitance devices (the Q gets better at higher voltages). NXP and Zetex are the major suppliers of varactors below 1 GHz—check out the BBY40 or MMBV609 and their relatives. The tuning rate is highest near 0 V—empirically, C goes pretty accurately as $\exp(-0.4V)$ for hyperabrupt devices and $(V + V_0)^{-2}$ for abrupt-junction ones[‡] (these approximations are good to a few percent over the full voltage range). Curve fitting can get you closer, but since you'll use digital calibration (and perhaps a phase-locked loop) anyway, this accuracy is usually good enough.

Keep the tuning voltage above 2 V or so for best Q and linearity, and watch out for parametric changes (where the signal peaks detune the varactor), which cause distortion and other odd effects. These parametric changes can be greatly reduced by using two matched varactors in series opposing (many are supplied that way), so when the signal increases one, it decreases the other, and the series combination stays nearly constant. Note also that the tuning voltage has to be very quiet.

The tuning curve of a varactor can be linearized by strategically chosen series and shunt inductances, as we'll see in Section 15.12.2, and its temperature coefficient is highish, about +150 ppm/°C in ordinary (CR = 2) devices to +400 ppm/°C in hyperabrupts.

14.3.7 Inductors

An inductor in its simplest form is just an N-turn helical coil of wire, called a solenoid. The **B** fields from each turn of the helix add together, which increases the magnetic energy stored (the energy density goes as B^2) and so the inductance goes as N^2. (Numerically the inductance is the coefficient L of the energy storage equation $\mathcal{E} = 1/2\, LI^2$, and $\mathcal{E}$ is proportional to the volume integral of B^2, so inductance isn't hard to calculate if you need to.)

Inductance can be increased by a further factor of 1.2 to 5000 by winding the coil around a core of powdered iron or ferrite, with the amount of increase depending on how much of the flux path is through air ($\mu_r = 1$) and how much through the magnetic material ($\mu_r = 4$ to 5000). This reduces ohmic losses in the copper, as well as helping to confine the fields (and so reducing stray coupling). The hysteresis losses in the core itself are normally much smaller than the copper losses would otherwise be. Unfortunately, μ is a strong function of B for ferromagnetic materials, so adding a core compromises linearity, especially at high fields, where the core will saturate and the inductance drop like a stone. In a high power circuit, the resulting current spikes will wreak massive destruction if you're not careful.

Toroidal coils, where the wire is wound around a doughnut-shaped core, have nearly 100% of the field in the ferrite, and so have high inductance together with the lowest loss and fringing fields at frequencies where cores are useful (below about 100 MHz). They are inconvenient to wind and hard to adjust, since the huge permeability of the closed-loop core makes the inductance almost totally independent of the turn spacing.

[†] If you're using a tuned photodiode front end, you can use the bias voltage as a peaking adjustment.

[‡] This characteristic produces nice linear tuning behavior, if the varactor is the only capacitance in an LC circuit.

There is some stray inductance; due to the helical winding, an N-turn toroid is also a 1-turn solenoid with a core that's mostly air. Clever winding patterns can more or less eliminate this effect by reversing the current direction while maintaining the helicity.

Pot cores interchange the positions of the core and wire (a two-piece core wraps around and through a coil on a bobbin) and are good for kilohertz and low megahertz frequencies where many turns are needed. Pot cores are easily made with a narrow and well-controlled air gap between the pole pieces. The value of $1/B$ goes as the integral of $(1/\mu) \cdot ds$ along the magnetic path, so a 100 μm gap can be the dominant influence on the total inductance of the coil. This makes the inductance and saturation behavior much more predictable and stable, and allows the inductance to be varied by a small magnetic core screwed into and out of the gap region.

High frequency coils are often wound on plastic forms, or come in small ceramic surface mount packages, with or without magnetic cores. The plastic ones usually have adjustable cores that screw into the plastic. Make sure the plastic form is slotted, because otherwise the high thermal expansion of the plastic will stretch the copper wire and make the inductance drift much worse.

A perfect inductor has a reactance $Z_L = jX_L = j\omega L$, which follows from its differential equation, $V = L\, dI/dt$. Inductors are lossier (i.e., have lower Q[†]) than capacitors, due to the resistance of the copper wires, core hysteresis, and eddy currents induced in nearby conductors. Loss is not always bad; it does limit how sharp you can make a filter, but lossy inductors are best for decoupling, since they cause less ringing and bad behavior. Ferrite beads are lossy tubular cores that you string on wires. While they are inductive in character at low frequency, they have a mainly resistive impedance at high frequency, which is very useful in circuits. Ferroxcube has some nice applications literature on beads—note especially the combination of their 3E2A and 4B beads to provide a nice 10–100 Ω resistive-looking impedance over the whole VHF and UHF range.

If you need to design your own single-layer, air-core, helical wire coils, the inductance is approximately

$$L(\mu\mathrm{H}) \approx \frac{a^2 N^2}{9a + 10b}, \tag{14.4}$$

where N is the total number of turns, a is the coil radius, and b is its length, both in inches.[‡] This formula is claimed to be accurate to 1% for $b/a > 0.8$ and wire diameter very small compared to a. It is very useful when you need to get something working quickly, for example, a high frequency photodiode front end or a diplexer (see the problems in the Supplementary Material). For high frequency work, the pitch of the helix should be around 2 wire diameters, to reduce the stray capacitance and so increase the self-resonant frequency of the inductor. You can spread out the turns or twist them out of alignment to reduce the value or squash them closer together to increase it; this is very useful in tuning prototypes. Make sure you get N right, not one turn too low (remember that a U-shaped piece of wire is 1 turn when you connect the ends to something). Thermal expansion gives air-core coils TCs of + 20 to + 100 ppm/°C depending on the details of their forms and mounting.

[†] Roughly speaking, Q is the ratio of the reactance of a component to its resistance. It has slightly different definitions depending on what you're talking about; see Section 14.3.9.

[‡] Frederick Emmons Terman, *Radio Engineers' Handbook*, McGraw-Hill, New York, 1943, pp. 47–64, and *Radio Instruments and Measurements*, US National Bureau of Standards Circular C74, 1924, p. 253.

Inductance happens inadvertently, too; a straight d centimeter length of round wire of radius a centimeters in free space has an inductance of about

$$L = (2\text{ nH})d\left(\ln\left(\frac{d}{a}\right) - 0.16\right), \tag{14.5}$$

which for ordinary hookup wire (#22, $a = 0.032$ cm) works out to about 7 nH for a 1 cm length.

14.3.8 Variable Inductors

Old-time car radios were tuned by pulling the magnetic cores in and out of their RF and LO tuning inductors.† Like capacitors, though, most variable inductors are for trimming; high inductance ones are tuned by varying the gap in a pot core, and low inductance ones by screwing a ferrite or powdered iron slug in and out of a solenoid.

You usually need trimmable inductors to make accurately tuned filters, but surface mount inductors are becoming available in tolerances as tight as 2%, which will make this less necessary in the future. This is a good thing, since tuning a complicated filter is a high art (see Section 16.11).

14.3.9 Resonance

Since inductors and capacitors have reactances of opposite signs, they can be made to cancel each others' effects, for example, in photodiode detectors where the diode has lots of capacitance. Since X_L and X_C go as different powers of f, the cancellation can occur at only one frequency for a simple LC combination, leading to a resonance at $\omega_0 = (LC)^{-1/2}$. If the circuit is dominated by reactive effects, these resonances can be very sharp. We'll talk about this effect in more detail in Chapter 15, because there are a lot of interesting things you can do with resonances, but for now they're just a way of canceling one reactance with another.

14.3.10 *L*-Networks and *Q*

One very useful concept in discussing reactive networks is the ratio of the reactance of a component to its resistance, which is called Q. In an LC tank circuit (L and C in parallel, with some small r_s in series with L), Q controls the ratio of the 3 dB full width to the center frequency ω_0, which for frequencies near ω_0 is

$$Q \approx \frac{\omega_0 L}{R_s} = \frac{1}{\omega_0 R_s C} = \sqrt{\frac{L}{R_s^2 C}}. \tag{14.6}$$

If we transform the LR series combination to parallel, the equivalent shunt resistor R_p is

$$R_p = R_s(Q^2 + 1). \tag{14.7}$$

†For readers not old enough to remember, these were the first pushbutton radios—you'd set the station buttons by tuning in a station, then pulling the button far out and pushing it back in; this reset the cam that moved the cores in and out. It was a remarkable design for the time; you didn't have to take your eyes off the road to change stations.

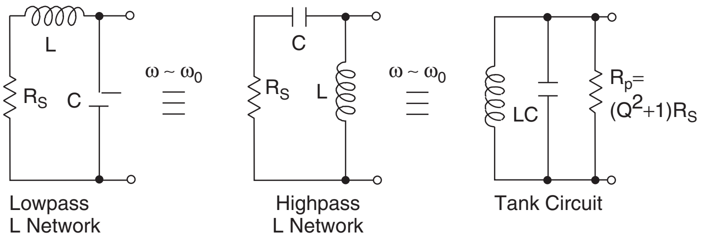

Figure 14.4. In the high-Q limit, near resonance, an L-network multiplies R_s by $Q^2 + 1$ to get the equivalent R_p. The three forms have very differently shaped skirts.

This is an L-network impedance transformer, the simplest of matching networks. We'll meet it again in Chapters 15 and 18. (See Figure 14.4.)

Aside: Definitions of Q and f_0. The competing definitions of Q and the resonant frequency f_0 make some of the literature harder to read. The most common is the reactance ratio, $Q = X_L/R = \omega L/R$ (assuming that the loss is mostly in the inductor, which is usually the case), but the center frequency to 3 dB bandwidth ratio is also commonly quoted ($Q = f_0$/FWHM). These are obviously different, because X_L is not constant in the bandwidth. Similarly, in a parallel-resonant circuit f_0 may be taken to be (a) the point of maximum impedance; (b) the point where the reactance is zero; or most commonly, (c) the point where $X_L = X_C$, which is the same as the series resonance of the same components. At high Q, all of these are closely similar, but you have to watch out at low Q.†

14.3.11 Inductive Coupling

It's worth spending a little time on the issue of inductive coupling, because it's very useful in applications. Figure 14.5 shows two inductors whose **B** fields overlap significantly. We know that the voltage across an inductor is proportional to $\partial\Phi/\partial t$, where Φ is the total magnetic flux threading the inductor's turns, and the contributions of multiple turns add. Thus a change in the current in L_1 produces a voltage across L_2. The voltage is

$$V_{2M} = M\frac{\partial I_1}{\partial t}, \tag{14.8}$$

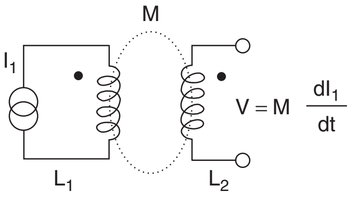

Figure 14.5. Coupled inductors.

† See F. E. Terman, *Radio Engineer's Handbook* (1943), pp. 135–148.

where the constant M is the mutual inductance. The ratio of M to L_1 and L_2 expresses how tightly coupled the coils are; since the theoretical maximum of M is $(L_1 L_2)^{1/2}$, it is convenient to define the *coefficient of coupling* k[†] by

$$k = \frac{M}{\sqrt{L_1 L_2}}. \tag{14.9}$$

Air-core coils can have $k = 0.6$–0.7 if the two windings are on top of each other, ≈ 0.4 if they are wound right next to each other, with the whole winding shorter than its diameter, and ≈ 0.2 or so if the winding is longer than its diameter. (Terman has lots of tables and graphs of this.) Separating the coils decreases k. Coils wound together on a closed ferromagnetic core (e.g., a toroid or a laminated iron core) have very high coupling—typical measured values for power transformers are around 0.99989. The coefficient of coupling K between resonant circuits is equal to $k(Q_1 Q_2)^{1/2}$, so that the coupling can actually exceed 1 (this is reactive power, so it doesn't violate energy conservation—when you try to pull power out, Q drops like a rock).

14.3.12 Loss in Resonant Circuits

Both inductors and capacitors have loss, but which is worse depends on the frequency and impedance level. For a given physical size and core type, the resistance of a coil tends to go as L^2 (longer lengths of skinnier wire), so that even if we let them get large, coils get really bad at low frequencies for a constant impedance level; it's hard to get Q values over 100 in coils below 1 MHz, whereas you can easily do 400 above 10 MHz. The only recourse for getting higher Q at low frequency is to use cores with better flux confinement and higher μ (e.g., ferrite toroids and pot cores).

Capacitors have more trade-offs available, especially the trade-off of working voltage for higher C by using stacks of thin layers; the ESR of a given capacitor type is usually not a strong function of its value. The impedance level at which inductor and capacitor Q values become equal thus tends to increase with frequency.

14.3.13 Temperature Compensating Resonances

The k of some ceramics is a nice linear function of temperature, and this can be used for modest-accuracy temperature compensation. Capacitors intended for temperature compensation are designated N250 (−250 ppm/°C), N750, or N1500; the TCs are accurate to within 5% or 10%, good enough to be useful but not for real precision. A resonant circuit can be temperature compensated by using an NPO capacitor to do most of the work, and a small N750 to tweak the TC (see Problem 14.3 at http://electrooptical.net/www/beos2e/problems2.pdf .). In a real instrument you'd use a varactor controlled by an MCU as part of a self-calibration strategy, but for lab use, N750s can be a big help.

14.3.14 Transformers

The normal inductively coupled transformer uses magnetic coupling between two or more windings on a closed magnetic core. The magnetization of the core is large, because it is

[†] Yes, we're reusing k yet again, for historical reasons.

the **B** field in the core threading both the primary and secondary windings which produces the transformer action (see Section 14.5.4 for a different transformer principle). Because of the tight confinement of the field in the core, k is very high: around 0.999 or better in low frequency devices. At RF, μ is smaller, so the attainable coupling is reduced. The strong coupling allows very wide range impedance transformations with very low loss: an $N{:}M$ turns ratio gives an $N^2{:}M^2$ impedance ratio. For instance, if we have a 10-turn primary and a 30-turn secondary, connecting 50 Ω across the secondary will make the primary look like a 5.5 Ω resistor over the full bandwidth of the transformer. The lower frequency limit is reached when the load impedance starts to become comparable to the inductive reactance; the upper limit, when interwinding capacitance, core losses, or copper losses start to dominate. In an ordinary transformer the useful frequency range can be 10 or even 100 to 1. The phase relationships of transformer windings matter: if there are multiple windings, connecting them in *series aiding* will produce the sum of their voltages; *series opposing*, the difference.

14.3.15 Tank Circuits

A tapped inductor functions like a transformer with the primary and secondary wired in series aiding. It looks like an AC voltage divider, except that the mutual inductance makes the tap point behave very differently; a load connected across one section looks electrically as though it were a higher impedance connected across the whole inductor. A parallel resonant circuit with a tapped inductor or a tapped capacitor can be used as an impedance transformation device; the tapped capacitor has no M to stiffen the tap, of course, so for a given transformation ratio you have to use smaller capacitances and therefore a higher Q than with a tapped inductor. Such a network is called a *tank circuit* because it functions by storing energy. Terman discusses tanks and other reactive coupling networks at length.

14.4 TRANSMISSION LINES

A transmission line is a two-port network whose primary use is to pipe signals around from place to place. You put the signal into one end, and it reappears at the other end, after a time delay proportional to the length of the line. Coaxial cable is the best known transmission line, but there are others, such as metal or dielectric waveguide, twin lead, and microstrip (like PC board traces over a ground plane). Like light in a fiber, the signal can be reflected from the ends of the transmission line and bounce around inside, if it is not properly terminated.

Transmission lines can be modeled as a whole lot of very small series inductors alternating with very small shunt capacitors. The line has an inductance L and capacitance C per unit length. The two define a characteristic impedance Z_0,

$$Z_0 = \sqrt{\frac{L}{C}}, \tag{14.10}$$

which for a lossless line is purely real.[†] An infinite length of line looks electrically exactly like a resistor of value Z_0 connected across the input terminals. If we sneak in

[†] There is a good analogy between transmission line impedance and the refractive index in optics. The formulas for reflection and transmission at a transmission line discontinuity are the same as the normal-incidence forms of the Fresnel formulas, when Z is replaced by $1/n$.

one night and cut off all but a finite length, replacing the infinite "tail" with a resistor of value Z_0, no one will notice a difference—there is no reflection from the output end of the line, so the input end has no way of distinguishing this case from an unbroken infinite line.

The voltage and current at any point on the line are in phase for a forward wave and 180° out of phase for a reverse wave—this is obvious at DC but applies at AC as well—which is what we mean by saying that the line looks like a resistor. We can find the instantaneous forward and reverse signal voltages V_F and V_R by solving the 2 × 2 system:

$$V_F = \frac{V + Z_0 I}{2} \quad \text{and} \quad V_R = \frac{V - Z_0 I}{2}. \tag{14.11}$$

14.4.1 Mismatch and Reflections

If the cable is terminated in an impedance Z different from Z_0, the reflection coefficient Γ and the normalized impedance $z = Z/Z_0$ are related by

$$\Gamma = \frac{z-1}{z+1}, \quad z = \frac{1+\Gamma}{1-\Gamma}. \tag{14.12}$$

This transformation maps the imaginary axis (pure reactances, opens, and shorts) onto the unit circle $|\Gamma| \equiv 1$, and passive impedances ($\mathrm{Re}\{Z\} > 0$) inside.

For example, if the far end of the cable is short-circuited, the voltage there has to be 0; this boundary condition forces the reflected wave to have a phase of π, so that the sum of their voltages is 0. At an open-circuited end, the current is obviously 0, so the reflection phase is 0;[†] the voltage there is thus twice what it would be if it were terminated in Z_0—just what we'd expect, since the termination and the line impedance form a voltage divider.

Seen from the other end of the line, a distance ℓ away, the reflection is delayed by the round trip through the cable, so its phase is different;[‡] this leads to all sorts of useful devices. Using (14.12), with an additional phase delay $2\theta = 4\pi\ell/v$, where v is the propagation velocity of the wave in the line, we can derive the impedance Z' seen at the other end of the cable:

$$Z' = Z_0 \left(\frac{z + j\tan\theta}{1 + jz\tan\theta} \right). \tag{14.13}$$

The one-way phase θ is called the *electrical length* of the cable. The impedance at any point on the line derives from successive round-trip reflections, and thus repeats itself every half wavelength, as in Figure 14.6. We know that a short length of open-circuited transmission line looks like a capacitor, so it is reasonable that the same line would look inductive when short-circuited, which is what (14.13) predicts. However, it may be less obvious that when $\theta = \pi/2$, a shorted line looks like an open circuit, and vice versa. This *quarter-wave section* turns out to be very useful in applications.

[†] The fields actually stick out into space a bit on the ends, which gives rise to a slight phase shift. The situation is somewhat similar to total internal reflection at a dielectric interface, except that here the coupling between the transmission-line mode and the free-space propagating modes is not exactly 0, so there is some radiation.

[‡] We're doing electrical engineering here, so a positive frequency is $e^{j\omega t}$ and a phase delay of θ radians is $e^{-j\theta}$.

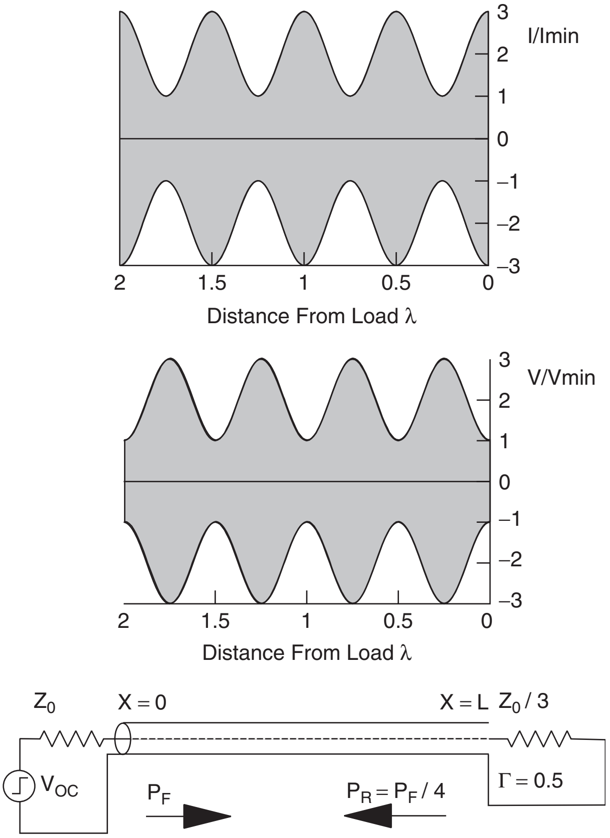

Figure 14.6. A transmission line can have forward and reverse waves propagating simultaneously.

The ratio of the incident to reflected wave powers is known as the return loss, RL $= 20\log_{10}(1/\Gamma)$, and is an important parameter in RF circuit design. A well-matched load might have a return loss of 25 dB or so, and an amplifier is often only 10 dB. The other commonly heard parameter is the *voltage standing wave ratio* (VSWR, pronounced "vizwahr"), which is the ratio of the peak and valley amplitudes of the standing wave pattern produced by the reflection,

$$\text{VSWR} = \frac{1+|\Gamma|}{1-|\Gamma|}. \tag{14.14}$$

See Table 14.2 for VSWR/RL conversion values.

14.4.2 Quarter-Wave Series Sections

A quarter-wave section is just a chunk of line in series with the load, whose electrical length is $\pi/2$. It's the electronic equivalent of a single-layer AR coating. Taking the limit

TABLE 14.2. Conversion Table from VSWR to Return Loss

VSWR	RL (dB)	VSWR	RL (dB)	VSWR	RL (dB)
1.01:1	46.1	1.25:1	19.1	5.00:1	3.5
1.05:1	32.3	1.50:1	14.0	3.00:1	6.0
1.10:1	26.4	2.00:1	9.5	10.0:1	1.7

of (14.13) as $\theta \to \pi/2$, we get

$$Z' = Z_0 \frac{1}{z} = \frac{Z_0^2}{Z}, \tag{14.15}$$

so that a short turns into an open, an open into a short, and reactances change in sign as well as in magnitude. This means that a capacitive load can be matched by using a quarter-wave section to change it into an inductive one, then putting in a series or parallel capacitance to resonate it out. If you have a resistive load R_L, you can match it to a resistive source R_S with a quarter-wave section of $Z_0 = (R_S R_L)^{1/2}$, just like the ideal λ/4 AR coating.

The AR coating analogy holds for more complicated networks as well; as in a Fabry–Perot interferometer, the discontinuity producing the compensating reflection to cancel the unwanted one can be placed elsewhere, providing that the electrical length is correct modulo π. Also as in a Fabry–Perot, the transmission bandwidth at a given order shrinks as the electrical length between the two reflections increases.

14.4.3 Coaxial Cable

Coax is familiar to almost everybody—a central wire with thick insulation, and an outer (coaxial) shield of braid, foil, or solid metal. You ground the shield and put your signal on the center conductor. Due to the dielectric, the propagation velocity is $c/\epsilon^{1/2}$ (just like light in glass). Current flows in opposite directions in the shield and center conductor, so there is no net magnetic field outside. The current in fact flows along the inside surface of the shield at high frequencies due to skin effect.

It's obvious that since the shield covers the inner conductor completely, there's no field outside, and so no crosstalk (i.e., unintended coupling of signals) can occur. Except that it does. This is another of those cases where failing to build what you designed produces gotchas. Most coax has a braided shield, which doesn't completely cover the inner conductor, and whose integrity depends on good contact occurring between wires merely laid on top of each other. You wouldn't trust that in a circuit, and neither should you here: single-shielded braided coax leaks like a sieve. RG-58A/U is especially bad—it talks to everybody in sight. If you're using coax cables bundled together, use braid-and-foil or double-shielded coax such as RG-223 for sensitive or high powered lines. For use inside an instrument, you can get small diameter coax (e.g., Belden 1617A) whose braid has been completely soaked with tin, forming an excellent electrical shield that is still reasonably flexible and highly solderable, and has matching SMA connectors available. (For some reason it has become very expensive recently—a few dollars per foot—but it's great stuff.) For microwaves, where the skin depth is small and therefore base-metal shields are lossy, there is semirigid coax (usually called hardline), made with a solid copper tube for an outer conductor. Special connectors are required for these lines, although you can solder them directly to the board as well. Solid-shield coax also has much better phase stability than braided—at 50 MHz and above, jiggling a cable can cause a few degrees' phase shift, which is obnoxious in phase-sensitive systems. Even in nominally amplitude-only setups, the resulting change in the relative phases of cable reflections can cause instability reminiscent of etalon fringes.

If a cable has multiple ground connections (e.g., a patch cord strung between two connectors on the same chassis), the return current will divide itself between them according to their admittances, which will reduce the shield's effectiveness at low frequency and cause ground loops (see Section 16.5.2).

Even at low frequency, coax has its problems. Flexing it causes triboelectric noise, where charge moves from shield to insulation and back, like rubbing a balloon on your hair. And flexing causes the cable capacitance to change slightly, which turns any DC bias into noise currents as the cable capacitance changes.

14.4.4 Balanced Lines

One of the primary functions of a transmission line is to keep the signal fields confined to the interior of the line, or at least to its immediate neighborhood. The mathematical way of saying this is that the overlap integral between the transmission line mode and any propagating wave outside the line has to be 0. This can be achieved by 100% shielding, as in waveguide or coax, or more elegantly by using balanced line such as TV twin-lead or twisted pair. These balanced lines work by making sure that a current i flowing in one conductor is balanced by a current of $-i$ flowing in the other. By symmetry, this guarantees that the AC voltages are v and $-v$ as well. An arm-waving way of looking at this is to consider the asymptotic falloff of the signal. A finite length of wire with an AC current flowing in it is an antenna: it produces a far-field amplitude that falls off asymptotically as r^{-1}, so that its energy flux across any large sphere is independent of radius. Two wires with equal and opposite currents will produce a far-field amplitude proportional to ${r_1}^{-1} - {r_2}^{-1}$. If the wires are separated by a distance $2d$ in the X direction, we get

$$E \propto \frac{1}{\sqrt{(x-d)^2+y^2+z^2}} - \frac{1}{\sqrt{(x+d)^2+y^2+z^2}} = \frac{2d\cos\theta}{r^2} + O(r^{-3}), \quad (14.16)$$

where $x = r\cos\theta$. This falls off faster than a propagating wave and thus can have no propagating component whatever, since you can make the energy flux as small as you like by choosing a big enough sphere. This is relevant to the coupling of adjacent lines, as we'll see in Section 16.3.3. A crucial and usually overlooked precondition for this to work is that the currents must really sum to 0. Failing to obey this rule is the usual reason for failure with balanced lines. If the two currents do not sum to 0, you can decompose them into balanced and common-mode parts:

$$i_{\text{bal}} = \frac{i_1 - i_2}{2} \quad \text{and} \quad i_{\text{CM}} = \frac{i_1 + i_2}{2}. \quad (14.17)$$

The balanced part works as before, but the common-mode part flows in the same direction in both wires; the two terms in (14.16) add instead of subtracting, so from a far-field point of view, it's just like having only one wire—that is, an antenna. (In fact, folded dipole antennas are commonly made from just this sort of balanced line—see the ARRL handbook.)

Since an appreciable field exists outside a balanced line, it is much more vulnerable to unintended coupling from very nearby objects than coax is. It doesn't like being near anything metal, in particular.

14.4.5 Twisted Pair

One particularly common balanced line is twisted pair. It is usually made in the lab by putting one end of a pair of hookup wires in the chuck of a hand drill, and spinning it while pulling gently on the other end (reverse the drill for a moment to avoid kinking when you release the tension). The twists serve to keep the pair at constant spacing and balance capacitive pickup; they also minimize the net loop area available for low frequency magnetic pickup, because the differential voltage induced in adjacent half-turns cancels. Twisted pair really works for that, provided it's truly used balanced. Too many people use it as a fetish instead and are surprised when it doesn't provide signal isolation.

Do not ground either side of a twisted pair. To reiterate: if you ground one conductor, you have built an antenna, not a transmission line. Use coax for unbalanced lines and twisted pair for balanced ones.[†]

14.4.6 Microstrip

An ordinary PC board trace of width w, with dielectric of thickness d between it and the ground plane, is an unbalanced transmission line called microstrip. You can look at it as a balanced line split in half, with the other side replaced by an image in the ground plane. Its characteristic impedance depends on the ratio of the width of the line to the thickness of the dielectric (the w/d ratio) and the dielectric constant of the material. You can make 50 Ω lines on ordinary G-10 or FR-4 fiberglass epoxy board ($\epsilon \approx 4.5$) by making $w/d = 2$. More detailed formulas are given in Section 16.2.2.

14.4.7 Termination Strategies

To avoid all reflections, both ends of the line must be terminated in Z_0. The problem with this is that we only get half the open-circuit voltage at the other end, and we use a lot of power heating resistors. On the other hand, if only one end is properly terminated, any reflections from the other will go away after only one bounce, and so no resonances will occur. Accordingly, we can *series-terminate* coax, where only the driving end is at Z_0. If we send a step function of open-circuit voltage V_{oc} down such a line, it initially sees a matched load of Z_0, so the voltage is $V_{oc}/2$. When the pulse reaches the (open-circuited) far end, it is reflected with $\Gamma = 1$, so the voltage rises to V_{oc}, and the step function starts back toward the source. (Since the forward and reflected wave arrive simultaneously at the open end, there is a single smooth step there.) When the reflected wave gets back to the source, it is not reflected, and the whole line is at a constant V_{oc}. Thus a series-terminated line can transmit the entire open-circuit voltage without annoyance from the reflection, provided only that the driver can handle the reflected signal without difficulty. (See Figure 14.7.)

[†] Oh, all right. The magnetic pickup reduction still applies if you ground one side. Just don't ground it on both ends, and don't say you weren't warned.

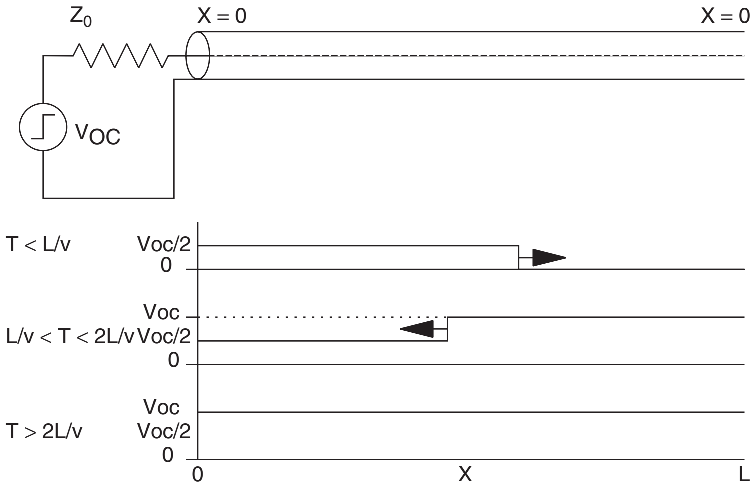

Figure 14.7. A series-terminated (or back-terminated) line. Although points along the line see the effects of the reflection, the open-circuited end sees a nice clean step, as though it were connected right to the source.

Aside: What Is the Worst Case? In testing a transmission system, we often want to verify that it is stable and works acceptably with a "worst case termination." However, what the worst case is may not be obvious. A line driver may work well into a mismatched 10 foot cable, and into a mismatched 1000 foot cable, but oscillate madly with a 250 foot cable. The loss in the very long cable makes it start to look like an infinite (i.e., resistive) line, whereas that 250 foot cable has both a huge delay and a strong reflection.

14.5 TRANSMISSION LINE DEVICES

14.5.1 Attenuators

Because a matched line looks like a resistor, we can use simple resistor networks to attenuate signals. They are three-element ladder networks (pi or T), so that we can set the attenuation, R_{in}, and R_{out} independently. Attenuators are usually called *pads*, because they reduce the effect of outside influences such as mismatches; if the source is isolated from the outside with an N dB pad, any reflected power has to cross it twice before reaching the source again; the return loss seen by the source thus goes up by $2N$ dB (provided the pad itself is that well matched—don't expect a 60 dB pad to have a 120 dB return loss). This is often useful to stop an amplifier from oscillating when badly mismatched. Don't use this idea on your low level circuitry, though, or the SNR will suffer; an N dB pad in front of an amplifier with a 3 dB noise figure will present an $N + 3$ dB noise figure to the world. (See Figure 14.8.)

14.5.2 Shunt Stubs

Since shorted pieces of coax can be made to look like capacitors and inductors, we can use them in matching networks. In particular, a shunt stub is very useful in resonating out

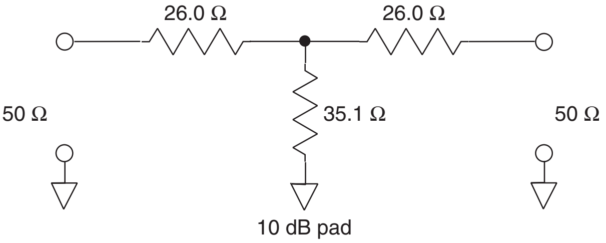

Figure 14.8. Three resistors give enough degrees of freedom to match specified input and output impedances as well as arbitrary attenuation values. This is a 10 dB tee-network attenuator for 50 Ω.

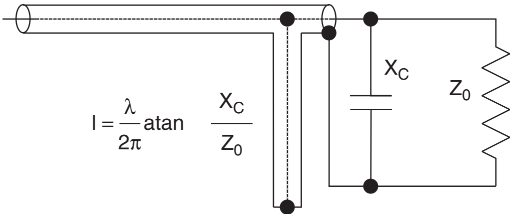
$l = \frac{\lambda}{2\pi} \operatorname{atan} \frac{X_C}{Z_0}$
X_C
Z_0

Figure 14.9. A shunt stub can tune out any reactance at a single frequency.

reactive loads, as shown in Figure 14.9. It's very convenient, too: put a coax patch cord on a tee connector, and stick a thumbtack through the shield into the center conductor, trying different places along the cable until your reactance goes away. If you'll be needing the setup for a while, cut the coax and solder the short. Thumbtack shorts are surprisingly stable: for temporary use, just leave the tack in there.

14.5.3 Trombone Lines

If you don't like thumbtacks, or need an adjustable series section, the line length can be adjusted with a telescoping coaxial section called a trombone line. New ones work best, as the contact gets flaky with age.

14.5.4 Transmission Line Transformers and Chokes

If you put a signal into a transmission line, it should emerge from the other end. The ends of the line have two terminals each, like resistors, and so can be wired in series or parallel with other circuit elements (e.g., termination resistors). The idea of a transmission line transformer is to wire two or more lines in parallel at one end and in series at the other. As shown in Figure 14.10, connecting two lines this way makes an input impedance of $Z_0/2$ and an output impedance of $2Z_0$, a 4:1 impedance transformation ratio. Series–parallel combinations can in principle achieve any rational impedance transformation value.

This sounds like getting something for nothing, because after all these supposedly independent Z_0 resistors are in fact connected together with wires, which will try to short out our nice idea; if you look carefully at it, our 4:1 transformer is a dead short at DC. On the other hand, if the sections are $\lambda/4$ long, the dead short at one end will look like an open at the other. (Should they be $\lambda/4$ in air or inside the line? Why?)

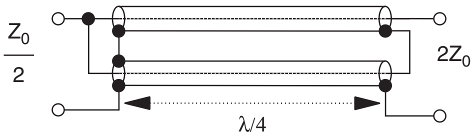

Figure 14.10. A 4:1 coaxial transformer.

Another thing to notice is that the current which acts to short out the transformer flows only on the shield. Its corresponding **E** and **B** fields therefore do not cancel in the region outside the line. We can prevent it from flowing by using a common-mode choke, which is nothing more than wrapping a few turns of line around a toroid core (both coax and twisted pair work well). Such a choke will kill the short circuit and allow us to build wideband transmission line transformers on this parallel-in, series-out principle. The differential mode signal produces no magnetization in the core, and so feels no inductance. The absence of core magnetization is their defining property, and because there are no eddy currents, hysteresis, or other magnetic losses to worry about, transmission line transformers are extremely stable, wideband, and low in loss.

14.5.5 Directional Couplers

As we saw in Section 8.3.3, if we couple two long lines together (e.g., by putting them side by side with slightly leaky shielding between them), a forward wave in one will couple into a forward wave in the other. Providing the interaction region is many wavelengths, it discriminates well between forward and reflected waves, because only the forward wave is phase matched. Such directional couplers are the easiest way to measure the forward and reflected power in your circuit. At frequencies where such long lines are inconvenient, we can take advantage of the opposite phase relationships of voltage and current for forward and reflected waves, making a balanced transformer device that does the same job.

14.5.6 Splitters and Tees

A tee connector wires two cables in parallel, creating an impedance mismatch and consequent reflection. A splitter is a transformer-based device that eliminates the problem and provides isolation between the two taps: as long as the input is matched to 50 Ω, reflections at one output don't affect the other.

14.6 DIODES AND TRANSISTORS

14.6.1 Diode Switches

We all know that a diode conducts electricity in one direction and not in the other. Its conductivity is a smooth function of applied voltage, going from 0 to a large value as the bias voltage goes from large reverse bias to about 1 V forward (anode positive). The forward current of a diode is approximately predicted by the *diode equation*,

$$I_F \approx I_S(e^{V_F/V_\gamma} - 1). \tag{14.18}$$

It is sometimes maintained that V_γ is the thermal voltage $V_{\text{th}} = kT/e$, but that isn't so for real diodes; at 300 K, V_{th} is 25.7 mV but V_γ is 30–50 mV for most devices. Diode-connected transistors (base shorted to collector) are an exception, so if you need accuracy, use those instead of real diodes. In any case, the conductance of a diode is approximately I_F/V_γ, so with a few milliamps of forward current you can get impedances of 10 Ω or so—not that great compared with modern FETs, but with nice low capacitances.

Although (14.18) is wrong for $V_F \ll V_\gamma$ due to second-order effects, it is accurate enough around zero bias that we can calculate the zero-bias shunt impedance r_0 by differentiating,

$$r_0 = V_\gamma / I_S \tag{14.19}$$

which in general is far from infinite—and since I_S increases exponentially with temperature, it can get down to the 10 kΩ range at 125 °C—beware of your protection diodes in high impedance circuits (see Section 14.6.3).

Diodes are ideal current switches: they pass currents in the forward direction but not in the reverse direction. A reverse-biased diode looks like a small capacitor. To take advantage of this ideal behavior, design circuits that are not sensitive to the exact voltages involved (e.g., AC-coupled, current mode, or translinear circuits).

You can get cheap diodes (PIN and Schottky) that have very low capacitance when reverse biased, and which conduct hard in forward bias; these are good for switching AC signals.[†] Unlike electromechanical or FET switches, their control currents are not separate from the signal, but if you're using DC to switch an AC signal, you can separate them again with a capacitor, as in Figure 14.11. Don't worry about the DC voltage offset unless you're at DC or the switching waveform has components that land in your passband.

Another way of getting rid of the offset is with balanced circuits, for example, the series diode clipper and diode bridge of Figure 14.12. In understanding these circuits, think in terms of current, and of the diodes as ideal current switches; matched diodes

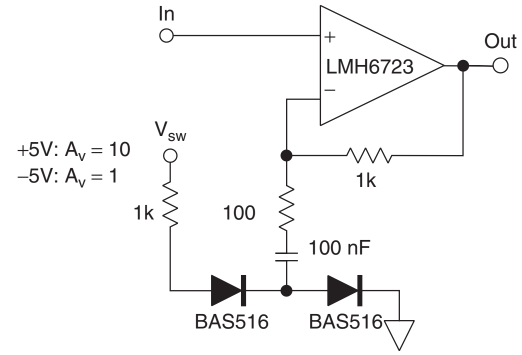

Figure 14.11. If you can separate the control current from the signal, diodes make excellent switches. Here two diodes make a gain switch for a current feedback amp.

[†] PIN diodes store lots of charge in their junctions, which has to be swept out before the diode turns off; this means that their conductance is modulated less on each half-cycle, and so they need less bias current for a given IP_2.

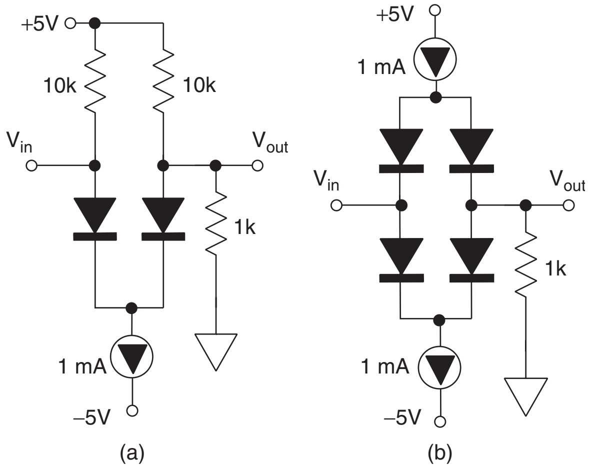

Figure 14.12. Another way of separating the switching current from the signal is to use balanced circuits: (a) series clipper and (b) diode bridge.

work best and are vital in the bridge configuration. This four-diode circuit is often used as an ultrafast sampling bridge; the diodes are biased off normally, but a fast pulse applied in place of the DC bias turns the switch on momentarily, charging up a capacitor on the output; this is how most sampling scopes work.[†]

For switching, Schottky diodes are best if your switch waveform may vary, as in diode mixers; PN diodes store charge, and so change speed with drive level, whereas Schottkys don't.

Aside: Diode Foibles. It's worth reemphasizing that odd diode pathology, their surprisingly low impedances when used near zero bias (e.g., as the protection diodes of a sensitive FET amplifier). We ordinarily think of a diode with zero bias as an open circuit, but in fact for small signals, it's a resistor of value $R_{\mathrm{d0}} = \partial I_F/\partial V_F \approx V_\gamma/I_0$; with gold-doped diodes such as 1N914s and 1N4148s, this can be as low as a few kilohms at elevated temperature, which starts to matter even with microvolt signals. Base−collector junctions of common transistors such as 2N3904s (or MMBT3904s) are much less leaky. For really demanding applications, wide bandgap diodes such as LEDs are enormously better for this use, because their much higher V_F for the same I_F means that I_0 is many orders of magnitude smaller. (Would you believe < 100 fA for biases between -5 V and $+0.5$ V?) Section 18.7.2 uses this nice property in a pyroelectric detector front end.

Diodes exhibit delays in turning on as well as off, and these delays vary widely among devices. A 1N4001 rectifier needs hundreds of microseconds, 1N914s a few nanoseconds, and 1N5711 Schottkys less than 1 ns. It is odd to see a diode circuit overshoot by a volt on turn-on, but some devices will do this. If you find one, complain to the manufacturer or change device types.

[†] Because of their very low duty cycles, sampling scopes precharge the capacitor to the previous value, using a feedback loop called the *sampling loop*, which greatly improves their accuracy.

14.6.2 Bipolar Transistors

The most flexible active device in electronics is the bipolar junction transistor (BJT). It is basically a three-terminal, voltage-programmable current source; the current between the collector and emitter terminals is controlled by the base–emitter voltage, V_{BE}, and is nearly constant over a very wide range of bias conditions. It has excellent characteristics and its highly repeatable from unit to unit, provided you use it properly. We can't devote nearly enough space to this remarkable device, so by all means look it up in Gray and Meyer. Instrument designers are always needing the occasional perfect circuit, but one that isn't available in the IC catalogs; thus we'll look at the BJT's virtues for signal processing.

The oldest, simplest, and still most useful mathematical description of a BJT is the Ebers–Moll model, which (in a somewhat simplified form, valid for normal bias[†] only) predicts that the collector current I_C is a simple function of V_{BE}:

$$I_C = I_S \exp \frac{eV_{\mathrm{BE}}}{kT}, \tag{14.20}$$

where e is the electron charge, k is Boltzmann's constant, and T is absolute temperature. When computing the gain, we care more about the transconductance

$$g_M = \frac{\partial I_C}{\partial V_{\mathrm{BE}}} = \frac{e}{kT} I_C, \tag{14.21}$$

which is I_C/25 mV at room temperature, a very high value.[‡] The exponential character is accurate to better than 0.1% over at least four decades of collector current in a good device, and the exponential constant is $V_T = kT/e$ to very high accuracy; what does vary unit-to-unit is I_S (and of course T, if we're not careful—another opportunity for common centroid design).

The collector and emitter currents are not exactly equal, because the base draws some small current. The current gain $\beta = I_C/I_B$ and is usually 30–500 in common devices, but (unlike g_M) varies widely between devices of the same type. It is thus a bad idea to design circuits that depend on β having some particular value. One other figure of merit for amplifier transistors is β linearity, that is, how much β changes with I_C—it may be flat to 5% over five orders of magnitude, or vary 2:1 over one order. We'll be very concerned with that in Chapter 18 when we talk about laser noise cancelers.

14.6.3 Temperature Dependence of I_S and V_{BE}

The Ebers–Moll model predicts that V_{BEon} is

$$V_{\mathrm{BE(on)}} = \frac{kT}{e} \ln \frac{I_C}{I_S}, \tag{14.22}$$

[†] Normal bias means reverse-biasing the CB junction and forward-biasing the BE junction.

[‡] No semiconductor device can have a higher transconductance than this, because it is limited by the thermal spread of the Fermi level in the semiconductor.

which looks proportional to T, but actually has a nearly constant -2 mV temperature coefficient from nitrogen temperature to 150 °C. The reason for this is that I_S is a strongly increasing function of T, roughly proportional to[†]

$$I_S \propto T^{3.25} \exp\left(\frac{-eV_{G0}}{kT}\right), \tag{14.23}$$

where V_{G0} is the zero-temperature bandgap of the semiconductor (1.205 V for silicon), and the constant 3.25 is somewhat device dependent. Above nitrogen temperature, this produces a TC of -1.8 to -2.2 mV/°C in V_{BEon},[‡] and since the zero-bias $r_0 = V_T/I_S$ decreases rapidly with T, we have to start worrying about shunt resistances at high temperatures. This is a problem mostly with diodes and diode-connected transistors. Note especially the extremely strong dependence of I_S on V_{G0}: high bandgap devices such as LEDs and SiC photodiodes have very high shunt resistances, whereas low V_{G0} IR diodes like InAs and InSb have very low r_0 and need cooling.

14.6.4 Speed

The β of a BJT tends to roll off as $1/f$ for large f (and therefore has a phase near $-90°$). The frequency at which $|\beta| = 1$ is called the *cutoff frequency*, f_T. As a first approximation, β goes as f_T/f down to where it flattens out at its DC value. The frequency f_T is a fairly strong function of collector current, going as I_C for low currents, up to a broad peak somewhere around a third of its $I_{C\,\max}$ spec limit. It also improves with increasing bias voltages, which reduce the interelectrode capacitances.

14.6.5 Biasing and Current Sources

That -2.1 mV/°C temperature coefficient is a nuisance sometimes. Since I_C goes up by a factor of e in 26 mV at 300 K, I_C goes up by 9%/°C if V_{BE} is fixed (the increase in V_T reduces it by only 0.3%). If the increased power dissipation from that 9% extra current raises the temperature another 1°C, the bias will be unstable, and the resulting *thermal runaway* will melt the device.

BJTs also exhibit Early effect, where I_C increases with collector–emitter voltage V_{CE}, and furthermore, different devices have different I_S, which makes accurate current sources hard to build. In addition, the collector current predicted by (14.20) exhibits full shot noise, which has to be dealt with in low noise design.

These problems are easily fixed by negative feedback, as shown in Figure 14.13. Feedback can be applied from collector to base with the emitter grounded, but a more common way to do it is by putting a resistor in series with the emitter. This technique is known as *emitter degeneration*.[§] Without getting mathematical, if the emitter resistor drops 2 V, I_C drifts only 0.1%/°C instead of 9%/°C. Problem 14.8 (at http://electrooptical.net/www/beos2e/problems2.pdf) has more on this. Similarly, the shot noise current produces an opposing collector–emitter voltage that reduces the shot noise current by the same factor.

[†] Paul A. Gray and Robert G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 2nd ed. Wiley, Hoboken, NJ, 1984, p. 340.

[‡] The TC increases as collector current goes down, reaching as much as -3 mV/°C at picoamp collector currents.

[§] Positive feedback is what keeps oscillators running, and so is said to be regenerative; negative feedback is then logically degenerative. No moral judgment is implied.

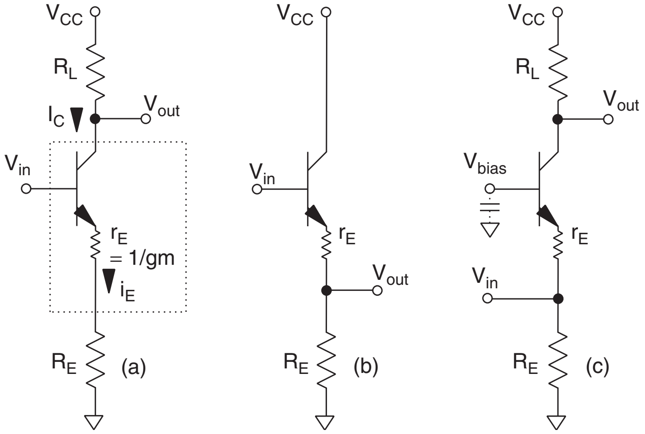

Figure 14.13. Transistor amplifier configurations: (a) common emitter, high A_V, medium Z_{in}; (b) common collector or emitter follower, high Z_{in}, low Z_{out}, $A_V = 1$; and (c) common base, low Z_{in}, high A_V. Emitter degeneration due to R_E stabilizes the bias and suppresses nonlinearity in each case.

14.6.6 Cutoff and Saturation

We often run transistors in nonlinear modes: cutoff, where the collector is biased correctly but V_{BE} is small or negative, so no collector current flows; or saturation, where $I_C R_L$ is so big that the CB junction becomes forward biased (about 200 mV of forward CB bias is the real limit of normal bias conditions). Transistors can go in and out of cutoff very fast, but come out of saturation very, very slowly, so avoid saturating if you need speed. The origin of this effect is that pulling V_{CE} below V_{BE} reduces the E field in the base region so much that the carriers coming in from the emitter don't fly into the collector as normal, but hang around in the base until they recombine, which shows up as a huge increase in the base current. All the stored charge has to be swept out before the transistor can turn off again, which is why it's slow.

If you drive the base hard enough ($I_B \approx 0.1 I_C$), you can get V_{CEsat} down to 50–100 mV in small signal devices, and around 0.4 V in power transistors. The CE bias necessary to avoid saturation is roughly proportional to absolute temperature (+0.33%/°C at room temperature), so check your circuits at high and low temperatures to make sure they still work properly.

Aside: Inverted Transistors. There is a sometimes-useful trick to get it lower: run the transistor upside down. That is, exchange the collector and emitter. The emitter is normally much more highly doped than the collector; this asymmetry makes the β of an inverted transistor very low (2–5, typically), but the V_{CEsat} can go as low as 10 mV.

14.6.7 Amplifier Configurations

Transistors are flexible devices; a BJT amplifier can use its base or emitter as input and its collector or emitter as output; since you need two terminals per port and there are only three available, one has to be common to both input and output; it can be any one of the three. (Being the control electrode, the base always has to connect to the input port.)

Figure 14.13a shows the collector as output, the common-emitter amplifier. Neglecting the Early effect, it has medium input impedance of $R_{\text{in}} = (1 + \beta)(R_E + r_E)$, high output impedance of $R_{\text{out}} = R_L$, and medium voltage gain, $A_V = R_L/(R_E + r_E)$. The BJT's shot noise contribution is reduced by a factor of $r_E/(R_E + r_E)$.

If the emitter is used as the output in Figure 14.13b (the common-collector or *emitter follower* connection), the same negative feedback that stabilized the bias stabilizes the output voltage, producing a very low output impedance ($R_{\text{out}} = R_L \parallel r_E$), higher input impedance (still $(1 + \beta)(R_E + r_E)$, but R_E is usually several times bigger). It has nearly unity voltage gain ($A_V = 1 - r_E/R_E$) since the emitter follows the base closely and (for the same reason) has a shot noise voltage contribution of $A_V i_{N\text{shot}} R_E$. BJTs follow their noise models very closely (see Section 18.4.5).

Finally, as in Figure 14.13c, if we keep the base still (common-base), any signal current i_E we shove into the emitter will reappear at the collector; the collector swing is then $R_L i_E$, whereas the emitter moves only $(r_E \parallel R_E)i_E$. For frequencies where the source impedance remains high, there is no additional shot noise contribution from the BJT. The low noise and near-total lack of voltage swing at the emitter makes the common-base amplifier very useful in reducing the effects of capacitance, as in *cascode* amplifiers and photodiode front ends (see Section 18.4.4 for a great deal more on these points.)

Example 14.1: Capacitance Multiplier. Figure 14.14 shows what we've already adumbrated: a slow *RC* lowpass filter with an emitter follower makes a really good supply filter for low level circuitry. The input impedance of the follower is about $(\beta + 1)R_L$, which means that for a given sized capacitor, the *RC* corner frequency can be $\beta + 1$ times lower, and the filtering correspondingly better. Considering that an MPSA-14 Darlington has a $\beta \approx 20,000$, this is a big deal. This improvement can be traded off for reducing the drop across the resistor, so a β of 300 can be apportioned as 30 times lower f_c and 10 times lower *IR* drop. If you need real help, try splitting the input resistor in two, and putting another capacitor to ground from the middle, as in Figure 14.14c. You can reach 100 dB suppression at low kilohertz frequencies with a couple of 22 kΩ resistors and 1

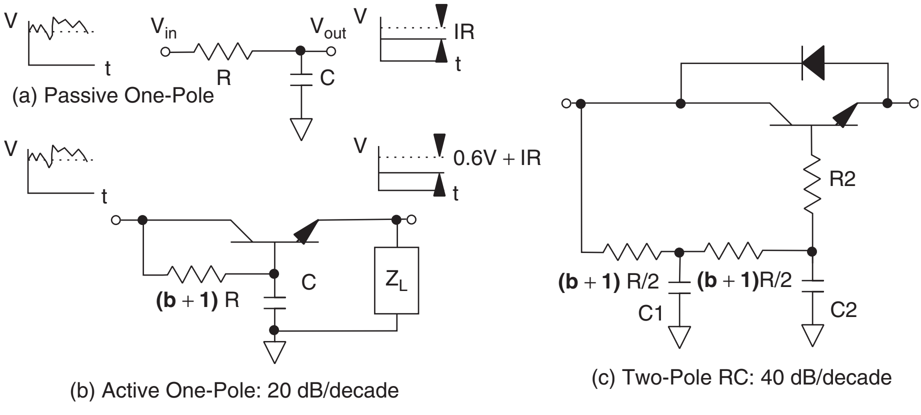

Figure 14.14. Capacitance multipliers. (a) *RC* lowpass, whose rejection is limited by the allowable voltage drop. (b) one-pole capacitance multiplier: the rejection improves $\beta + 1$ times. (c) Realistic two-pole version: the diode and R2 are for protection from input shorts, which otherwise will destroy the transistor (make sure the bypasses on the circuits you're driving are at least 10^6 times bigger than the diode capacitance).

μF capacitors. For the highest suppression, you can replace the single diode with two diodes with a big bypass capacitor in between.

Somewhat similar tricks can be played with currents. For instance, by taking the circuit of Figure 14.14b and connecting the capacitor between the base and emitter of the transistor, you can make a simulated inductor.

14.6.8 Differential Pairs

One of the most important elementary circuits is the differential pair, formed by connecting the emitters of two (notionally identical) BJTs, as shown in Figure 14.15. The Ebers–Moll model predicts that collector currents I_{C1} and I_{C2} obey

$$\frac{I_{C2}}{I_{C1}} = \exp\left(\frac{e\Delta V_{\mathrm{BE}}}{kT}\right), \tag{14.24}$$

which specifies the ratio of I_{C1} and I_{C2} but doesn't depend on how big they are: an ideal bipolar diff pair is a perfect voltage-controlled current splitter. (The TC of I_S cancels out.) This comes in very handy, especially in noise canceling front ends (see Section 18.6.3). Since ΔV_{BE} need not be large (±60 mV gets you a 90:10 current ratio at 300 K), the currents can be switched without large voltage swings. If the input swing is a bit bigger than this (e.g., 500 mV p-p), the diff pair is an excellent current switch; this is how emitter-coupled logic (ECL) works.

When building differential amps, we usually care more about the difference between I_{C1} and I_{C2}, since that's what the output voltage depends on; it is

$$I_{C2} - I_{C1} = I_{\mathrm{Bias}} \tanh\left(\frac{e\Delta V_{\mathrm{BE}}}{2kT}\right). \tag{14.25}$$

Usually you bias a diff pair with a current source, because r_E of each transistor serves as the other one's emitter resistor, and using a current source keeps the magnitudes of I_{C1} and I_{C2} independent of the common-mode base voltage $V_{\mathrm{CM}} = (V_{B1} + V_{B2})/2$. You can also use a large resistor if you don't need good common-mode rejection away from $\Delta V_{\mathrm{BE}} = 0$.

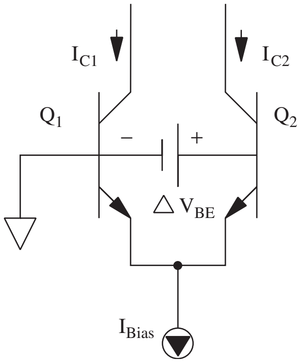

Figure 14.15. A BJT differential pair makes a near-ideal current splitter, since I_{C2}/I_{C1} depends only on ΔV_{BE}.

14.6.9 Current-Mode Circuitry

The diff pair is the simplest example of a *translinear circuit*. Translinear circuits such as the Gilbert cell multiplier of Figure 14.16 use bipolar transistors for current steering, which is what they're best at. The cross-connected differential pairs at the top of the Gilbert cell commutate the output currents of Q_1 and Q_2, making a nearly ideal multiplier.

Like the differential pair, translinear circuits switch at very low input voltages, that is, at a fundamentally minimal circuit impedance for a given power dissipation. This greatly reduces the effect of device capacitances, making them very fast. They're not that easy to debug, but they work very well and are especially suitable for low supply voltages. A typical example is an operational transconductance amp with linearizing diodes on its inputs (e.g., the LM13700).

You have to watch out for shot noise in these circuits, because the transistor junctions generate it, and it takes feedback to suppress it. In the limit of infinite β, if the emitter current of a differential pair has exactly full shot noise, each collector current also has exactly full shot noise, independent of the splitting ratio. Thus if you don't keep copying the current with undegenerated current mirrors,[†] there's no serious buildup of shot noise in the signal path.

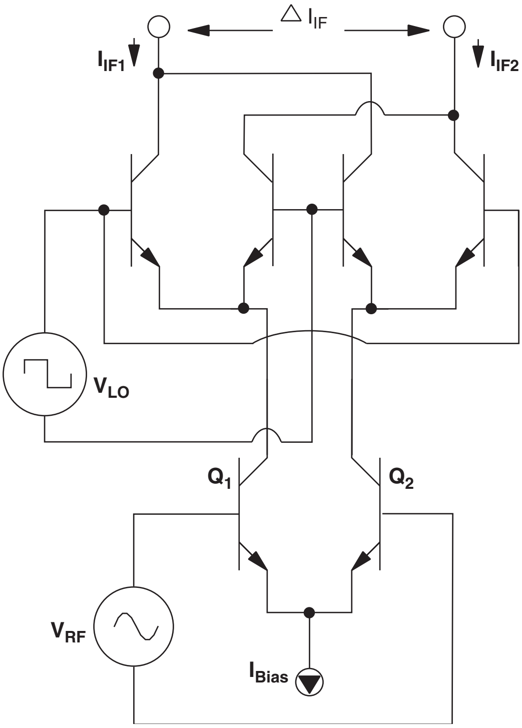

Figure 14.16. Gilbert cell multiplier.

[†] A current mirror uses a diode-connected BJT to provide V_{BE} for another identical device; assuming $\beta \gg 1$, the second BJT's collector current will be a replica of the first one's, with additional full shot noise.

Aside: FETs. It may seem strange that FETs haven't been given equal time with BJTs, especially with the number of FET integrated circuits around. The reason is that apart from isolated instances, e.g., using power FETs for switching, dual-gate GaAs FETs for isolation amplifiers, or MOSFETs for zero input current buffers, small-signal discrete FETs have little to recommend them for circuit design. Besides low transconductance, FETs have disgracefully poorly specified biasing properties; a typical discrete N-channel JFET has a threshold voltage $V_{gs(\mathrm{Th})}$ specified as -2 to -5 V (a typical BJT might have a range of 50 mV). That means that your source follower circuit has an offset that may vary by 3 V from device to device, and it only gets worse if you want voltage gain. Not many circuits can conveniently accommodate that sort of sloppiness, especially on a single 5 V supply. Dual matched JFETs are a partial exception, but good luck finding them these days.

There do exist some nice quiet FETs that are good for some things. For audio frequencies, the Toshiba 2SK368 is very good, with 1 Hz v_N below 0.8 nV/Hz$^{1/2}$, but it has input capacitance $C_{gs} = 80$ pF and feedback capacitance $C_{dg} = 15$ pF, which is fairly horrible. At RF, the BF862 makes good bootstraps and discrete input stages for TIAs. The author has designed with them several times, but they never seem to make the final cut.

Heterojunction FETs (HJFETs) such as the NE3508 series have much better DC specs than JFETs and are about 100 times faster than a BF862, even—ten times the g_m and a tenth the C_{dg}. At 2 GHz, the NE3509 can reach noise temperatures of 30 K at room temperature in a properly tuned amplifier.

14.7 SIGNAL PROCESSING COMPONENTS

We'll talk about the ins and outs of ICs and modules shortly, but first of all it's worth making a high level tour of how to choose components wisely.

14.7.1 Choosing Components

Make sure you design your circuit around widely available components, which are likely to be obtainable and not be discontinued for the life of your product. The author has had his favorite unique components discontinued a number of times, especially the TL011 series of three-terminal current mirrors, which were hard to replace, the MRF966 dual-gate GaAs FET, the MRF9331 micropower RF transistor, the AD639 sine converter, the MC13155 FM IF chip.... Especially dubious are devices intended for specific fast-changing markets, for instance, read channel amplifiers for hard discs or DVD player motor control ICs. The likelihood of that chip being available in five years is next to zilch.

14.7.2 Read the Data Sheet Carefully

There's a lot of data in a data sheet. Some absolutely crucial things are usually specified in small print in one corner of it; for example, the variation of phase margin versus feedback resistance in current feedback amplifiers, the tendency for multiplexers to connect all their inputs together if even one is overdriven, FET op amps' habit of pegging their outputs when their input common-mode range is exceeded, the maximum allowable voltage drop between analog and digital ground on an ADC, that sort of thing. Read the data sheet carefully, and don't ignore the stuff you don't understand; look it up. Bob Pease of National Semiconductor once wrote an article on reading data sheets, which it's worth getting.[†]

14.7.3 Don't Trust Typical Specs

Try to design your circuits so that parts meeting only the minimum spec will work without difficulty. If you need something a bit more special, for example, the typical spec as a guaranteed minimum, you can do that one of two ways; first, get the manufacturer to test them for you, which is expensive, or order a couple of months' stock ahead of time, and test them yourself. The extra inventory is to cover you when you get a batch that is good, but doesn't meet the typical spec; keep a month's worth of known good parts in a safe somewhere for emergencies.

Above all, don't trust any spec in a data sheet marked "Preliminary" or "Advance Information." By way of example, the MC34085 quad JFET op amp had its typical GBW and slew rate numbers reduced by nearly a factor of 2 between the advance information and product release data sheets. This is embarrassing for the manufacturer but may be catastrophic for you.

14.7.4 Specsmanship

Don't expect that all the nice specifications of a given component will necessarily apply at the same time. It is reasonable for the flatness and accuracy of a switched-capacitor filter chip, for example, to degrade somewhat at the extremes of its specified frequency range. Dig through the data sheet to find out how it works under the actual conditions anticipated, and don't try to push more than one limit at a time. The classical example of this is analog lock-in amplifiers, which have accuracies of 0.1% in amplitude and 0.1°in phase, with a bandwidth of 1 Hz to 100 kHz. You have to read the manual carefully to find out that the near-perfect accuracy specs only apply near 1 kHz, and that the error at the band edges is more like 2 dB and 1 radian (most digital ones are better).

14.7.5 Watch for Gotchas

The second most important reason to learn about discrete circuit design is to be able to understand IC data sheets, especially the "simplified schematics" that farsighted manufacturers like National print in their data sheets (at least for bipolar parts). You can spot a lot of circuit funnies that way, for example, how the output swing will vary with the supply voltage, and whether you can do anything about it. In data sheets especially, if you notice something strange, stop and figure it out before using that part—treat it as though it were a large machine with unfamiliar controls.

[†] Bob Pease, How to get the right information from a datasheet, Appendix F of *Operational Amplifiers Data Book*. National Semiconductor Inc., 1988–1993. Reprinted in *Troubleshooting Analog Circuits*, Butterworth-Heinemann, Woburn, MA, 1991.

14.7.6 Mixers

The main application of diode switching is in diode ring double-balanced mixers (DBMs, not to be confused with dBm). This simple but remarkably effective combination of four matched Schottky diodes and two toroidal transformers is shown in Figure 14.17.

The diodes are arranged differently than in a bridge rectifier; both ports conduct on both half-cycles. The RF port transformer produces 0° and 180° phases of the weak RF input, and the strong LO signal switches the diodes to commutate between these two phases. The output is the resulting current; since it flows 180° out of phase through both halves of both transformer secondaries, it suffers no inductance and no coupling to the RF or LO port—that's why the mixer is called double-balanced.†

How well a mixer does this separation trick is specified in its LO–RF and IF–RF isolation numbers. Typically, a good double-balanced mixer will have LO–RF isolation of 30–45 dB, and LO–IF isolation of 20–35 dB. This is of more than cosmetic importance. If we have unintended signal paths, such as the LO of one mixing stage going out its RF port and getting into the previous mixing stage, the signals will get intermodulated in the other stage and cause spurious signals.

A diode bridge mixer typically has a conversion loss of about 5 dB, so that the desired IF is 5 dB below the RF input level. Its noise figure is usually approximately the same as its conversion loss, which means that it contributes very little noise of its own. The upper end of its dynamic range is set by intermodulation; its input-referred $P_{1\mathrm{dB}}$ is typically $P_{\mathrm{LO}} - 7$ dB.

14.7.7 LO Effects

Mixers typically work well only over a narrow range of LO powers (e.g., the Mini Circuits SRA-1 likes +7 dBm, although it will work from +4 to +10 dBm). The third-order nonlinearity of diode bridges and FET mixers is dominated by the nonlinear conductance of the devices during turn on and turn off. Using square wave drive makes the active devices switch faster, and so improves the mixer's linearity. At frequencies for which good logic devices exist, use gates or inverters to drive the mixer. (Use analog power for that stage, and don't use the output of a clocked logic block such as a counter or flip-flop for anything analog.)

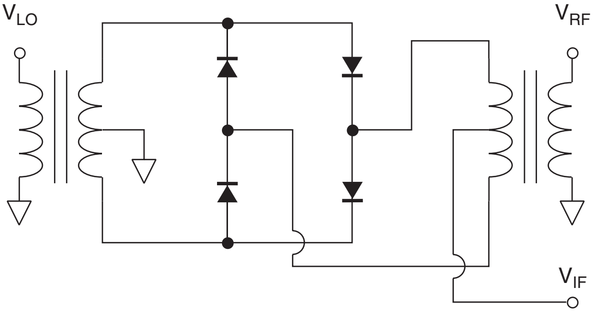

Figure 14.17. Diode bridge double-balanced mixer (DBM).

† Single-balanced mixers keep the LO signal out of the IF and RF, and double-balanced mixers also keep the RF out of the IF.

The symmetry of the mixing characteristic prevents LO amplitude noise from being converted into IF phase noise, and if the mixer is driven hard, such noise won't amplitude modulate the IF, either. On the other hand, decreasing the LO amplitude will cause the diodes to switch more slowly, since it takes more of a cycle to get enough LO current to saturate them. When the diodes are only partly turned on, the RF input can significantly change the diode conductance, causing nonlinearities and attendant signal distortion. This effect is symmetric, so it ideally causes no phase shift, but the increased nonlinearity can make LO noise cross-modulate the signal components, and will certainly increase the intermodulation products. This is not too obnoxious unless the LO is weak and the RF has more than one strong signal.

14.7.8 Op Amps

You've probably designed at least one op amp circuit already, and know the ideal op amp rules, virtual ground, and so forth, so we won't belabor them. (If you haven't, you can learn about them in Horowitz and Hill.) Decent op amps have noise voltages of 0.9 to 30 nV/Hz$^{1/2}$, a 30 dB range. Current noise ranges much more widely, from a few dozen electrons/s/Hz$^{1/2}$ up to tens of picoamps (a 120 dB range). Very low noise voltages come at the price of large noise currents, large bias currents in bipolar amplifiers, and, most insidiously, high input capacitance. This input capacitance loads the summing junction (where feedback is applied) and can cause all sorts of mysterious instabilities. This severely limits the impedance levels you can use; in a front end amplifier, it hurts your SNR with small photocurrents. Chapter 18 has an extended practical example.

The other thing to mention is that the inverting input of an op amp is not really virtual ground, not at AC. If the output swings ±5 V at frequency f, the inverting input will be moving by about (5 V) f/f_c, where f_c is the closed-loop 3 dB bandwidth (the exact value depends on the frequency compensation, but it won't be less than this). If f is 1 MHz and f_c is 10 MHz, the "virtual ground" will be swinging by a good volt peak-to-peak. Whether an op amp can manage that depends on its input circuit—a BJT diff pair can only manage ±60 mV. (This also shows the limits of "virtual ground" in providing isolation between signal sources connected to the summing junction.)

In our current enthusiasm for low supply voltages, op amps whose outputs swing from rail to rail are popular. Their output stages cannot be emitter followers, which lose at least 0.6 V on each side, and so they come from the collectors or drains of the output devices. As we saw earlier, these are high impedance points, and indeed a rail-to-rail op amp has an open-loop output impedance at least an order of magnitude higher than a good emitter-follower output. This makes them much more sensitive to weird loads and funky reactive feedback networks, so beware.

14.7.9 Differential Amps

Voltage is a relative measure, like altitude. All amplifiers measure a voltage difference between the input terminals, and turn it into another voltage between the output terminals. A single ended amplifier has a common reference point for both input and output, whereas a differential amplifier allows both input terminals to swing back and forth, and provides a separate terminal for the output reference voltage.

The figure of merit for a diff amp is its common-mode rejection ratio (CMR or CMRR), which is notionally its differential gain divided by its common-mode gain (inputs shorted). CMR is a subtle thing to measure; Pease gives a good discussion.

Differential pairs make good diff amps, since their outputs move symmetrically and ignore common-mode voltages, but their outputs are up near the supply rail. Another stage is normally required to convert their differential output to a ground-referred single-ended one.

Besides rejection of common-mode signals, a truly differential amplifier should not load down the measurement points, and certainly should not load them asymmetrically. Operational amplifiers (op amps) are good differential amps but have the drawback that their feedback networks are connected directly to their inverting inputs, resulting in low and unbalanced input impedances. An instrumentation amp has good differential performance (120 dB rejection of common-mode signals near DC), together with high input impedance. It works by buffering each input of an op amp with its own noninverting stage, with a clever circuit trick to greatly improve CMR at high gain (see Horowitz and Hill).

At RF, differential amps are often used to reduce coupling via grounds, and for their inherently better overload behavior due to their symmetry.

Gotcha: Floating Sources. Beginners are often frustrated by differential measurements. If you put a floating voltage source such as a transformer secondary or a thermocouple across the inputs of a high impedance differential amplifier such as an instrumentation amp or the inputs of a balanced A/D converter board, it won't work. This is because the amplifier inputs will always source or sink some bias current. In a single-ended measurement, this current flows through the source to ground, causing few problems unless the source resistance is extremely high. In a floating measurement, this current has nowhere to go, because the bias currents of the two inputs will never sum to zero. The result is that the floating source floats toward one of the supply rails until the net current going into it becomes zero, and of course the amplifier ceases to work properly. This leads to even more mysterious problems when there is significant hum on the source (e.g., a remote thermocouple input). The solution is simple, fortunately: put a big (≈ 1 MΩ) resistor from one side to signal ground.

14.7.10 RF Amps

Radio frequency processing is usually done in the context of transmission lines, which have very low characteristic impedances (50–300 Ω, nearly always 50 or 75). RF amplifiers are designed to match into these lines reasonably well, so that they can be strung together with cables and other 50 Ω devices such as power splitters, mixers, attenuators, and directional couplers, to form a complete RF signal processing system. This is a very convenient way to proceed if you don't require low power consumption. These devices come packaged with connectors, or in transistor packages for PC board mounting. Most are *cascadable*, meaning that you can string them together without having them oscillate (unless you do something dumb, like folding the string in a U shape so that the output is next to the input).

14.7.11 Isolation Amps

Ordinary RF amps have resistive feedback, meaning that inside the package, a low resistance is connected from the input to the output. As a result, if you put a signal into the output of an amplifier, some will appear at its input. As a corollary, mismatching the output will change the input impedance. This is usually not too objectionable, but sometimes the reverse signal path will introduce artifacts into your measurement. For example, if you need to produce two signals whose relative phases are accurately specified, having the two leak into one another can cause noticeable phase errors, as we saw in Section 13.6.9. This is often a concern when we are doing time and frequency measurements, which are attractive because of the extreme accuracy that is possible with current technology, or in a buffer stage following a sensitive resonator such as a SAW device or quartz oscillator.

One way to get more isolation is to cascade amplifiers with pads in between. An amplifier with 25 dB isolation followed by a 10 dB pad has 35 dB isolation, at the price of reduced gain and output power. Two sets in a row will get you 70 dB, which is usually enough. This approach is easy and is sure to work, but it is expensive and eats power and board space. It's great for prototyping, because the right combination of connectorized amplifiers and pads can be bayoneted together in a few minutes. Watch out for cable crosstalk, supply coupling, radiation, and other sneak paths when you're doing this—isolation amps plug one leak, but there are always more.

It is possible to build good amplifiers that have almost perfect isolation of the input from the outputs. A garden variety common-source amplifier using a $0.75 dual-gate GaAs FET with no feedback to the gate can achieve 70 dB isolation in one stage at 100 MHz, with very small power consumption and board area. Apply feedback using a source degeneration resistor.

14.7.12 Radio ICs

The are a lot of wireless devices sold, and although a lot of wireless ICs are very specialized, a fair number are useful in instruments, too, especially the FM IFs, mixer/oscillators, and VCO/dividers; they are intended for communications use, and since communications puts surprisingly stringent requirements on RF signal processing, they have to perform pretty well. For a dollar or two, you get a miniature signal processing toolkit that runs from 3 V at 10 mA. Especially useful are the quadrature FM detector and logarithmic meter (RSSI) outputs, which have excellent performance and quick response. Before designing your signal processor, make sure you have the latest RF and wireless product data from NEC, Maxim, On Semiconductor, NXP (formerly Philips), National, and Texas Instruments at least, and that you've spent some time comparing the block diagrams of the parts with the one for your system.

Apart from their meter outputs, which are amazingly useful, the performance of these parts is never as good as you'd get from a collection of Mini Circuits stuff. Their noise figures tend to be around 6 dB, and their BJT mixers don't have as good dynamic range as a high level Schottky diode mixer. This matters a lot in optical instruments, since our dynamic range is often enormous. On the other hand, radio ICs tend to be much cheaper, and they chew up less power and board space. There are often circuit hacks to get round the dynamic range problem; for example, since the radio ICs are so cheap, use two complete systems, one with a 30 dB attenuator in front of it. Digitize both, and use the low gain system when the high gain one rails, as in Section 15.2.1.

14.7.13 Stability

Cascadability and isolation are related to the whole notion of stability of RF amplifiers. Most amplifiers have a significant amount of feedback (intentional or not) from output to input. This will usually have a phase lag. If you put a capacitive load on the output, the feedback will be further delayed, until it starts to have a negative conductance. If this negative conductance gets too large, the net input resistance will go negative at some frequency, and the amplifier will potentially become an oscillator. Amplifiers whose feedback is sufficiently small that this cannot occur are said to be unconditionally stable. Isolation amplifiers are always unconditionally stable. If your amplifier is not unconditionally stable, don't use it for an input or output stage where somebody might hang an unterminated cable, or you'll be sorry. A pad on the input will usually solve the problem at the expense of gain and noise. There's more in Sections 15.4.1 and 18.4.1.

14.7.14 Slew Rate

The output of any device can move only so fast. The maximum *dV/dt* (or *slew rate*) is set by how fast the device's internal bias currents can charge its internal capacitances. Thus, as shown in Figure 14.18, a large-signal transient consists of a straight-line slew followed by an exponential settling from the step response, and possibly thermal and dielectric contributions later.

14.7.15 Settling Time

When the output of any device moves, it takes a while (the *slewing time*) to get there, and another while to settle down to a given accuracy; the time from the input transient to the final entry into the output error band is called the *settling time*.

For sufficiently small signals, the settling time of a linear device is easily deducible from the step response, as we saw with filters in Chapter 13. For larger signals, or with

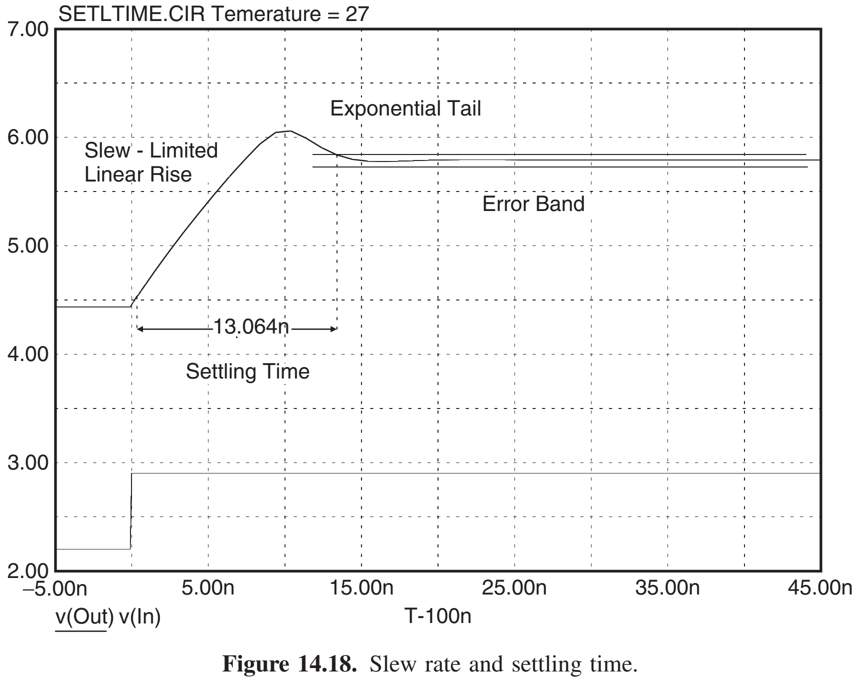

Figure 14.18. Slew rate and settling time.

intrinsically nonlinear or time-dependent devices such as track/holds or DACs, this is not the case—the settling time must be characterized carefully, and surprises may lurk; it is easily possible for important settling transients to occur on time scales 10^4 times longer than the initial slew of the output.

Dielectric absorption and thermal transients within the IC are two of the most common causes of this, but there are also transmission line reflections and load effects to worry about. Make sure you scrutinize the settling behavior of your finished circuit very carefully—those long-tailed effects are not that easy to see, and almost never come out in simulations (SPICE is even worse at multiple time scales than a digital oscilloscope, because it can't skip stuff—it must laboriously follow every wiggle).

14.7.16 Limiting Amplifiers

A limiter is an amplifier that is intentionally severely overdriven, so that its output bangs between its high and low limits on each cycle. A typical limiter stage is one or more differential pairs, as in Figure 14.19, where the collector loads are small enough that the transistors cut off rather than saturating on each half-cycle. Heuristically, a limiter preserves only the timing of the signal's zero crossings, which is related to its frequency.

The relation is not trivial, however; if two signals at slightly different instantaneous frequency are fed into the limiter (e.g., two different FM stations on the same channel), the limiter exhibits a *capture effect*: the frequency of the signal coming out of the limiter is that of the stronger signal, even if it's only a very small amount stronger (1 dB, say). This behavior is a gift, because it allows us to coexist with our spurious signals unusually easily in FM systems. The phasor diagram in Figure 14.20 should make this more plausible (see Section 13.3.2). The two analytic signals are represented by phasors, which are vectors in the complex plane representing the instantaneous signal voltages. As the longer phasor spins around the origin, the smaller one cannot change the average

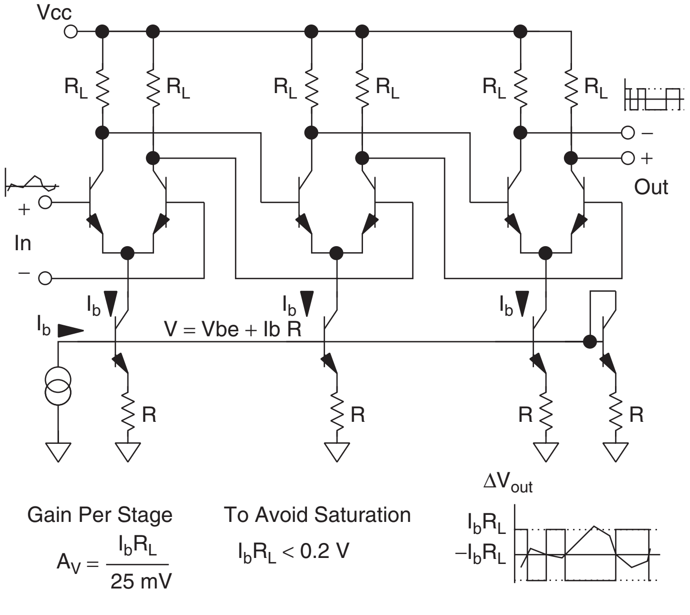

Figure 14.19. Three-stage BJT limiter with ultrasimple biasing.

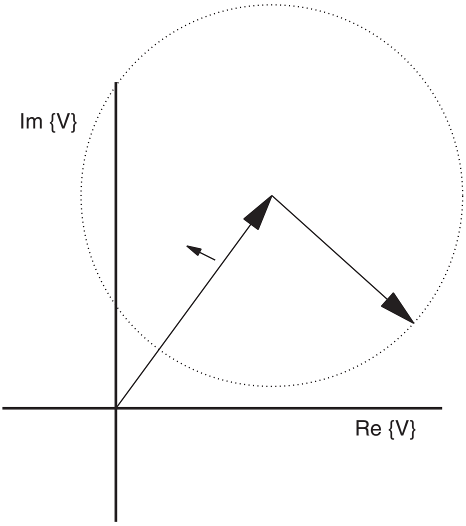

Figure 14.20. Phasor diagram of two signals entering a limiter. The smaller is too short to reach the origin, so the larger sets the output frequency.

rate at which the total signal voltage goes around, because it's too short to reach. Thus the only way it can cause extra zero crossings is to make the total voltage briefly turn round and go backwards enough to cross and recross the imaginary axis. This requires the smaller signal to have a steeper slope than the larger one, and so cannot occur as long as

$$\omega_{\text{sig}} V_{\text{sig}} > \omega_{\text{spur}} V_{\text{spur}}. \tag{14.26}$$

We saw in Section 13.6.9 that additive spurs and noise cause phase jitter, which is preserved by limiting, so do your limiting after the bandwidth setting filter. Bandwidth limiting also prevents extra zero crossings, as shown by (14.26). Note that this is a bit of a two-edged sword; as the SNR in the IF bandwidth drops below 15 dB, you'll get momentary dropouts due to noise peaks bigger than the signal, and when it drops below 0 dB, the limiter will completely suppress the signal in favor of the noise.

Limiters suffer from AM–PM conversion (see Section 13.5.6), so limit at low frequency if possible, where the slowing down of the limiter with strong signals doesn't cause a large phase error. If you build your own, use bipolar diff pairs with small collector loads, as shown in Figure 14.19, so they cut off instead of saturating, which slows them down abysmally. You need overall DC feedback for stable operation, which is not shown but is easily added with an *RC* network.

Majority-carrier devices like GaAs FETs are somewhat better than bipolars here, if they're well enough matched. It's often worthwhile to convert, limit, and convert back if the phase accuracy must be high. Alternatively, a calibration of the AM–PM conversion can be made, and the phase data adjusted afterwards. It is also possible to compensate for it in real time with a phase shifter and some circuitry, if the data rate is too large or the latency requirements too tight for postprocessing. Section 15.7.2 has a good on-the-fly calibration method applicable to limiters.

Linear rectifying amplitude detectors also exhibit a weaker capture effect. Except near the strong signal's zero crossings, the weaker signal mostly just puts wiggles on the DC from the stronger one. Assuming that the two signals are at sufficiently different frequencies that their beat note is outside the baseband frequency response of the system, the effect of the weaker signal is reduced. The weaker signal can dominate the detector's output near the strong signal's zero crossings, so this capture effect isn't as pronounced as with a limiter.

14.7.17 Lock-in Amplifiers

A lock-in is basically a radio that measures the phase and amplitude of its input using two phase detectors driven in quadrature from a reference signal you supply (see Section 13.7.5).

Lock-ins have good dynamic range and contain circuitry that verifies that no stage of the amplifier is being overdriven, provided that the signals are slow enough (most lock-ins will be driven berserk by 2 volts of 100 MHz, but few will inform you of the fact). The value of lock-ins, and their fatal allure, is their apparent ability to conjure signal-to-noise improvements out of nowhere, and to provide phase measurements of high accuracy (0.1°) at frequencies up to 100 kHz. Users who do not expect more than their front ends will provide, and who read the spec sheets carefully, will find these devices very useful. Pay attention to the way the specified accuracy varies with frequency and amplitude, and remember that narrowing the bandwidth by a factor $1/\alpha$ in a continuously swept measurement increases the measurement time by a factor of α^2 (so that 10 dB better SNR costs 100× in measurement time), because you've got α times as many points, each of which takes α times as long to settle). See Sections 13.10.4 and 17.11.5 for a good alternative.

One thing you have to be careful about in lock-in amplifiers is their calibration. Lock-ins using switching phase detectors produce an output corresponding to the average value of the commutated signal voltage, including both the signal frequency and all its harmonics. This means that you aren't just measuring the amplitude of the fundamental signal, and that your noise bandwidth may be quite different from the $1/(4\tau)$ of a continuous-time *RC* rolloff. Lock-ins using analog multipliers really measure just the fundamental. Digital lock-ins may do a better job, but there are no guarantees.

14.8 DIGITIZERS

Nearly all electro-optical instruments use digital data acquisition or digital control for something. We saw how to calculate the noise degradation caused by an ideal digitizer in Chapter 13; here we'll see what a real one looks like.

14.8.1 Digital-to-Analog Converters

The simplest place to begin is with DACs. A DAC in its purest form is a binary-weighted array of current sources or resistors, switched in and out bitwise in order to produce a current or a voltage ratio corresponding to the binary number input. Errors caused by switch impedance are solved by switching currents instead of voltages, or by scaling the die areas of the switches in the MSBs so that the switch resistance terms in the voltage ratio cancel out. A subtler method is the *R–2R network*, where all the resistors have the same value, and the switch impedances therefore don't need to be scaled. The R–2R network can be inverted as well, with identical current sources connected to the taps and the output taken from the V_{ref} pin. A third method, used mainly in 8 bit applications, is to take 256 identical resistors wired in series and select a tap using a gigantic multiplexer and a noninverting buffer. This has the advantage of inherent monotonicity and stability, but tends to be slow.

Besides speed, resolution, and accuracy, the main figure of merit is *glitch energy*, which is the area on the $V(t)$ graph between the actual curve and some (poorly specified) smooth monotonic curve—basically it tells you how much total charge the glitch is going to deliver to the load.[†]

To get high speed and good linearity with low glitch energy, modern DACs often use combinations of these techniques, rather than being pure R–2R or binary-weighted; for example, the Analog Devices AD9760 (12 bits, 160 MS/s), which uses 31 identical current sources for the 5 MSBs, another 15 identical current sources of 1/16 the value for the next 4 bits, and binary-scaled sources for the 3 LSBs. (See Figure 14.21.)

Since the binary input is dimensionless, the scale for the DAC output must be set by an independent reference; thus a DAC inherently multiplies the reference by the given binary fraction to produce the output. A *multiplying* DAC is one that is especially good at it, with good linearity and reasonable bandwidth for V_{ref}.

Get the Analog Devices and Maxim catalogs to see the wide variety of DAC types, and pay close attention to the octal serial DACs such as the AD8801; they're cheap and very useful for trimming, auto-zeroing, and low-speed control applications.

DACs are pretty trouble-free devices, providing you give them a clean, low-Z reference, a jitter-free clock, well-bypassed supplies, and really solid analog and digital grounds (see Chapter 16). There are audio-type 24 bit dual DACs available for a dollar or two, as well as much better ones (e.g., the Cirrus CS4398), which will solve most problems of DAC dynamic range, if not linearity.

One partial exception to this rosy picture is noise. An ideal DAC is a totally noiseless device. Unlike an ADC, a DAC contributes no quantization noise, since its output voltage corresponds exactly to the input digital code. Thus a DAC does not add the

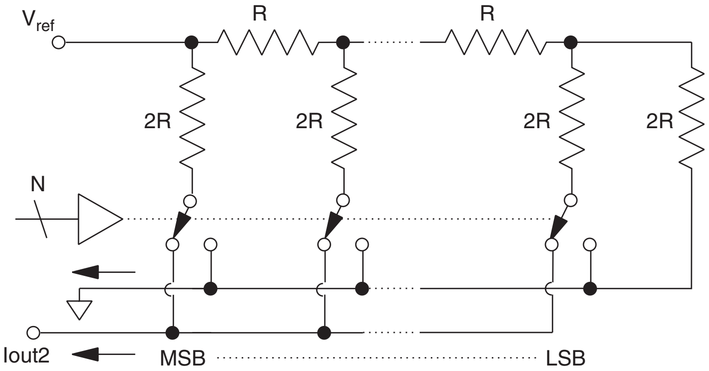

Figure 14.21. R–2R current-output DAC.

[†] You're right, it's not really an energy—it's specified in picovolt-seconds, so it's an impulse.

$1/\sqrt{12}$ ADU quantization noise of an ADC. Of course, since no DAC has perfectly spaced steps, switches instantaneously without jitter or glitches, or produces noiseless levels, the real situation is somewhat worse.[†] Even with constant DC level outputs, voltage-mode CMOS DACs can be significantly noisier than current-mode or bipolar voltage DACs—100–200 nV/Hz$^{1/2}$ noise, with $1/f$ noise corners as high as 10 kHz, which is a disaster in high accuracy situations since it can't be filtered out. Some DACs also exhibit bad popcorn noise. Serial DACs have a lot of clock feedthrough—for best performance, use a parallel-access DAC and an external op amp.

14.8.2 Track/Hold Amplifiers

Track-and-holds[‡] are the workhorses of A/D conversion. The track/hold does just what it says: tracks the signal and then holds its last value when told to. Good track/holds (T/Hs) are very linear, settle rapidly in both states, don't load the signal source unduly, have low and stable output impedances, don't produce big glitches on their inputs or especially their outputs, and of course hold the level without drooping at all. We're still waiting for one like that, but modern T/Hs are very good.

Track/holds are used to hold a voltage so that it can be digitized by an A/D, or to deglitch a DAC. Deglitching requires only modest timing accuracy in the T/H, but of course its own glitches must be small. ADC front ends require excellent timing accuracy and low droop. We expect high slew rates and fast settling during track mode, and good isolation of the stored voltage in hold mode. The crucial timing parameter in a track/hold is the *aperture uncertainty*, or *aperture jitter*.[§] This is the uncertainty in just when the sampling occurs. It's usually random, but not necessarily, as it may depend on signal swings. Manufacturers aren't very good at specifying how they measure this, so we'd better be conservative and assume that the published spec is the rms error, rather than the peak to peak.

We can estimate the voltage noise caused by an rms aperture uncertainty t_{ap} with a full scale signal at f,

$$\left\langle \frac{\Delta V}{V_{\text{FS}}} \right\rangle = \frac{t_{\text{ap}}}{V_{\text{FS}}} \sqrt{\left\langle \left(\frac{dV}{dt} \right)^2 \right\rangle} = \frac{\pi f t_{\text{ap}}}{\sqrt{2}}. \tag{14.27}$$

In order to hold this to less than $1/\sqrt{12}$ ADU, so that it doesn't dominate the noise for large signals, the jitter must be less than

$$t_{\text{ap}} < \frac{2^{-N}}{\pi\sqrt{6}f} \approx \frac{2^{-N}}{8f}. \tag{14.28}$$

For 12 bits at 1 MHz, $t_{\text{ap}} < 30$ ps, even allowing aperture jitter to contribute as much noise as the quantization itself. If you know that you will never have signals varying that

[†] There are also digital signal processing issues such as the use of zero-order holds—see Chapter 17.

[‡] In a simpler age, these used to be called *sample/hold* amps. Now we have to worry more about what happens during acquisition.

[§] Aperture uncertainty is occasionally confused with the *aperture time*, which is the time window in which the sampling occurs. This should stay still, so it isn't the same as jitter.

fast, you can get away with poorer jitter, but don't push it too far—bad performance there can look harmless in an FFT spectrum (merely raising the noise floor a bit), but still have horrendous time-domain artifacts (e.g., glitches and missing codes). Extensive digital postprocessing puts a huge premium on the quality of the data.

The maximum aperture uncertainty is much less than the rise time of the sampling clock edge, and slow or noisy clock edges make it worse. Clock jitter of course adds in RMS fashion as well; if you're using PLL clock recovery, for instance, that jitter counts too. Aperture uncertainty is an almost impossible limitation in fast, high resolution converters; if our 12 bit ADC were 16 bit instead, the jitter spec would be 2 ps. There are T/Hs available that are this good, but remember that this spec has to be met by your entire clocking system. This is really impracticable without great care. It may be sufficient to resynchronize the T/H control line directly from the master clock oscillator using a fast flip-flop (e.g., a 74AC74A), but in some instances the clock has to be filtered first. The timing accuracy of the rest of the system has to be considered too, especially if your high frequency component is ultimately derived from another clock somewhere.

Gotcha: T/H Outputs Are Also Inputs. Track-and-holds tend to be vulnerable to transients at their outputs. To achieve simultaneous sampling of several inputs, it's common to use several T/Hs with a multiplexer following, and a single ADC. If the multiplexer is charged up to a very different voltage, its capacitance will have to discharge very rapidly through our T/H's output pin, and some of this current is liable to find its way onto the hold capacitor, producing a spurious step in the stored voltage. A resistor between the T/H and mux will help, but in any such system, make sure you test for this problem.

14.8.3 Analog-to-Digital Converters

There are lots of different kinds of ADCs, but the ones you'll probably use most are half-flash for high speed and $\Delta\Sigma$ for high accuracy at low speed.

Flash ADCs. The easiest kind of ADC to understand is the flash converter (Figure 14.22), where you have $2^N - 1$ separate comparators (why not 2^N?), with all their $+$ inputs connected together and their $-$ inputs connected to a $2^N - 1$ tap voltage divider. The output is constructed as the binary address of the highest comparator that goes active. Flash converters are very fast, but their complexity increases exponentially with N. You might think that they would have excellent aperture jitter, but it isn't as good as all that, due to a fixed-pattern effect; because of chip gradients and the different source impedance of each tap, the comparators don't all have the same delay. Accordingly, you should use a T/H anyway.

Half-Flash ADCs. A half-flash converter (Figure 14.23) uses a two-stage process where a very accurate $N/2$ bit flash converter gets the MSBs. An equally accurate DAC subtracts their contribution off, leaving a residual that is amplified by $2^{N/2}$ times and digitized by the same or another flash converter, yielding the LSBs. This is complicated and error-prone, but it's worth it for the reduced circuit complexity and power consumption. Modern half-flash units use an extra bit or two in the flash section, plus some on-chip smarts for self-calibration, relaxing the accuracy requirements. As an aside, the DACs inside ADCs are often based on capacitive voltage division instead of resistive, which has precision and power consumption advantages, but can't hold a fixed level.

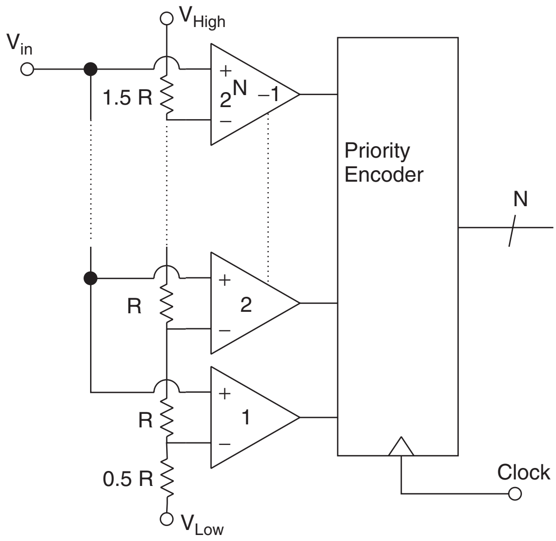

Figure 14.22. Flash analog-to-digital converter.

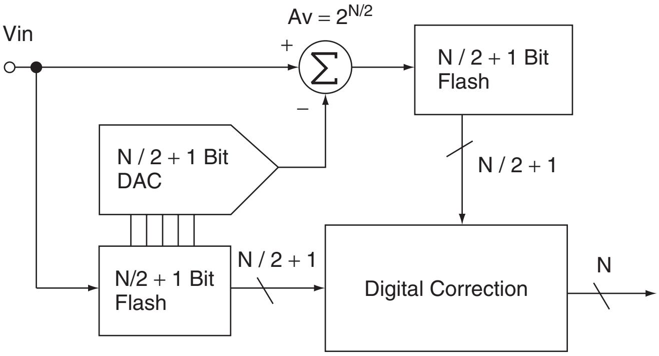

Figure 14.23. Half-flash ADC.

Successive Approximation ADCs. Occasionally, you'll need to use successive approximation, which uses a DAC, a single comparator, and some logic (a successive approximation register, or SAR). In $N + 1$ clock cycles, the SAR constructs the output word one bit at a time (MSB first) by doing a binary search of the DAC's code space to find the exact point at which the comparator switches. The beauty of SAR circuits is that since the step size varies with the uncertainty (cf. slew rate in an analog loop) they are inherently nulling without being exponentially slow for large N as $\Delta\Sigma$ or integrating converters are. This combination makes them good for nulling-type phase and amplitude digitizers, which let us avoid two-dimensional calibrations.

Delta-Sigma ($\Delta\Sigma$) ADC. A $\Delta\Sigma$ ADC, also called a delta-sigma, bitstream, or "1 bit" ADC (Figure 14.24), uses a single comparator in a digital feedback loop, to continually

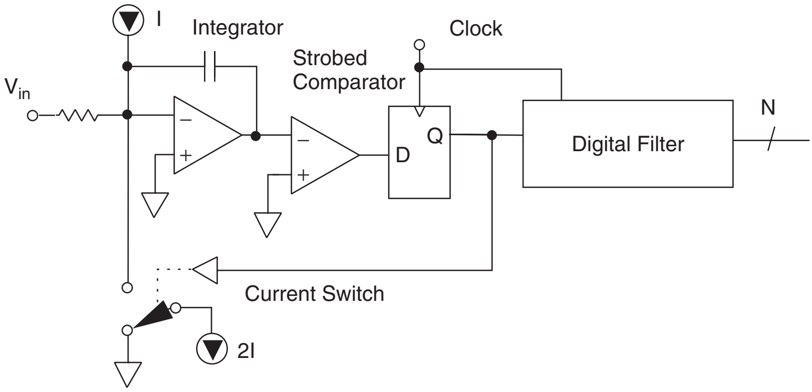

Figure 14.24. Delta-sigma ADC (bitstream ADC).

null out the input signal with a single switched current source.[†] Its output is the duty cycle required, expressed as a binary fraction after digital filtering. Only one threshold and one current source are involved, so there are no matching problems, and the capacitor and comparator voltages are always near 0. Thus its linearity is excellent. Besides, they come with up to 24 bits—8 more than any other technique—so what's not to like?

Due perhaps to their association with "high end" audio, or to the ease of adding lots of extra bits—useless, but always irresistible to marketeers—delta-sigmas have attracted about as much hype and specsmanship as anything in electronics, so you have to be really careful in reading their data sheets. This is a pity, because they're amazingly good for some things. It's partly explained by the fact that the theory of $\Delta\Sigma$s is much richer (read, more difficult) than that of ordinary quantizers because of the quantized, and hence nonlinear, feedback. This means that not all the usual theorems about quantization noise being equally distributed and white apply.[‡]

They aren't very fast; $\Delta\Sigma$ ADCs slow down exponentially with N, needing 2^N clocks per independent measurement. SARs need N clocks, and an analog feedback loop settles to N bits in about $0.7N$ time constants, because the slew rate is proportional to the size of the error, whereas in a $\Delta\Sigma$ or slope converter it is constant. A $\Delta\Sigma$ integrates continuously, so that each measurement merely samples its output. Self-calibration is usually used to null out the current source errors and voltage offsets here too. The slowness of $\Delta\Sigma$ converters limits their usefulness, because you generally can't put a multiplexer in front to sample several different sources, as you can with devices whose conversion time is shorter and better defined. (Some $\Delta\Sigma$s use multibit front-end converters, almost like a half-flash; if more than 1 bit feedback is used, this trades off accuracy for a bit more speed.)

Their linearity is very good; $\Delta\Sigma$s have INLs about $\pm 10^{-5}$ (and typically the rest of your circuit won't be this good). However, a 24 bit A/D has 16 million codes, so 10^{-5} of that is ±160 ADUs. Similarly, they're usually claimed to be monotonic to their maximum resolution, but due to their noise, typically they won't even *sit still* to that accuracy, even

[†] Another oldie in new clothes: H. Inose et al., New modulation technique simplifies circuits. *Electronics* **36** (4), 52–55 (1963).

[‡] Robert M. Gray, Quantization noise in $\Delta\Sigma$ A/D converters, in S. Norsworthy et al., *Delta-Sigma Data Converters*, Wiley-IEEE, Hoboken, NJ, 1996.

with their inputs shorted. (Don't expect a $\Delta\Sigma$ to get anywhere near ADU/$\sqrt{12}$ noise.) So basically your 24 bit $\Delta\Sigma$ is a very nice 18 bit converter with six nearly useless extra bits attached by the marketing department.

Their popularity is explained largely by their low prices, good accuracy, freedom from range switching, and (secretly) the effortless ease with which they allow a designer to finesse a stupidly written accuracy specification by waving their apparent 60-parts-per-billion resolution. This makes them the circuit equivalent of switching to double precision floating-point without changing your algorithm.

14.8.4 DAC and ADC Pathologies

DACs and ADCs never have the nice even staircase characteristic we'd like, and the characteristic degrades somewhat with fast signals. For slowly varying signals, the main types of error are the static *differential* and *integral nonlinearity* (DNL and INL, respectively), defined visually in Figure 14.25 for a DAC. The INL is the maximum deviation of the output voltage from a perfect staircase connecting the upper and lower reference voltages, in units of ADU. The DNL is the deviation of each individual step from being exactly 1 ADU tall, which is easily expressed in finite differences:

$$\mathrm{DNL}(n) = \frac{V(n+1) - V(n)}{2^{-N} V_{FS}} - 1 \tag{14.29}$$

A DNL of ≤ -1 ADU makes the DAC nonmonotonic, which is very undesirable in control applications.

The ADC case is a bit more complicated, because we're looking at Figure 14.25 sideways, coming in with a voltage, and coming out with a number. From what we know about inverse functions, we expect problems if the DAC isn't monotonic; here, a DNL of ≤ -1 ADU at code M will make the inverse function multivalued, and that makes M a *missing code*; a slowly varying input will produce an output that jumps directly

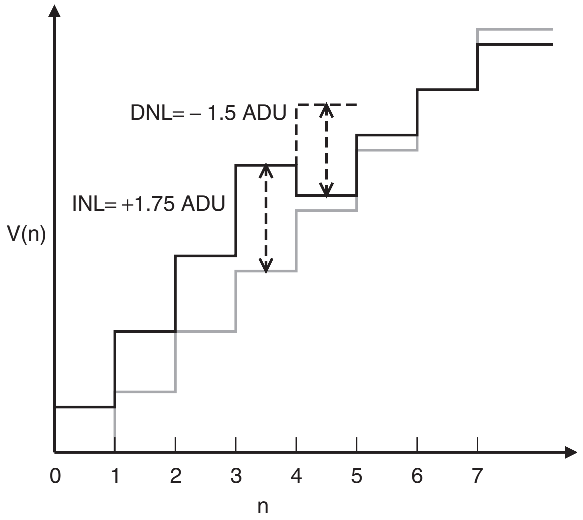

Figure 14.25. DAC pathologies: integral and differential nonlinearity.

from $M - 1$ to $M + 1$ (3 to 5 here), which is even worse than DAC nonmonotonicity, especially in feedback systems, where a set point of M leads to the loop hunting forever. (Big positive DNL leads to deadbands, which are just as bad.) Many of the highest resolution ADCs have many ADU of input noise, which causes them not to sit still, even with their inputs shorted; this smears out the DNL and may make it less significant in servo applications.

INL and DNL have effects identical to putting a nonlinear amplifier in front of a perfect digitizer (or after a perfect DAC). Broadly speaking, DNL causes time-domain glitches and raises the noise floor, while INL causes harmonics and intermod; thus you should test for DNL in the time domain and INL in the frequency domain.

You test ADCs in the frequency domain with two sine waves, just like a two-tone IMD test for an amplifier. Do the frequency analysis with an unwindowed FFT—don't use a DAC plus a spectrum analyzer. Make sure the test signals and sampling clock are very stable, and placed so the tones and IM products land exactly at sample frequencies (e.g., 65,536 samples, with f_1 going through 100.0000 cycles and f_2 103.0000 cycles in that interval—see Section 17.4.7). Phase-lock them together if possible—these conditions force all the products to concentrate where they're easily measured. INL is usually gentle and creates low-lying harmonics that show up well.

Check very carefully for DNL in the time domain, by subtracting off what you ought to have got (samples of two sine waves) and looking at the residuals. It's often a very illuminating test.

The same tests run backwards are good for testing DACs; use a spectrum analyzer to look for the intermodulation products and harmonics. The DNL test is best done with a good oscilloscope and a second DAC driven with the twos complement; you can sort out which DAC is being nonlinear with a digital adder in series with one of the DACs' inputs; adding 1 to the DAC code will shift its glitches over by a clock period. Subtracting the two traces leaves only the DNL of the shifted DAC, which is very convenient. Another point is that the dynamic performance of ADCs will depend on the relative phases of the clock and test tones; if you have phase-shiftable synthesizers, experiment with different phases to see if it makes any difference.

14.8.5 Differential Nonlinearity and Histograms

Differential nonlinearity is especially obnoxious when we're doing histograms, since the expectation value of the count in each histogram bin is proportional to the bin width, so that any differential nonlinearity introduces a systematic error into the histogram. This is not even slightly subtle—$\pm\frac{1}{2}$ ADU DNL corresponds to a 3:1 range of bin widths. There are ADCs with intrinsically identical bin widths, such as voltage-to-frequency or time-to-amplitude converters and $\Delta\Sigma$ ADCs, which are a better match for histogramming applications; an inferior expedient is to use a higher resolution converter and sum adjacent bins. Another approach is *dithering*: use a fine-resolution DAC to add pseudorandom noise to the signal, and subtract it again numerically after digitizing. This homogenizes out much of the DNL at some sacrifice in complexity and dynamic range (since you don't want to clip the signal + dither waveform). (Hint: Since DNL is localized, it takes *lots* of noise, not just a few ADU, and requires very close attention to the impulse response so the undithering can be done accurately—test the daylights out of it.)

If you're oversampling significantly, the error-smearing approach can be used in a much simpler way: add out-of-band noise at the ADC input, and filter it out again digitally afterwards (see Section 17.6). Again, you need to use a fair amount—at least several ADU, and maybe 1/4 of full scale.

14.8.6 Dynamic Errors

The two-tone test we used earlier only tests for low speed errors. High speed testing is done just the same way, but aliased (see Chapter 17). The measurement conditions are typically something like this: $f_{\text{sample}} = 10.01$ MHz, $f_{\text{in}} = 10.00$ MHz, $V_{\text{in}} = 95\%$ of full scale. The desired output shows up at 10 kHz, and the spurious signals wind up at harmonics of 10 kHz. Two-tone testing might have $f_2 = 10.031$ MHz. (See Figure 14.26.)

ADCs intended for signal processing usually have AC specs quoted in terms of spurious-free dynamic range (SFDR), which is the power ratio of the largest sinusoidal signal the ADC can handle to the largest of these spurs (usually the second or third), under given testing conditions. Another way of specifying it is the effective number of bits (ENOB), which is based on the SINAD (SIgnal to Noise And Distortion, a generalized SNR); you take the largest AC swing you can accommodate, $V_{\text{rms}} = 1/(2\sqrt{2})$ times the reference voltage, and divide it by the $1/\sqrt{12}$ ADU additive noise of a perfect digitizer, take 20 log to get decibels, and you get the mysterious-looking formula

$$\text{SINAD}|_{\text{ideal}} = 20\left[N\log 2 + \log\left[\frac{\sqrt{12}}{\sqrt{8}}\right]\right] = 6.02N + 1.76 \text{ dB}. \tag{14.30}$$

The ENOB is what you get by plugging in the observed SINAD and solving (14.30) for N. It isn't that useful for calculations, but the number is easily memorable and gives a good seat-of-the-pants feel for how much worse the ADC gets at high frequency. Data sheet preambles and signal processing books will give you precise definitions for all these things, but when you look at the actual specifications, you find that the specs are uncertain by factors of 3 to 10 in SFDR anyway, so don't sweat it too much.

These aliased tests are pretty stringent, although they don't tell you a lot about the time-domain performance; all those spurs might add up to one great big time-domain

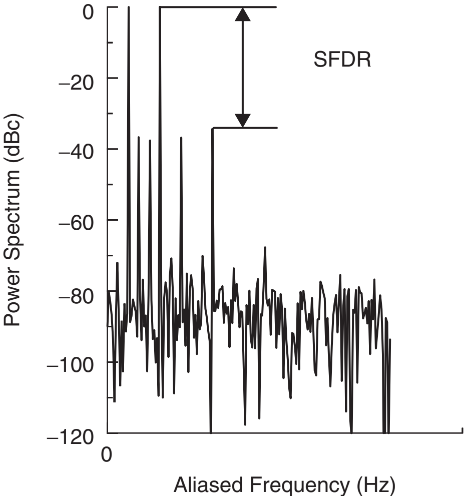

Figure 14.26. Dynamic errors: spurious-free dynamic range.

glitch due to 3 missing codes right at the zero crossing, for example; or it might have horrible cross-modulation due to an internal node that can't slew rapidly; or it might have lots of jitter, which will raise the noise floor. You care about these things, even if the SFDR looks OK, because the effects are nonlinear and the test signal isn't the same as your real data. Make sure that you test for what you care about, including at least the noise floor, missing codes, intermodulation, offset, and harmonics. Leave a big fat safety factor if you can possibly afford to—ADCs can surprise you in a lot of unpleasant ways. Ignore their idiosyncrasies at your peril.

14.8.7 Dynamic Range

The SINAD formula (14.30) is appropriate for ADCs working on AC signals, where both half-cycles have to be digitized, and the maximum rms signal is $1/(2\sqrt{2})$ of full scale. A converter dealing with unipolar signals (e.g., the output of an AM detector) has things considerably better; an ideal N bit ADC has a dynamic range (DR) of

$$\mathrm{DR(dB)} = 20\left[N \log 2 + \log\sqrt{12}\right] \approx 6.02N + 10.78\ \mathrm{dB}, \tag{14.31}$$

which is 9 dB better than the SINAD number (14.30) due to not needing to accommodate the peak-to-peak swing of a sine wave.[†] Keeping these straight is not difficult if you bear in mind what's really going on at the ADC input (see Section 15.2.1).

The dynamic range of a DAC is somewhat thornier philosophically, because an ideal DAC introduces no error whatever; its output voltage exactly corresponds to the input word.[‡] The quantization is what introduces the noise, and that has to have occurred elsewhere before the DAC word could have been generated. Accordingly, we usually talk about the *resolution* of a DAC instead. Measurements of the output spectrum of a DAC, as in the two-tone SINAD measurement above, inherently include the quantization noise contributed by the computation of the approximate sinusoids used as the input. It's worth keeping the two distinct in our minds, however, because in error propagation calculations we don't want to add in the quantization error twice. Once again, *ADCs add quantization noise, DACs do not*.

14.8.8 ADC Noise

The noise of a DAC is easily measured—sit it at various levels and look at the AC part of its output. ADC noise is a bit more problematical; it is a random motion of the comparator thresholds, causing inputs near the edge of a bin to show up randomly in the two adjoining bins. The easiest way to see it is with a low amplitude ramp signal; sweep repetitively and look for fuzziness on the edges. Remember that the conversion itself contributes $1/\sqrt{12}$ ADU of noise, so noise at least 10 dB below that is probably OK, since we aren't letting the quantization noise dominate in the first place. Sometimes we even add noise intentionally, to smear out the staircase and allow signal averaging.

[†] Another way of looking at this is the full-scale DC signal of a converter in a unipolar circuit has $2\sqrt{2}$ times the rms value of a full-scale sine wave, for the same converter used in bipolar mode.

[‡] Real DACs, of course, do produce output noise, but it's easy to measure, and usually not highly relevant.

14.8.9 Ultrafast ADCs

Really quick ADCs, such as 25 MHz at 14 bits, 50 MHz at 12 bits, or 100 MHz at 10 bits, are about as easy to use as a recipe for unicorn stew. Their performance changes dramatically with tiny changes in conditions.† If you can't possibly avoid these monolithic slabs of misery, get the manufacturer's evaluation board, and bang on it. Copy his layout and component choices slavishly, except *don't split the ground plane* (see Section 16.4.4). Do not attempt to economize on the circuit card or decoupling components when using one of these. There exist even more ridiculously fast ADCs—like 16 bits at 130 MHz—but their SNR levels are far worse than the quantization limit; they're mostly used in *software-defined radio*, where digitizing is done early in the signal processing chain and the ADC and clock jitter problems are spread out over a wide bandwidth and subsequently filtered out digitally. These usually have optional dither and noise injection to cover up the bad ADC behavior.

14.9 ANALOG BEHAVIOR OF DIGITAL CIRCUITS

14.9.1 Frequency Dividers

Frequency dividers are usually digital counters, though there are other kinds (parametric dividers and injection-locked oscillators, in particular). You can use a binary counter for sequencing your data acquisition and control tasks; a walking-ring (Johnson) counter for generating equally spaced phases for an SSB mixer; or a loop divider in a phase-locked loop frequency multiplier. There are a couple of analogish things to worry about with counters. The first is the usual sensitivity of the logic transition to supply noise and poor clock edges, which lead to timing jitter and to fuzz on the output voltage levels that will get into subsequent stages—pay attention to your bypassing if you care about that, and consider resynchronizing with an additional flip-flop stage at the end of the logic chain. (Microprocessor and FPGA outputs are especially bad for this, with poorly defined logic levels and lots of spurious junk; zero-power programmable logic and CMOS gates and flip-flops are much better.) The second is more mathematical: a divide-by-N counter has an N-fold phase ambiguity. This is seriously annoying in situations where separate units have to agree on the phase (e.g., in time-of-flight ranging). The reason is that the counter can wake up in any of N states, and even if you reset it right away, state 0 is no more likely to be correct than any of the other states. One thing to remember about clocked logic is that it is vulnerable to cable reflections at its outputs, which can cause extra clocks, missed clocks, and metastability.

14.9.2 Phase Noise and Jitter of Logic

Digital circuits have noise just like analog ones, but (as in comparators) it shows up as timing jitter. A good HCMOS system can have 10 ps jitter, 10KH ECL a few picoseconds, and ECLinPS Lite can be in the 100 fs range. Chaining functions together makes this worse, generally faster than the rms sum since variations in temperature and supply

†For carefully documented fast ADC horror stories and discussion, see Bill Travis, Demystifying ADCs: EDN Hands-On Project. *EDN* **42**, 7 (March 27, 1997), and its sequel, Remystifying ADCs. *EDN* **42**, 21 (October 9, 1997).

voltage cause systematic changes. Resynchronizing with a flip-flop running from a quiet supply and a good clock will help a lot. In bad cases, two stages of resynchronization can be better than just one.

14.9.3 Analog Uses of Gates and Inverters

CMOS gates especially have lots of analog uses. One of the best is in driving diode bridge mixers. A lot of the intermodulation spurs from diode mixers are generated during the transition times, when the diodes aren't fully on or off, so using a square wave LO improves the intermod performance. Most mixers have transformer inputs and like to have their ports properly terminated; drive them via a 47 ohm resistor and a capacitor in series. They're also good for driving MOSFET switches. Section 18.7.2 shows a multiplexer for a pyroelectric imager, in which diode switches are driven by CMOS shift register outputs. Three-state outputs can be used to switch channels, for example, by grounding one of *N* photodiode anodes, when the cathodes are all connected to single transimpedance amplifier. Zero-power programmable logic swings reliably from rail to rail, which makes it good for analog jobs.